

Polycyclic Aromatic
Hydrocarbons Modify DNA
Methylation in Animals and
Humans

Francesca Galea

Department of Epidemiology and Biostatistics,
School of Public Health, Imperial College London

A thesis submitted for the degree of Doctor of Philosophy

March 2019

Declaration of Originality

I, Francesca Galea, declare that all the work presented in this thesis is my own unless otherwise stated.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a [Creative Commons Attribution-NonCommercial 4.0 International Licence](#) (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work.

This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

There are a number of people without whom this work would not have been possible. Firstly, I would like to thank the MRC-PHE Centre for the Environment and Health for granting me studentship that allowed me to undertake this project. Next, I am grateful to my primary supervisor Professor Paolo Vineis for the opportunity to undertake this project and the guidance given over the years. Huge thanks also go to my second supervisor Dr James Flanagan for taking me under his wing, pushing me and challenging me, and for always being there. Thank you to Dr Volker Arlt for the opportunity to investigate DNA methylation in mice by providing the funding and DNA samples which were used to prepare the RRBS libraries. Thanks also goes to the investigators of EPIC-Itlay, EPIC-NL, and ESCAPE who gave me access to the data that made the epidemiological studies possible, especially Dr Giovanni Fiorito for his help and ideas.

Thank you to various members of Paolo and James' groups: Dr Michelle Plusquin and Dr Florence Guida for teaching me the ropes when it came to R scripts, beta regression and EWAS; Dr Annelie Johansson for being there to share ideas with and debate the merits of various statistical methods with; John Gallon for all his help with understanding bioinformatics analyses and tools; Dr Eric Loomis with his help troubleshooting the preparation of the RRBS libraries; Caitriona Tyndall for the lab advice and support; Nahal Masrour for helping find my feet in the Epigenetics Unit lab, for sharing protocols and for patiently training me to use the pyrosequencer; Dr Ed Curry for being an excellent sounding board and source of new ideas; and Marina Natoli and Dr Emma Bell for your cheerleading, support and for listening.

This work would not have been completed without the support of my wonderful family: Ian, Nathalie, and Clara, thank you for your unwavering positivity, for keeping me going, and for always believing me and everything I do. I'm so glad to have had such a wonderful and loving home to go to when I needed a break. And thank you for enduring all the rants I subjected you to. And thank you to Nannu Val for your cheeriness and constant interest in my work, even if you didn't always understand it.

To Graeme, I would never have submitted this work without you. You kept me going, and pushed me to give just a little bit more. Thank you for all the little and big ways that you have made my life easier since coming into my life – I am so very grateful to have had you on this journey with me.

Abstract

Human exposure to the ubiquitous environmental carcinogens polycyclic aromatic hydrocarbons (PAHs) occurs predominantly through the diet, tobacco smoke, and air pollution. While the genotoxic effects of these compounds have been well characterised, it was hypothesised that DNA methylation changes induced by PAHs could be a potential mechanism of their carcinogenicity. This study aimed to identify DNA methylation changes associated with PAH exposure in a mouse model and in human cohorts. Reduced representation bisulphite sequencing was carried out on lung tissue from Benzo[*a*]pyrene exposed mice. Additionally, PAH8 exposure from air and dietary sources estimated from land use regression models and food frequency questionnaires were used with data from Illumina Infinium HumanMethylation450 in EPIC-Italy (Training subset: N = 493; Testing subset: N = 208) and EPIC-NL (N = 132) cohorts.

Several differentially methylated CpG sites (Treated vs Untreated: N = 430; $p < 0.05$, Δ methylation > 25%), 500 b.p. windows (Treated vs Untreated: N = 1780; $p < 0.05$, Δ methylation > 25%), and probes (Air PAH8 exposure: N = 204; Dietary PAH8 exposure: N = 171; Combined air and dietary PAH8 exposure: N = 274; FDR $q < 0.05$) were identified in the analyses carried out. Although there were little to no overlaps between mouse and human studies at the CpG or gene level, in both the mouse and human analyses significantly fewer changes than expected by chance occurred at promoter regions. Additionally, the three human EWAS showed that different routes of PAH exposure may have different effects on DNA methylation, and when these exposures were combined, the methylation changes observed represented the separate exposures. These observations require further validation, but the results suggest that PAH-DNA adduct formation, which does not occur in a gene-specific manner, could be driving DNA methylation changes.

Table of Contents

1.	Chapter 1 - Introduction	26
1.1	Polycyclic Aromatic Hydrocarbons.....	26
1.1.1	Occurrence of PAHs in Air.....	28
1.1.2	Routes of Exposure	32
1.1.3	Absorption and Metabolism	33
1.1.4	Carcinogenicity and Genotoxicity	35
1.1.5	Other Consequences of PAH Exposure	52
1.2	Epigenetics	55
1.2.1	DNA Methylation	55
1.2.2	Histone Modifications.....	57
1.2.3	Epigenetics and the Environment.....	59
1.2.4	Epigenetic Changes and Cancer	60
1.3	PAH Exposure and Epigenetics.....	62
1.4	Hypothesis and Aims.....	64
2	Chapter 2 - Methods	66
2.1	RRBS of lung tissue from mice exposed to B[a]P.....	66
2.1.1	Mouse Samples.....	66
2.1.2	Preparation of Reduced Representation Bisulphite Sequencing (RRBS) Libraries	66
2.1.3	Pre-processing of Data	70
2.1.4	Finding Associations Between DNA Methylation and B[a]P Exposure	71
2.2	Air and Dietary Exposure to PAHs in EPIC Subjects	74
2.2.1	EPIC Cohort	74
2.2.2	DNA Methylation Data	76
2.2.3	Air PAH8 Exposure Estimation	77
2.2.4	Dietary PAH8 Exposure Estimation.....	78
2.2.5	Combined Air and Dietary PAH8 Exposure	81
2.2.6	Data and Models.....	81
2.2.7	Global Methylation, Methylation at Genomic Regions and EWAS.....	84
2.2.8	Building a Methylation Index of PAH8 Exposure	86
3	Chapter 3 – RRBS of Lung Tissue from Mice Exposed to B[a]P.....	89
3.1	Introduction	89
3.1.1	Measuring DNA Methylation	89
3.1.2	The Effects of PAH Exposure on DNA Methylation.....	93
3.1.3	Aims and Hypothesis.....	96

3.2	Results.....	97
3.2.1	Sequencing Statistics.....	97
3.2.2	Genomic Distribution of RRBS Libraries.....	100
3.2.3	Inter-Sample Variation.....	102
3.2.4	Principal Components Analysis.....	102
3.2.5	EWAS.....	104
3.3	Discussion.....	131
3.3.1	Inter-Sample Variation.....	131
3.3.2	Genomic Distribution of B[a]P-induced DNA Methylation Changes.....	133
3.3.3	Methylation and Gene Expression.....	134
3.3.4	Comparison of Differentially Methylated Genes to Previously Published Results.....	136
3.3.5	Potential Explanations for the Effects of B[a]P on DNA Methylation.....	136
3.3.6	Conclusions.....	138
4	Chapter 4 - DNA Methylation and Air PAH8 Exposure.....	139
4.1	Introduction.....	139
4.1.1	Policy and Compliance.....	139
4.1.2	Measuring exposure to PAHs and Air Pollutants.....	140
4.1.3	PAHs in Air and DNA Methylation.....	142
4.1.4	Effects of Other Air Pollutants on DNA Methylation.....	148
4.1.5	Aims.....	148
4.2	Results.....	151
4.2.1	Comparison of Cohort Characteristics.....	151
4.2.2	Effects of Air PAH8 Exposure on Global Methylation.....	151
4.2.3	Effects of Air PAH8 Exposure on Methylation at Genomic Regions.....	155
4.2.4	EWAS Results.....	158
4.2.5	Building a Methylation Index of Air PAH8 Exposure.....	170
4.3	Discussion.....	172
4.3.1	Air PAH8 Exposure.....	172
4.3.2	Cohort Differences.....	173
4.3.3	Statistical and Other Considerations.....	174
4.3.4	Comparison of EWAS Findings to Previously Published Results.....	177
4.3.5	Conclusions.....	179
5	Chapter 5 - DNA Methylation and Dietary PAH8 Exposure.....	180
5.1	Introduction.....	180
5.1.1	Recommendations and Policies.....	180
5.1.2	Incidence of PAHs in Food.....	182

5.1.3	Human Exposure to PAHs in Food	183
5.1.4	Measuring Human Exposure to Dietary PAHs.....	185
5.1.5	Minimising Human Exposure to PAHs from Food.....	186
5.1.6	Aims.....	186
5.2	Results.....	187
5.2.1	Presence of PAHs in Foods Dataset	187
5.2.2	Estimation of Dietary PAH8 Exposure.....	192
5.2.3	Effects of Dietary PAH8 Exposure on Global Methylation.....	197
5.2.4	Effects of Dietary PAH8 Exposure on Methylation at Genomic Regions	197
5.2.5	EWAS Results	201
5.2.6	Building a Methylation Index of Dietary PAH8 Exposure.....	210
5.3	Discussion.....	212
5.3.1	Exposure Estimates of Dietary PAH8 Exposure.....	212
5.3.2	Factors Affecting Estimation of Dietary PAH8 Intake	213
5.3.3	Comparison of Estimated Dietary PAH8 Intakes to Previous Reports.....	215
5.3.4	Comparison of EWAS Results to Previously Published Findings	216
5.3.5	Statistical and Other Considerations	217
5.3.6	Conclusions	218
6	Chapter 6 – Combined Air and Dietary PAH8 Exposure	219
6.1	Introduction	219
6.1.1	The Effects of Smoking on DNA Methylation.....	219
6.1.2	Aims.....	220
6.2	Results.....	222
6.2.1	Combined Air and Dietary PAH8 Exposure	222
6.2.2	Effects of Combined Air and Dietary PAH8 Exposure on Global Methylation.....	222
6.2.3	Effects of Combined Air and Dietary PAH8 Exposure on Methylation at Genomic Regions	225
6.2.4	EWAS Results	228
6.2.5	Building a Methylation Index of Combined Air and Dietary PAH8 Exposure.....	238
6.2.6	Comparison of the EWAS Results of Air, Dietary, and Combined Air and Dietary PAH8 Exposure	240
6.3	Discussion.....	253
6.3.1	Using Z-scores to Represent Combined Air and Dietary PAH8 Exposure	253
6.3.2	Comparison of EWAS Results to Previously Published Findings	254
6.3.3	Statistical and Other Considerations	256
6.3.4	Conclusions	256

7	Chapter 7 – General Discussion, Future Work, and Conclusions.....	258
7.1	General Discussion.....	258
7.1.1	Influence of PAH Exposure on DNA Methylation in Animals and Humans.....	258
7.1.2	Effect of Route of PAH Exposure on DNA Methylation	259
7.2	Strengths and Limitations	260
7.2.1	Power and Other Statistical Considerations	260
7.2.2	Technical Considerations and Underlying Differences in DNA Methylation	264
7.2.3	Methods for Assessing PAH Exposure	265
7.3	Future Work.....	267
7.3.1	Animal Studies	267
7.3.2	Human Epidemiological Studies.....	269
7.4	Final Conclusions.....	270
8	References	271
9	Appendices.....	299
9.1	Appendix 1 - Chapter 3 Supporting Tables and Figures.....	299
9.1.1	Model Results: Control mice vs mice exposed to low dose of B[a]P (25 mg/kg b.w.)	299
9.1.2	Model Results: Control mice vs mice exposed to medium dose of B[a]P (50 mg/kg b.w.)	304
9.1.3	Model Results: Control mice vs mice exposed to a high dose of B[a]P (75 mg/kg b.w.)	309
9.2	Appendix 2 – Chapter 4 Supporting Tables and Figures	314
9.3	Appendix 3 – Chapter 5 Supporting Tables and Figures	345
9.4	Appendix 4 - Chapter 6 Supporting Tables and Figures.....	369

Abbreviations

<u>Abbreviation</u>	<u>Meaning</u>
3' UTR	Three prime untranslated region
5' UTR	Five prime untranslated region
5-hmC	Hydroxymethylated cytosine residue
5mC	5-methylchrysenes
5-mC	Methylated cytosine residue
8-oxo-G	7,8-dihydro-8-oxoguanine
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
B[a]A	Benz[a]anthracene
B[a]P	Benzo[a]pyrene
B[b]Fl	Benzo[b]fluoranthene
B[c]F	Benzo[c]fluorene
B[ghi]P	Benzo[ghi]perylene
B[j]Fl	Benzo[j]fluoranthene
B[k]F	Benzo[k]fluoranthene
B[k]Fl	Benzo[k]fluoranthene
BER	Base excision repair
Bp	Base pairs
BPDE	B[a]P-diol epoxide
Bw	Body weight
C[cd]P	Cyclopenta[cd]pyrene
CEE	Central and Eastern Europe
ChIP-seq	Chromatin immunoprecipitation with sequencing
Chr	Chrysenes
CONTAM	EFSA Panel on Contaminants in the Food Chain
DB[ae]P	Dibenzo[a,e]pyrene
DB[ah]A	Dibenzo[a,h]anthracene
DB[ah]P	Dibenzo[a,h]pyrene
DB[ai]P	Dibenzo[a,i]pyrene
DB[a]P	Dibenzo[a,l]pyrene
DMC	Differentially methylated CpG site
DMW	Differentially methylated window
DMW	Differentially methylated window
DNA	Deoxyribose nucleic acid
DNMT	DNA methyltransferase
EC	European Commission
EFSA	European Food Safety Authority
EPA	U.S. Environmental Protection Agency
EPIC	European Prospective Investigation into Cancer and Nutrition
EPIC-NL	Subjects from the Dutch component of the EPIC cohort
ESCAPE	European Study of Cohorts for Air Pollution Effects
EWAS	Epigenome-wide association study
FAO	Food and Agriculture Organisation of the United Nations
FFQ	Food frequency questionnaire
Fl	Fluoranthene
GC-MS	Gas chromatography couple with mass spectroscopy
GG-NER	Global genomic nucleotide excision repair
GI	Gastrointestinal
GIS	Geographic information system

GLM	Generalised linear model
GNMT	Glycine N-methyltransferase
H3K27me3	Histone H3 lysine27 trimethylation
H3K36me2	Histone H3 lysine36 dimethylation
H3K36me3	Histone H3 lysine36 trimethylation
H3K4me2	Histone H3 lysine4 dimethylation
H3K4me3	Histone H3 lysine4 trimethylation
H3K9ac	Histone H3 lysine9 acetylation
H4K20me3	Histone H4 lysine20 trimethylation
HAT	Histone acetyltransferase
Hb	Haemoglobin
HBE	Human bronchial epithelial cells
HCP	High-CpG-density promoter
HDAC	Histone deacetylase
HPLC	High performance liquid chromatography
HPLC	High performance liquid chromatography
HSA	Human serum albumin
I[<i>cd</i>]P	Indeno[1,2,3- <i>cd</i>]pyrene
IARC	International Agency for Research on Cancer
IQ	Intelligence quotient
JECFA	Joint FAO/WHO expert Committee on Food Additives
LC-MS	Liquid chromatography coupled with mass spectroscopy
LCP	Low-CpG-density promoter
LINE	Long interspersed nuclear element
LOOCV	Leave-one-out cross validation
LTR	Long terminal repeat
miRNA	MicroRNA
MORGEN	Monitoring Project on Risk Factors for Chronic Diseases
NDAMR	<i>N</i> -methyl-D-aspartate receptor
NER	Nucleotide excision repair
NGS	Next generation sequencing
PAH	Polycyclic aromatic hydrocarbon
PAH8	Collective term for B[<i>a</i>]A, B[<i>b</i>]Fl, B[<i>k</i>]Fl, B[<i>ghi</i>]P, B[<i>a</i>]P, Chr, DB[<i>ah</i>]A, and I[<i>cd</i>]P
PC	Principal component
PM10	Particulate matter with a diameter of 10 µm
PM10	Particulate matter with an aerodynamic diameter less than 10 µm in size
PM2.5	Particulate matter with an aerodynamic diameter less than 2.5 µm in size
PM25	Particulate matter with a diameter of 2.5 µm
RMSE	Root mean squared error
ROS	Reactive oxygen species
RRBS	Reduced representation bisulphite sequencing
SCF	Scientific Committee on Food
SCOOP	Scientific Cooperation on Food
SD	Standard deviation
ShPARG	PARG-deficient human bronchial epithelial cells
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
SNP	Short nucleotide polymorphism
TCDD	2,3,7,8-tetrachlorodibenzo- <i>p</i> -dioxin
TC-NER	Transcription-coupled nucleotide excision repair.
TCR	Transcription-coupled repair

TDS	Total diet study
TDS	Total Diet Study
TEF	Toxic equivalency factor
TET	Ten-eleven transferase
TSS	Transcription start site
TTS	Transcription termination site
WBC	White blood cell
WGBS	Whole genome bisulphite sequencing
WHO	World Health Organisation

List of Tables

Table 1.1. Breakdown of PAH Exposure of a "Reference Man" ²¹	32
Table 1.2. Table of common PAHs, their structure, molecular weight, and carcinogenic classification in order of increasing molecular weight.....	36
Table 2.1. Matrix used in Fisher's Exact Tests to assess enrichment or depletion of methylation changes at genomic regions compared to distribution of all tested sites/windows	72
Table 2.2. Matrix used in Fisher's Exact Tests to compare hypomethylation and hypermethylation events for each genomic annotation to the overall ratio.	73
Table 2.3. Table of number of CpG probes at each genomic region analysed in the EPIC-Italy derived Training and Testing Datasets, and the EPIC-NL dataset	85
Table 2.4 Matrix used in Fisher's Exact Tests to assess enrichment or depletion of methylation changes at genomic regions compared to distribution of all tested sites/windows	86
Table 2.5 Matrix used in Fisher's Exact Tests to compare hypomethylation and hypermethylation events for each genomic annotation to the overall ratio.	87
Table 3.1. Table of sequencing statistics for each mouse RRBS library.	98
Table 3.2 Table of Fisher's test results comparing the number of DMWs (N = 1780) and DMCs (N = 430) to all tested windows (N = 152,720) and CpG sites (N = 38,874) at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred.	107
Table 3.3 Table of Fisher's test results comparing the number of hypermethylation changes (DMWs: N = 660; DMCs: N = 145) to hypomethylation changes (DMWs: N = 1120; DMCs: N = 285) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected.	109
Table 3.4 Table summarising the number of differentially methylated CpG sites and differentially methylated windows for each model.	111
Table 3.5 Summary of characteristics and results of significant DMCs across all models.	113

Table 3.6. Table showing summary of correlation between gene expression and methylation levels of DMCs. The gene expression data were generated from Agilent 4x44K oligonucleotide microarrays as part of the study carried out by Labib et al.¹⁴². The results for different microarray transcripts for the same gene are separated by a semicolon. 116

Table 3.7 Summary of characteristics and results of significant DMWs across all models. The position column indicates the position of the first base in each 500 bp window. 121

Table 3.8. Table showing summary of correlation between gene expression and methylation levels of DMWs. The gene expression data were generated from Agilent 4x44K oligonucleotide microarrays as part of the study carried out by Labib et al.¹⁴². The results for different microarray transcripts for the same gene are separated by a semicolon. 128

Table 3.9 Table summarising overlaps between chapter results and B[a]P gene interactions reported in the 135

Table 4.1 Table summarising published results of associations between PAH exposure and DNA methylation. 144

Table 4.2 Table summarised published results of associations between air pollutants and DNA methylation 149

Table 4.3 Table of cohort characteristics for the training, testing and EPIC-NL EPIC subjects. 152

Table 4.4 Table of beta regression results looking for differences in global methylation between quartiles of air PAH8 exposure. The lowest quartile (Q1) was used as the reference quartile. 154

Table 4.5 Table of beta regression results looking differences in methylation levels at CpG islands, shores and shelves between quartiles of air PAH8 intake. The lowest quartile (Q1) was used as the reference quartile. Entries in bold indicate statistical significance ($p < 0.05$). 156

Table 4.6. Table of beta regression results looking differences in methylation levels at promoter, 3' UTR, gene body, and intergenic regions between quartiles of air PAH8 intake. The lowest quartile (Q1) was used as the reference quartile. Entries in bold indicate statistical significance ($p < 0.05$). 157

Table 4.7 Model results for the Bonferroni significant ($p < 1.38e^{-7}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship between air PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status. 160

Table 4.8 Table of characteristics of probes found to be significantly associated with air PAH8 exposure at the Bonferroni level ($p < 1.38e-7$) in the training dataset.	163
Table 4.9. Table of Fisher’s test results comparing the number of differentially methylated probes (N = 274) and all tested probes (N = 362,394) in the training dataset EWAS at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred.	165
Table 4.10. Table of Fisher’s test results comparing the number of hypermethylation changes (N = 99) to hypomethylation changes (N = 175) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected.	166
Table 5.1 Table showing groups of PAHs as described by various European committees. Shading indicates that a PAH is included in that group.....	181
Table 5.2 Median dietary intake in European consumers ²⁸³	182
Table 5.3 Table of the number of food items and number of type entries in the datasets.....	188
Table 5.4 List of polycyclic aromatic hydrocarbons included in the datasets.....	189
Table 5.5 Table of cohort characteristics for the Training, Testing and EPIC-NL EPIC subjects.	194
Table 5.6 Table of beta regression results looking for differences in global methylation between quartiles of dietary PAH8 intake. The lowest quartile (Q1) was used as the reference quartile.	199
Table 5.7. Table of beta regression results looking differences in methylation levels at CpG islands, shores and shelves between quartiles of dietary PAH8 intake. The lowest quartile (Q1) was used as the reference quartile. Entries in bold indicates statistically significant results ($p < 0.05$).....	199
Table 5.8. Table of beta regression results looking differences in methylation levels at promoter, 3’UTR, gene body, and intergenic regions between quartiles of dietary PAH8 intake. The lowest quartile (Q1) was used as the reference quartile. Entries in bold indicates statistically significant results ($p < 0.05$).....	200
Table 5.9 Model results for the Bonferroni significant ($p < 1.38e^{-7}$) EWAS probes in the three datasets: Training, Testing and EPIC-NL. All results are from beta regression models assessing the relationship between dietary PAH8 exposure and the methylation beta values for each probe. The Training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and	

subject centre. The Testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status. 204

Table 5.10 Table of characteristics of probes found to be significantly associated with dietary PAH8 exposure at the Bonferroni level ($p < 1.38e-7$) in the Training dataset. 206

Table 5.11. Table of Fisher’s test results comparing the number of differentially methylated probes ($N = 171$) and all tested probes ($N = 362,310$) in the Training dataset EWAS at various genomic regions. An $OR < 1$ indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an $OR > 1$ indicates that more changes than expected occurred. 207

Table 5.12. Table of Fisher’s test results comparing the number of hypermethylation changes ($N=86$) to hypomethylation changes ($N=85$) compared to the overall ratio of hypermethylated to hypomethylated probes. An $OR < 1$ indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected. 207

Table 6.1 Table of genes most frequently reported to be differentially methylated in association with smoking. The total number of reports column includes the total number of times each CpG probe was reported by each study, with some studies reporting multiple probes associated with the same gene, and in many instances the same probe was reported by more than one study. 221

Table 6.2 Table of cohort characteristics for the training, testing and EPIC-NL EPIC subjects. 223

Table 6.3 Table of beta regression results looking for differences in global methylation between quartiles of combined air and dietary PAH8 exposures. The lowest quartile (Q1) was used as the reference quartile. 224

Table 6.4 Table of beta regression results looking differences in methylation levels at CpG islands, shores and shelves between quartiles of combined air and dietary PAH8 exposures. The lowest quartile (Q1) was used as the reference quartile. 226

Table 6.5. Table of beta regression results looking differences in methylation levels at promoter, 3’ UTR, gene body and intergenic regions between quartiles of combined air and dietary PAH8 exposures. The lowest quartile (Q1) was used as the reference quartile. 227

Table 6.6 Model results for the Bonferroni significant ($p < 1.38e^{-7}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship between combined air and dietary PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status. 231

Table 6.7 Table of characteristics of probes found to be significantly associated with combined air and dietary PAH8 exposure at the Bonferroni level ($p < 1.38e^{-7}$) In the training dataset..... 234

Table 6.8. Table of Fisher’s test results comparing the number of differentially methylated probes (N = 274) and all tested probes (N = 362,394) in the training dataset EWAS at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred. 235

Table 6.9. Table of Fisher’s test results comparing the number of hypermethylation changes (N=99) to hypomethylation changes (N=175) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected. 236

Table 6.10 Model results for probes that were FDR significant ($q < 0.05$) in both the combined air and diet PAH8 exposure EWAS model and the air PAH8 exposure only EWAS model. All results are from beta regression models run on the training dataset and were adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. 244

Table 6.11. Table of characteristics of probes found to be significantly associated with combined air and dietary PAH8 exposure and air PAH8 exposure only at the FDR significance levels ($q < 0.05$). 247

Table 6.12. Model results for probes that were FDR significant ($q < 0.05$) in both the combined air and diet PAH8 exposure EWAS model and the dietary PAH8 exposure only EWAS model. All results are from beta regression models run on the training dataset and were adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre..... 249

Table 6.13. Table of characteristics of probes found to be significantly associated with combined air and dietary PAH8 exposure and dietary PAH8 exposure only at the FDR significance levels ($q < 0.05$). 251

Table 8.1 Table of Fisher’s test results comparing the number of DMWs (N = 1910) and DMCs (N = 699) to all tested windows (N = 152,720) and CpG sites (N = 38,874) at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred. 300

Table 8.2 Table of Fisher’s test results comparing the number of hypermethylation changes (DMWs: N = 993; DMCs: N = 478) to hypomethylation changes (DMWS: N = 917; DMCs: N = 221) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected. 302

Table 8.3 Table of Fisher’s test results comparing the number of DMWs (N = 2671) and DMCs (N = 768) to all tested windows (N = 152,720) and CpG sites (N = 38,874) at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred. 305

Table 8.4 Table of Fisher’s test results comparing the number of hypermethylation changes (DMWS: N = 584; DMCs: N = 263) to hypomethylation changes (DMWs: N = 2087; DMCs: N = 505) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected. 307

Table 8.5 Table of Fisher’s test results comparing the number of DMWs (N = 1952) and DMCs (N = 664) to all tested windows (N = 152,720) and CpG sites (N = 38,874) at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred. 310

Table 8.6 Table of Fisher’s test results comparing the number of hypermethylation changes (DMWS: N = 813; DMCs: N = 248) to hypomethylation changes (DMWs: N = 1139; DMCs: N = 416) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected. 312

Table 8.7. Model results for the FDR significant ($p < 2.8e^{-5}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship between air PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status. 314

Table 8.8. Table of characteristics of probes found to be significantly associated with air PAH8 exposure at the FDR level ($p < 2.8e^{-5}$) in the training dataset. 326

Table 8.9. Table comparing results published by Tryndyak et al. (2018)³²⁵ and the air PAH8 exposure EWAS results 336

Table 8.10. Table showing overlaps between results of the air PAH8 exposure EWAS, and results from published smoking EWAS. Overlaps were identified by looking for exact CpG probes and by looking for probes with the same genes. 337

Table 8.11. Model results for the FDR significant ($p < 2.4e^{-5}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship

between dietary PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status..... 345

Table 8.12. Table of characteristics of probes found to be significantly associated with dietary PAH8 exposure at the FDR level ($p < 2.4e-5$) in the training dataset. 355

Table 8.13. Table comparing results published by Tryndyak et al. (2018)³²⁵ and the combined PAH8 exposure EWAS results..... 363

Table 8.14. Table showing overlaps between results of the dietary PAH8 exposure EWAS, and results from published smoking EWAS. Overlaps were identified by looking for exact CpG probes and by looking for probes with the same genes. 364

Table 8.15 Model results for the FDR significant ($p < 3.8e^{-5}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship between combined air and dietary PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status. 369

Table 8.16. Table of characteristics of probes found to be significantly associated with combined air and dietary PAH8 exposure at the FDR level ($p < 3.8e^{-5}$) in the training dataset..... 385

Table 8.17. Table comparing results published by Tryndyak et al. (2018)³²⁵ and the combined PAH8 exposure EWAS results..... 398

Table 8.18. Table showing overlaps between results of combined air and dietary PAH8 exposure, and results from published smoking EWAS. Overlaps were identified by looking for exact CpG probes and by looking for probes with the same genes..... 399

List of Figures

Figure 1.1. Overview of the four carcinogenic mechanisms of PAHs. A: The diol epoxide mechanism; B: One electron oxidation; C: Meso-region biomethylation and benzylic oxidation; D: Ortho-quinone/reactive oxygen species mechanism.	38
Figure 1.2. Metabolism of B[a]P via the diol epoxide mechanism resulting in the formation of DNA adducts.....	41
Figure 1.3. A: Chemical structure of B[a]P as an example of a PAH with a bay region (or cove region based on a more recent definition). B: Chemical structure of DB[a,l]P as an example of a PAH with a fjord region.....	46
Figure 2.1 Schematic of MspI recognition sites (A) and cleavage pattern to form sticky ends (B).	68
Figure 2.2. Schematic showing ends preparation on MspI digested fragments.	68
Figure 2.3. Schematic showing ligation of adapter sequences to DNA fragments.	68
Figure 2.4. Schematic illustrating the effects of bisulphite conversion. Methylated cytosine bases are unaffected while unmethylated cytosines are converted to thymines.	69
Figure 2.5.Schematic showing where the first enrichment PCR primers bind (A.), the amplicons for each strand and where the second set of PCR primers bind (B.) to produce fragments to be sent for sequencing after clean up and quality control.	70
Figure 3.1 Bar charts showing the number of sequencing reads (A), the number of CpG sites (B), and the number of 500 bp windows (C) for each mouse sample coloured by B[a]P dose.	99
Figure 3.2.Bar chart of percentage methylated cytosine bases for each mouse sample coloured by B[a]P dose.	100
Figure 3.3 A: Bar chart showing the genomic distribution of CpG sites in the mm10 genome. B: Bar chart of genomic distribution of CpG sites from one of the RRBS libraries. C: Bar chart of the genomic distribution of the CpG sites common to all samples (N = 38,874). D: Bar chart of the genomic distribution of the 500 bp windows common to all samples (N = 152,691).	101
Figure 3.4 A&C: Bar charts showing the percentage variance explained by each PC for the 500 bp window (A) and the CpG site (C) PCA analyses. B&D: Scatterplot of the first and second PCs for the 500 bp (B) and CpG sites (D) PCA analyses coloured by B[a]P exposure dose.....	103

Figure 3.5 Heatmaps of the 430 differentially methylated CpG sites (A) and the 1780 differentially methylated 500 bp windows (B) from the treated vs untreated models. 105

Figure 3.6 Comparison of the genomic distribution of differentially methylated sites (N = 430) and all tested sites (N = 38,874) (A) and of differentially methylated windows (N = 1780) and all tested windows (N = 152,691) (B). The coloured bars show the proportion of significant sites/windows, the grey outline bars show the proportion of all sites/windows tested, i.e. the expected distribution. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.005$ following Fisher’s Exact test. 108

Figure 3.7 Comparison of the genomic distribution of hypermethylated (N = 145) and hypomethylated (N = 285) sites (A), and hypermethylated (N = 660) and hypomethylated (N = 1120) windows (B). The coloured bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.005$ following Fisher’s Exact test. 110

Figure 3.8 Venn diagrams showing the overlaps between the four models run for the sites analysis (A) and the 500 bp windows analysis (B). 112

Figure 3.9 Scatterplot of the Log_2 expression of one of the transcripts of the D5Ertd579e gene against the methylation levels of the 500 bp window located at position 36696501 on chromosome 5. Expression and methylation were found to be significantly positively correlated using Spearman’s correlation. 129

Figure 3.10 Summary of the analysis looking for enrichment or depletion of differential methylation in different genomic regions across the four models for the sites analysis (A) and windows analysis (B). 130

Figure 4.1 Summary of cohort characteristics. A: Distribution and air PAH8 exposure in the training, testing and EPIC-NL cohorts. B: Distribution of age in the three datasets. C: Gender proportions within each dataset. D: Smoking status proportions within each dataset. E: Proportion of subjects with cancer case or control status in each dataset. 152

Figure 4.2 Mean methylation distributions of global methylation (A), shores (B), shelves (C), CpG islands (D), promoters (E), 3’ UTRs (F), gene bodies (G), and intergenic regions (H) in the Training (yellow), Testing (pink), and EPIC-NL (purple) datasets 153

Figure 4.3 A: Manhattan plot showing the $-\log_{10}$ transformed p values of the 362,404 CpG probes tested arranged by chromosome. The red line indicates the threshold for Bonferroni correction for multiple testing, and the blue line indicates the FDR threshold. B: Volcano plot showing the $-\log_{10}$ transformed p values of the 362,404 CpG probes against the β -coefficient for dietary PAH8 intake.

Coloured points indicate significance after FDR correction, with red points indicating a decrease in methylation and orange points indicating an increase in methylation. C: QQ plot showing the observed $-\log_{10}$ transformed p values against the expected $-\log_{10}$ transformed p values from the EWAS. 159

Figure 4.4 A: Comparison of the genomic distribution of differentially methylated probes (N = 204) and all tested probes (N = 362,404) in the training dataset EWAS. The filled yellow bars show the proportion of significant probes, the grey outline bars show the proportion of all probes tested, i.e. the expected distribution. B: Comparison of the genomic distribution of hypermethylated (N = 74) and hypomethylated (N = 130) probes. The filled yellow bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. For both plots, * indicates $p < 0.05$, ** indicates $p < 0.01$ and *** indicates $p < 0.001$ following Fisher’s Exact test. 167

Figure 4.5. Modified coMET plot of a 2kb region on chromosome 8 where multiple CpG sites were found to be hypomethylated. The top panel of the figure is a regional association plot showing the beta coefficients from the beta regression models of these probes from the EWAS by genomic position. The colour of the points corresponds to the \log_{10} P value, where orange indicates FDR-significant probes. The central panel shows the genomic landscape of the region with respect to genes, CpG islands, chromatin state, clusters of DNase, SNPs and regulatory features. The bottom panel shows a correlation matrix of the methylation values for each probe where red indicates a strong positive correlation, blue a strong negative correlation, and white a lack of correlation between the probes. 168

Figure 4.6. Modified coMET plot of a 2kb region on chromosome 8 where multiple CpG sites were found to be hypomethylated. The top panel of the figure is a regional association plot showing the beta coefficients of these probes from the EWAS by genomic position. The colour of the points corresponds to the \log_{10} P value, where orange indicates FDR-significant probes. The central panel shows the genomic landscape of the region with respect to genes, CpG islands, chromatin state, clusters of DNase, SNPs and regulatory features. The bottom panel shows a correlation matrix of the methylation values for each probe where red indicates a strong positive correlation, blue a strong negative correlation, and white a lack of correlation between the probes. 169

Figure 4.7 Plots showing the correlation between the air PAH8 exposure predicted by the elastic net model against the real air PAH8 exposure for each subject. Figures A and B show the results from the full model for the training and testing sets respectively. 171

Figure 5.1 Data by country. A&B: Number of food items (A) and food types (B) by country. C&D: Number of studies within the food items (C) and food types (D) datasets by country. E&F: Heatmaps representing the proportion of data in each food class coming from each country for the food items (E) and food types (F) datasets. 191

Figure 5.2 A: Boxplot showing PAH8 distributions from the food items (lighter colour on the left) and food types (darker colour on the right) datasets grouped by food class. B: Bar chart showing the median PAH8 concentration from the food items (lighter colour on the left) and food types (darker colour on the right) datasets grouped by food class. 193

Figure 5.3 Distribution of dietary PAH8 intake in ng/day for each of the three datasets: Training (yellow), Testing (red) and EPIC-NL (purple)..... 195

Figure 5.4 Correlation between Air (ng/m³) and Dietary (ng/day) PAH8 exposures in the Training (A), Testing (B), and EPIC-NL (C) datasets. 196

Figure 5.5 A: Manhattan plot showing the $-\log_{10}$ transformed p values of the 362,310 CpG probes tested arranged by chromosome. The red line indicates the threshold for Bonferroni correction for multiple Testing, and the blue line indicates the FDR threshold. B: Volcano plot showing the $-\log_{10}$ transformed p values of the 362,310 CpG probes against the β -coefficient for dietary PAH8 intake. Coloured points indicate significance after FDR correction, with red points indicating a decrease in methylation and orange points indicating an increase in methylation. C: QQ plot showing the observed $-\log_{10}$ transformed p values against the expected $-\log_{10}$ transformed p values from the EWAS. 202

Figure 5.6 A: Comparison of the genomic distribution of differentially methylated probes (N = 171) and all tested probes (N = 362,310) in the Training dataset EWAS. The filled yellow bars show the proportion of significant probes, the grey outline bars show the proportion of all probes tested, i.e. the expected distribution. B: Comparison of the genomic distribution of hypermethylated (N = 86) and hypomethylated (N = 85) probes. As in A, the filled yellow bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. For both plots, * indicates $p < 0.05$ and ** indicates $p < 0.01$ following Fisher’s Exact test. 208

Figure 5.7. Modified coMET plot of a 2kb region on chromosome 15 where multiple CpG sites were found to be hypomethylated. The top panel of the figure is a regional association plot showing the beta coefficients of these probes from the EWAS by genomic position. The colour of the points corresponds to the \log_{10} P value, where orange indicates FDR-significant probes. The central panel shows the genomic landscape of the region with respect to genes, CpG islands, chromatin state, clusters of DNase, SNPs and regulatory features. The bottom panel shows a correlation matrix of the methylation values for each probe where red indicates a strong positive correlation, blue a strong negative correlation, and white a lack of correlation between the probes..... 209

Figure 5.8 Plots showing the correlation between the dietary PAH8 exposure predicted by the elastic net model against the real combined PAH8 exposure for each subject. Figures A and B show the results from the full model for the Training and Testing sets respectively. 211

Figure 6.1 Distribution of Z-scores for the combined air and dietary PAH8 exposures for each of the three datasets: EPIC-Italy Training (yellow), EPIC-Italy - Testing (red) and EPIC-NL (purple)..... 223

Figure 6.2 A: Manhattan plot showing the $-\log_{10}$ transformed p values of the 362,394 CpG probes tested arranged by chromosome. The red line indicates the threshold for Bonferroni correction for multiple testing ($p < 1.38 \times 10^{-7}$), and the blue line indicates the FDR threshold ($p < 3.76 \times 10^{-5}$). B: Volcano plot showing the $-\log_{10}$ transformed p values of the 362,394 CpG probes against the β -coefficient for dietary PAH8 intake. Coloured points indicate significance after FDR correction, with red points indicating a decrease in methylation and orange points indicating an increase in methylation. C: QQ plot showing the observed $-\log_{10}$ transformed p values against the expected $-\log_{10}$ transformed p values from the EWAS..... 229

Figure 6.3 A: Comparison of the genomic distribution of differentially methylated probes (N = 274) and all tested probes (N = 362,394) in the training dataset EWAS. The filled yellow bars show the proportion of significant probes, the grey outline bars show the proportion of all probes tested, i.e. the expected distribution. B: Comparison of the genomic distribution of hypermethylated (N = 175) and hypomethylated (N = 99) probes. As in A, the filled yellow bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. For both plots, * indicates $p < 0.05$, ** indicates $p < 0.01$ and *** indicates $p < 0.001$ following Fisher's Exact test. 237

Figure 6.4 A and B: Plots showing the correlation between the combined PAH8 exposure predicted by the elastic net model against the real combined PAH8 exposure for each subject in the training and testing datasets respectively. 239

Figure 6.5 Venn diagram showing the overlap between the FDR-significant CpG sites between the air PAH8 exposure model (green), the dietary PAH8 exposure model (red), and the model of combined PAH8 exposure (blue)..... 242

Figure 6.6. Heatmap summarising the Fisher's test results looking for enrichment or depletion of methylation differences at particular genomic features across the three models: Air PAH8 exposure, dietary PAH8 exposure and combined air and dietary PAH8 exposure. White boxes indicate non-significant results ($p > 0.05$), and coloured boxes indicate significant results ($p < 0.05$). Orange indicates less methylation differences than expected (depletion) and purple indicates more methylation differences than expected (enrichment)..... 243

Figure 8.1 Heatmaps of the 699 differentially methylated CpG sites (A) and the 1910 differentially methylated 500 bp windows (B) from the control vs low dose model..... 299

Figure 8.2 Comparison of the genomic distribution of differentially methylated sites (N = 699) and all tested sites (N = 38,874) (A) and of differentially methylated windows (N = 1910) and all tested windows (N = 152,691) (B). The coloured bars show the proportion of significant sites/windows, the grey outline bars show the proportion of all sites/windows tested, i.e. the expected distribution. In all

cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.005$ following Fisher's Exact test. 301

Figure 8.3 Comparison of the genomic distribution of hypermethylated (N = 145) and hypomethylated (N = 285) sites (A), and hypermethylated (N = 660) and hypomethylated (N = 1120) windows (B). The coloured bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. In all cases, these results are for the treated vs untreated models. For all plots, *** indicates $p < 0.005$ following Fisher's Exact test. 303

Figure 8.4 Heatmaps of the 768 differentially methylated CpG sites (A) and the 2671 differentially methylated 500 bp windows (B) from the control vs medium dose model. 304

Figure 8.5 Comparison of the genomic distribution of differentially methylated sites (N = 768) and all tested sites (N = 38,874) (A) and of differentially methylated windows (N = 2671) and all tested windows (N = 152,691) (B). The coloured bars show the proportion of significant sites/windows, the grey outline bars show the proportion of all sites/windows tested, i.e. the expected distribution. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.005$ following Fisher's Exact test. 306

Figure 8.6 Comparison of the genomic distribution of hypermethylated (N = 263) and hypomethylated (N = 505) sites (A), and hypermethylated (N = 584) and hypomethylated (N = 2087) windows (B). The coloured bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. In all cases, these results are for the treated vs untreated models. For all plots, *** indicates $p < 0.005$ following Fisher's Exact test. 308

Figure 8.7 Heatmaps of the 664 differentially methylated CpG sites (A) and the 1952 differentially methylated 500 bp windows (B) from the control vs high dose model..... 309

Figure 8.8 Comparison of the genomic distribution of differentially methylated sites (N = 664) and all tested sites (N = 38,874) (A) and of differentially methylated windows (N = 1952) and all tested windows (N = 152,691) (B). The coloured bars show the proportion of significant sites/windows, the grey outline bars show the proportion of all sites/windows tested, i.e. the expected distribution. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$ and *** indicates $p < 0.005$ following Fisher's Exact test. 311

Figure 8.9 Comparison of the genomic distribution of hypermethylated (N = 248) and hypomethylated (N = 416) sites (A), and hypermethylated (N = 813) and hypomethylated (N = 1139) windows (B). The coloured bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. In all

cases, these results are for the treated vs untreated models. For all plots, *** indicates $p < 0.005$ following Fisher's Exact test. 313

1. Chapter 1 - Introduction

1.1 Study Scope

This thesis aims to understand the effects of exposure of the ubiquitous environmental carcinogens, polycyclic aromatic hydrocarbons (PAHs) on DNA methylation which is the most commonly studied epigenetic mechanism. A number of *in vitro*, *in vivo* and epigenetic studies have provided evidence of the carcinogenicity of PAHs, and the DNA methylation landscape has been shown to be dysregulated in cancer. The evidence supporting a link between DNA methylation and a number of environmental exposures such as air pollutants is ever-increasing, and has been well-established in relation to tobacco smoke, a major source of PAH exposure. Additionally, there is a growing body of evidence suggesting an association between PAH exposure and DNA methylation, particularly in *in vitro* and some *in vivo* studies. However, the link between environmentally-relevant doses of PAHs and DNA methylation in humans is, as yet, not understood.

The studies presented in this thesis aim to identify any associations between PAHs and DNA methylation, and shed further light on this emerging area of research. A study in the lung tissue of mice exposed to benzo[*a*]pyrene, a known carcinogen, was carried out to attempt to link DNA methylation data from reduced representation bisulphite sequencing (RRBS) to other genotoxic data such as gene expression and DNA adducts. Additionally, the relationship between PAH exposure from air and dietary sources estimated in a number of subjects from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort and DNA methylation measured in the blood of these subjects was analysed.

The review of the literature presented in the introductory chapter of this thesis was conducted in waves across the period of the PhD project. The first two sections of the introduction covering PAHs and epigenetics respectively are well-researched areas, with a number of authoritative reviews published relating to both topics. The literature review aimed

to summarise the current state of the knowledge with respect to both topics by covering the most important aspects of both topics, using a number of key publications. The third section of the introduction covering the link between PAHs and epigenetics was initially researched within the first 9 months of the PhD project and was refreshed periodically throughout the remainder of the project. The following search terms were used on Pubmed to identify newly emerging research to be incorporated into the literature review:

- “Polycyclic aromatic hydrocarbons epigenetics”
- “Polycyclic aromatic hydrocarbons DNA methylation”
- “Benzo[a]pyrene epigenetics
- Benzo[a]pyrene DNA methylation

1.2 Polycyclic Aromatic Hydrocarbons

Polycyclic aromatic hydrocarbons (PAHs) are ubiquitous environmental procarcinogens that are formed from the incomplete combustion of organic materials, mainly saturated hydrocarbons. PAHs are the largest groups of chemical compounds known to be carcinogenic, with over 10,000 compounds believed to belong to this group¹. They are procarcinogens because their genotoxic properties are only apparent following metabolism and prior to metabolism they are chemically inert. Pyrosynthesis and pyrolysis are the two mechanisms by which PAHs are formed, with low molecular weight hydrocarbons preferring the former mechanism, and higher alkanes, such as those present in fossil fuels, the latter^{2,3}. Formation of PAHs may also result from petrogenic mechanisms which include the storage and transportation of crude oil and its products, or biologically from a number of sources such as biological degradation of vegetation³. Increased molecular weight has been associated both with increased carcinogenicity, lower cytotoxicity and decreased acute toxicity^{2,4,5}. Approximately 500 PAHs and their related compounds have been detected in air⁶. PAHs are found in coal tar, roofing tar, crude oil, and creosote, with some also used in medicines, dyes, plastics and pesticides^{2,3}. The vast majority of the discussion in this thesis will focus only on unsubstituted PAHs,

however it is important to note that substituted PAH compounds exist and have been reported to be toxic. Additionally, several reviews about PAHs have already been published ^{2,7-13} and this introduction aims to summarise the current state of the knowledge.

In 2007, 504,000 tonnes of PAHs were produced globally, with residential and commercial biomass burning accounting for over 60% of this ¹⁴. More than half of these global emissions came from East Asia, South Asia and Southeast Asia. Over 6 % were compounds with higher molecular weights which are the most carcinogenic, with developed countries having less than developing countries ¹⁴.

European PAH emissions have fallen by about 50% in the recent past ², decreased emissions have been observed from developed countries which produced 38,000 tonnes in 2008 ¹⁴, and the levels of PAHs are expected to continue to decline for the most part ^{14,15}. A Spanish province has also shown a steady decline in PAH levels between 2006 and 2011 ¹⁶. Antarctica has been reported to have the lowest concentrations of PAHs amongst other organic pollutants, and this is possibly explained due to it being relatively underdeveloped compared to the other continents ¹⁷.

1.2.1 Occurrence of PAHs in Air

Lighter PAHs are usually found in the vapour phase of air, but larger molecules tend to be adsorbed on to particles, particularly particles with a diameter smaller than 2.5 µm (PM 2.5) ¹⁸. Chrysene (Chr), benzo[ghi]perylene (B[ghi]P) and benzo[b]fluoranthene (B[b]Fl) have been reported to be the predominant compounds in urban air samples ¹⁹, while another study reported that fluoranthene (Fl) was the most abundant PAH in air samples and that dibenzo[a,l]pyrene (DB[a,l]P) is the most carcinogenic ²⁰. Benzo[a]pyrene (B[a]P) is the most commonly studied carcinogenic PAH and about half of the outdoor B[a]P particulate concentration usually penetrates indoors, with indoor levels of carcinogenic PAHs ranging from 1-80 ng/m³ ²¹. Once emitted, PAHs may undergo changes such as binding to particulate matter, oxidation reactions, activation by ultraviolet radiation and degradation ²².

The annual average levels of B[a]P in the 1960's in Europe were greater than 100 ng/m³, but by the early 90's these ranged from less than 1 ng/m³ in rural areas to around 6 ng/m³ in busy streets with many traffic and emission sources ⁶. Comparison of the concentrations from studies measuring multiple PAHs is difficult, since not all studies measure the concentrations of the same compounds. In the USA in the 1980's, the concentrations of carcinogenic PAHs were highly variable ranging from 0.2-3 ng/m³ in rural areas to 15-50 ng/m³ in urban cities ²¹. A more recent study has used air samples from Fresno, California to represent high exposure and levels were 4.4 ng/m³, while samples taken from Stanford, California to represent a low exposure had PAH concentrations of 0.6 ng/m³ ²³. An assessment of 16 PAHs in roadside samples from Hangzhou, China in 2014-2015 found concentrations of 750-1142 ng/m³ during the summer and 1050-1483 ng/m³ in the winter, with low molecular weight compounds accounting for 77-86% of the samples ⁵. In 1994 in Sweden, the concentration of 14 PAHs was 100-200 ng/m³ ²⁰. A more recent study in Europe found that the London/Oxford area had the lowest levels of a sum of 8 carcinogenic PAHs (1 ng/m³) with the highest levels reported in Copenhagen, Athens and Rome (2-2.1 ng/m³) ²⁴. Another recent study of a Spanish province found the concentration of a sum of 6 PAHs ranged from 0.3 – 8.29 ng/m³, and B[a]P levels ranged from 0.05 to 0.88 ng/m³ ¹⁶. Copenhagen was found to have the highest levels (6.43 ng/m³) of a sum of 8 PAHs in a study carried out between 2008 and 2011 in cities from 10 European countries ²⁵. Lastly, concentrations of PAHs in air have been found to be higher in the winter than the summer probably due to increased fuel burning during the winter ^{5,15,16,24}.

The emission source of PAHs greatly affects the profile composition ^{5,6,20,26}. There are several emission sources, but the main ones are domestic, mobile, industrial, agricultural and natural ^{2,26}. Natural emissions include those from forest fires caused by natural sources such as lightning strikes, and volcanic eruptions. Roadside air samples have found PAH profiles from multiple sources including fuel combustion, and residential and industrial emissions ⁵. Specifically with respect to high molecular weight compounds, different roadside profiles have been reported which depend on various traffic-related variables such as traffic volume and engine loading ⁵. The contribution of each emission

source to the total emission of PAHs may vary from one country to another and when comparing rural and urban areas. The correlation between PAHs and other air pollutants has been found to be highly variable across different parts of Europe which supports the theory that the emission source/s of PAHs have unique compositions which could also affect interactions with other compounds and pollutants ^{15,24}.

In 1976, the U.S. Environmental Protection Agency (EPA) made a list of 16 “priority PAHs” in order to assess human health risks from drinking water ²⁷. The criteria used in the creation of this list were as follows: analytical standards for the compounds had to be available, the compounds must be known to occur in the environment, and the compounds must be known to be toxic. Acenaphthene, naphthalene and fluoranthene made the list because they were already included in a previous list of 65 toxic pollutants; Benz[a]anthracene (B[a]A), B[a]P, B[b]Fl, benzo[k]fluoranthene (B[k]Fl), Chr, dibenz[a,h]anthracene (DB[a,h]A), and indeno[1,2,3-cd]pyrene (I[cd]P) were included due to the availability of analytical standards as well as meeting some or all of the inclusion criteria; Acenaphthalene, fluorene (F), and phenanthrene were included due to being suspected carcinogens in water supplies; Anthracene and pyrene were included due to their prevalence in coal tar and other dyes; and lastly B[ghi]P to represent a six-ringed member of the class ²⁷. While this list was never introduced to legislation, it became popular in many countries. Andersson and Achten (2015) ²⁸ reviewed the usefulness of this list and found that there were several advantages to having such a list including time- and cost-effectiveness of analysis, the practicality of being able to analyse the sum of all 16 compounds, and a standardised list allows for global comparability. However, in the four decades since the original list was published, several more standards are now commercially available, not to mention that more toxicological research is available for other PAHs ²⁸. This suggests that the list should be reviewed and possibly updated to reflect the current state of the knowledge.

The first European Parliament Directive that proposed monitoring and maintaining the level of PAHs in ambient air below certain thresholds was published in 2004 ²⁹. The Council proposed that B[a]P

should be used as a marker of the carcinogenic risk of PAHs and set the target value for B[a]P levels at 1 ng/m³ for the total content in the PM10 fraction averaged over a whole year. Furthermore, recommendations were made to promote research into the effects of PAHs on human health and the environment, to standardise accurate measurement techniques to ensure comparability across multiple measuring sites, and to monitor the levels of a further 6 PAHs: B[a]A, B[b]Fl, benzo[j]fluoranthene (B[j]Fl), B[k]Fl, I[cd]P, and DB[a,h]A.

Very often, B[a]P is used as an indicator of PAH levels even at policy level^{6,29}, however many argue^{2,4,6} that this is not correct for several reasons. The proportion of B[a]P in various PAH containing mixtures is highly variable, and it has been well-established that the effects of individual PAHs are not necessarily additive when in mixtures^{4,30–33}. Additionally, the concentration of B[a]P in ambient air may be lower or higher than that of other PAHs and may underestimate the carcinogenic potential of total PAH exposure. For similar reasons, there are also criticisms of the use of lists of “priority PAHs” such as those suggested by the EPA. Additional arguments state that such lists are not exhaustive, do not include all compounds with high toxicity, can lead to underestimation of toxicity, and do not include other known toxic heterocyclic aromatic compounds or alkylated derivatives²⁸. A recent study that compared the 16 EPA PAHs to the results from 13 studies measuring 88 PAHs, found that the 16 EPA PAHs underestimated the carcinogenic potency of complex mixtures by 85.6% on average³⁴. The use of relative potency factors and toxic equivalency factors which estimate the toxicity of compounds relative to B[a]P are also inadequate for predicting the behaviour and toxicity of PAH mixtures³⁵. However a recent study in mice contradicts this³⁶. Given the ubiquitous nature of PAHs in the environment and the highly variable composition of mixtures, it is unlikely that a single list of compounds would satisfactorily cover all relevant study areas. Statistical methods have been used to attempt to address this problem. Principal components analysis (PCA), hierarchical clustering analysis (HCA), and neural networks have been used to define carcinogenic activity based on the relationship between chemical structure and activity³⁷. Models have been developed to predict PAH interactions

based on the integration of probability matrices from pathways known to be enriched following PAH exposure which allows for prediction based on mechanistic information ³⁵.

1.2.2 Routes of Exposure

The main routes of PAH exposure are diet, smoking, inhalation and, to a lesser extent, drinking water.

Dietary exposure has been reported to have the widest range and largest magnitude – 2 – 500 ng B[a]P/day, while inhalation only ranged between 10-50 ng/day ⁶. In non-smokers, diet is the major contributor to PAH exposure and often is one to two orders of magnitude higher than the other sources ^{38,39}. Table 1.1 shows the PAH exposure estimates from the four major sources of a “reference man” between the ages of 19 and 50 years old ²¹. The study also added that smoking 1 packet of cigarettes per day would add 1-5 µg to their daily exposure depending on the type of cigarettes smoked. Based on an assumed respiration rate of 20 m³/day, 0.16 µg of carcinogenic PAHs have been reported be inhaled daily ²¹. Assuming that the average person consumes approximately 2 L of water per day, the PAH exposure attributable to this is 0.006 µg/day ²¹. A study from Beijing, China found that 85% of low molecular weight PAH exposure came from the diet, while 57% of the high molecular weight PAH exposure was attributable to inhalation ¹⁸. Exposure from air inhalation and diet are described in detail in Chapters 4 and 5 respectively.

Table 1.1. Breakdown of PAH Exposure of a "Reference Man" ²¹

<u>Source</u>	<u>Proportion</u>
Food	96.2%
Air	1.6 %
Water	0.2 %
Soil	1.9 %
Total	3.12 mg/day

Since the 60's, the levels of B[a]P in cigarette smoke have decreased from 35 ng/cigarette to 10 ng/cigarette in "low tar" tobacco. PAHs in tobacco smoke do not only affect smokers but anyone exposed to the smoke. It has been reported that the levels of B[a]P in a room highly polluted with cigarette smoke were 22 ng/m³⁶. Mainstream tobacco smoke from unfiltered cigarettes may contain between 0.1 and 0.25 µg of carcinogenic PAHs per cigarette, while the side-stream smoke may cause indoor levels to be 3-29 ng/m³ over and above PAHs from any other sources²¹. Ten PAHs have been characterised in tobacco smoke from 9 different American cigarette brands⁴⁰. Of these, B[a]A had the highest levels ranging from 38.2 to 66.6 ng per cigarette, with the overall levels of all 10 PAHs also being highly variable (76.3-140.9 ng per cigarette). A more recent study of 50 American cigarette brands found that lower molecular weight compounds were more prevalent than high molecular weight PAHs and that the composition varied from one brand to another⁴¹.

The levels of B[a]P in drinking water are significantly lower than the levels from other sources and range between 0.0002 – 0.024 µg/l and the levels of 6 common PAHs do not exceed 0.1 µg/l in 90% of drinking water samples^{6,42}. The concentrations of carcinogenic PAHs in surface water in the USA have been reported to range from 0.0001-0.83 µg/l²¹ which are higher than groundwater and drinking water levels due to the presence of PAHs adsorbed on to particles suspended in surface water. These PAHs are naturally filtered out and adsorbed onto organic soil before and so only very low concentrations reach groundwater (0.0002-0.007 µg/l)²¹.

1.2.3 Absorption and Metabolism

Exposure to PAHs leads to their absorption from the lung, gut and skin due to their lipid-solubility¹¹. Absorption from the lungs is swift and diphasic, with an initial rapid phase with a half-life of 5 minutes followed by a slow phase which has a half-life of 116 minutes^{43,44}. Respiratory uptake can be significantly slower when B[a]P is adsorbed onto particles, since the particles remain in the respiratory tract for longer. Consequently, the second phase of absorption in these cases has been reported to have a half time of 18 days and under such circumstances the compounds may be metabolised in the

lungs resulting in increased metabolite-DNA adduct formation particularly in the case of particulate-adsorbed B[a]P^{44,45}. This seems to indicate some co-carcinogenic effect of PAHs and urban air particles, either due to increased pulmonary retention time or by affecting the resulting metabolite pattern which may promote metabolite-DNA binding^{44,45}. Bioavailability of PAHs inhaled as particulates or in aerosol form is higher than of those ingested through food²¹. PAHs are usually present in food as solutes making them readily absorbed from the gastrointestinal (GI) tract with the help of bile salts. In rats, 30-50 % of a low oral dose of B[a]P was absorbed with most of it metabolised efficiently in the liver⁴⁶.

Studies in rats and mice have also shown that B[a]P can cross the placenta, but levels in the embryo/foetuses have been reported to be one to two orders of magnitude lower than those in maternal organs^{47,48}. In mice, both embryo and foetal tissues have been reported to be able to metabolise B[a]P albeit at a lower rate than adult animals⁴⁸. Consecutive administration of B[a]P in these mice did not lead to bioaccumulation but rather accelerated elimination⁴⁸.

In mice and rats, the pattern of distribution of B[a]P has been shown to be the same irrespective of subcutaneous, intra-tracheal or intravenous methods of exposure^{42,46}. Post-exposure, the highest levels were present in the liver, as well as fatty tissues such as mammary tissue however no significant bioaccumulation takes place due to efficient metabolism. Measureable, albeit lower levels have been reported to be found in the intestines, kidneys and blood of rats administered B[a]P by intra-tracheal instillation⁴³. However other sources pose that due to the lipophilicity of PAHs, they accumulate in adipose tissues from which they are then slowly released³.

The primary PAH-metabolising tissues are the liver, lung, intestine, skin and kidneys, but many other tissues like white blood cells (WBCs) are also capable of breaking down these compounds¹¹. The estimated half-life of PAHs in WBCs is estimated to be 3-4 months⁴⁹. Once absorbed, PAHs are activated and metabolised into compounds that are highly genotoxic by Phase I enzymes such as cytochrome P450 enzymes and peroxidases, however there are multiple mechanisms by which these

metabolites induce and promote carcinogenesis which are discussed in detail in the following section. Increased expression of these enzymes is known to occur via AHR-mediated transcription. Phase II metabolising enzymes are responsible for the formation of polar PAH conjugates and include glutathione S-transferases, uridine 5'diphosphate glucuronosyltransferases and uridine 5'-diphosphate sulfotransferases ¹¹. These enzymes add sugars, sulphates or amino acids to the PAH-metabolites which increases their solubility enabling excretion ¹².

Removal of PAHs from the body is triphasic and occurs mainly through hepatobiliary excretion followed by faecal elimination, with higher metabolite concentrations in faeces after oral administration ^{46,50}. A lesser-used excretory route is urine but this usually only represents a small fraction of the administered dose ⁴³. Considering both routes, the maximum number of metabolites removed from the body occurs during the first and second days following oral exposure, and second and third days following intraperitoneal administration ⁵⁰ with the delay possibly attributable to increased uptake from the intestine to the blood resulting in faster metabolism. Higher levels of metabolites and mutagenic by-products have been reported in male rat urine and faeces compared to female rats ⁵⁰.

1.2.4 Carcinogenicity and Genotoxicity

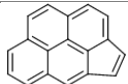
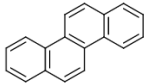
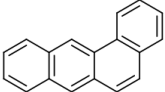
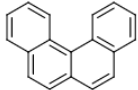
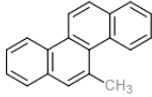
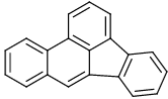
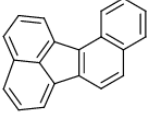
The first report of the association between PAH exposure and cancer dates back to 1775 when Sir Percival Pott noted a higher incidence of scrotal cancer in chimney sweeps which he attributed to exposure to soot ⁵¹. Varying levels of evidence as reviewed by the International Agency for Research on Cancer (IARC) have linked occupational PAH exposures to 9 different cancer types: lung cancer, renal cell carcinoma, urinary bladder cancer, laryngeal cancer, skin cancer, pancreatic cancer, stomach cancer, oesophageal cancer and prostate cancer in order of decreasing evidence ¹¹.

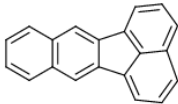
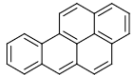
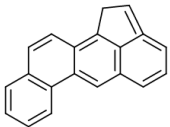
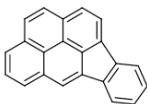
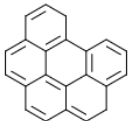
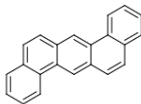
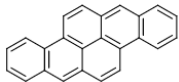
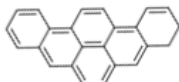
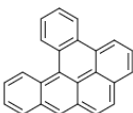
Interestingly, many of these tissues are sites of PAH absorption which would suggest the presence of higher levels of PAH-activating enzymes at these sites which might explain the carcinogenic activity of PAHs at their site of entry. IARC classified PAHs by their carcinogenicity based on the available literature in 2010. Only B[a]P is classified as a known human carcinogen (Group 1) and is the most

commonly researched PAH, with other common PAHs being classified as probable (Group 2A) or possible (Group 2B) human carcinogens as summarised in the most recent IARC monograph ¹¹. Table 1.2 shows the structure and classification of some of the most common PAHs.

The process of carcinogenicity as described by Hanahan and Weinberg (2000 and 2011) ^{52,53} is a multi-step progression which involves the deregulation of several processes at the genetic, epigenetic, protein, and cellular levels. Additionally, carcinogenesis is often broken down into the following steps: initiation, promotion and progression. PAHs are considered to be complete carcinogens because they are capable both of tumour initiation and promoting tumour progression ^{20,30}. The next sections outline the mechanisms by which PAHs induce carcinogenesis.

Table 1.2. Table of common PAHs, their structure, molecular weight, and carcinogenic classification in order of increasing molecular weight.

<i>Compound</i>	<i>Molecular Weight</i>	<i>IARC Classification</i>	<i>Structure</i>
<i>Cyclopenta[cd]pyrene</i>	226.29 g/mol	Probable carcinogen (Group 2A)	
<i>Chrysene</i>	228.29 g/mol	Possible carcinogen (Group 2B)	
<i>Benz[a]anthracene</i>	228.29 g/mol	Possible carcinogen (Group 2B)	
<i>Benzo[c]phenanthrene</i>	228.29 g/mol	Possible carcinogen (Group 2B)	
<i>5-methylchrysene</i>	242.31 g/mol	Possible carcinogen (Group 2B)	
<i>Benzo[b]fluoranthene</i>	252.31 g/mol	Possible carcinogen (Group 2B)	
<i>Benzo[j]fluoranthene</i>	252.31 g/mol	Possible carcinogen (Group 2B)	

<i>Benzo[k]fluoranthene</i>	252.31 g/mol	Possible carcinogen (Group 2B)	
<i>Benzo[a]pyrene</i>	252.31 g/mol	Carcinogen (Group 1)	
<i>Benzo[j]aceanthrylene</i>	252.33 g/mol	Possible carcinogen (Group 2B)	
<i>Indeno[1,2,3-cd]pyrene</i>	276.33 g/mol	Possible carcinogen (Group 2B)	
<i>Benzo[ghi]perylene</i>	276.34 g/mol	Not classifiable	
<i>Dibenz[a,h]anthracene</i>	278.35 g/mol	Probable carcinogen (Group 2A)	
<i>Dibenzo[a,h]pyrene</i>	302.38 g/mol	Possible carcinogen (Group 2B)	
<i>Dibenzo[a,i]pyrene</i>	302.38 g/mol	Possible carcinogen (Group 2B)	
<i>Dibenzo[a,l]pyrene</i>	302.38 g/mol	Probable carcinogen (Group 2A)	

1.2.4.1 Mechanisms of Carcinogenicity and Genotoxicity

Metabolic activation of PAHs is the initial step in their carcinogenicity²⁰. There are four mechanisms by which PAHs are metabolically activated: diol epoxide formation, one electron oxidation, meso-region biomethylation and benzylic oxidation, and ortho-quinone and reactive oxygen species (ROS) formation, with the former two being the predominant pathways. The ortho-quinone pathway leads to the formation of PAH-o-quinones which form stable DNA adducts that contribute to the formation of ROS by undergoing enzymatic and non-enzymatic redox cycles⁵⁴. A summary of each of these

mechanisms and their genotoxic outcomes can be found in Figure 1.1. While all four mechanisms result in the formation of DNA adducts, the reactive electrophilic PAH intermediates may also react with RNA and proteins.

The diol epoxide mechanism is the best characterised of the four. The aryl hydrocarbon receptor (AhR) is a ligand-activated transcription factor that is responsible for the cellular response to PAH exposure. In unexposed cells, the AhR is bound to several gene clusters associated with gene expression and differentiation, but this pattern changes upon ligand binding⁵⁵. Binding of PAHs to the AhR results in the activation of several immune and inflammation pathways⁵⁶. The signalling potency of the AhR plays an important role in determining the downstream consequences of PAH exposure⁵⁷.

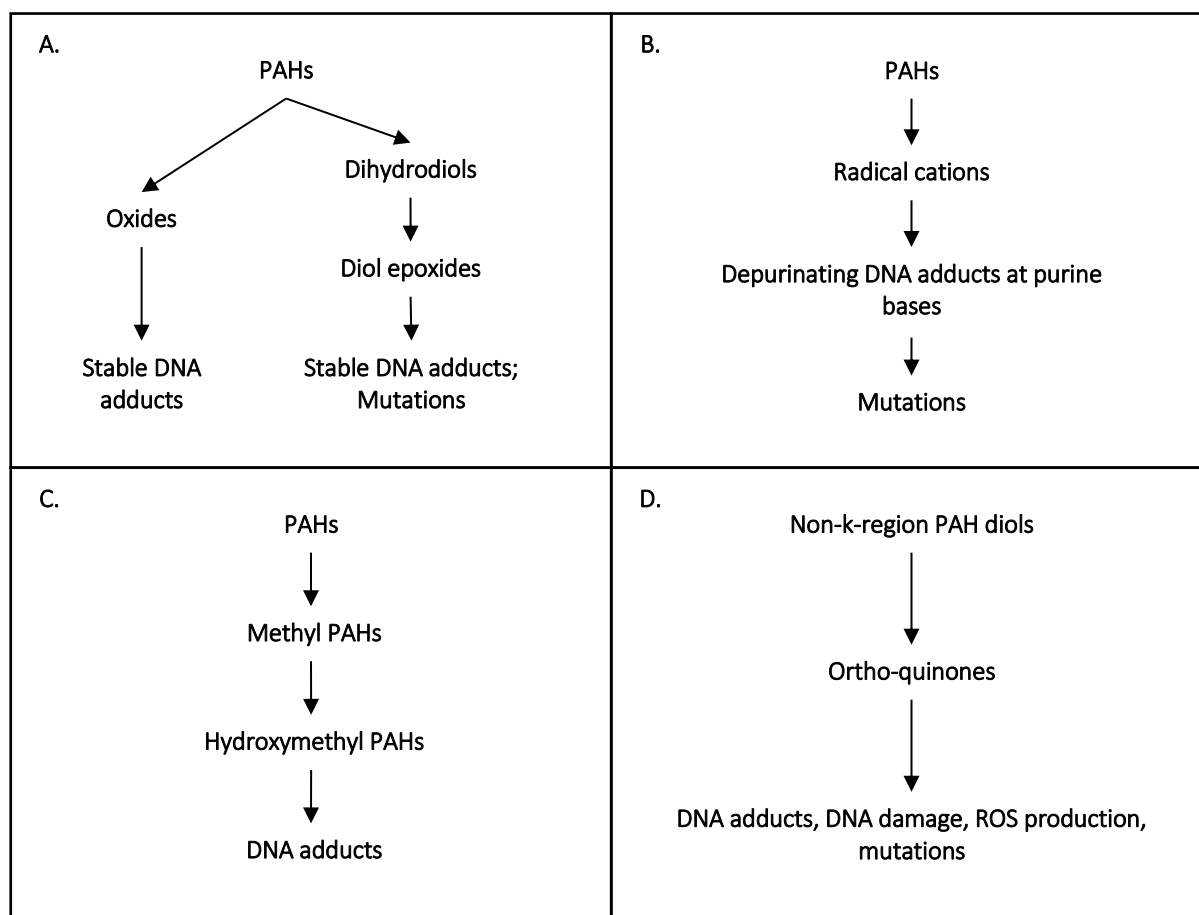


Figure 1.1. Overview of the four carcinogenic mechanisms of PAHs. A: The diol epoxide mechanism; B: One electron oxidation; C: Meso-region biomethylation and benzylic oxidation; D: Ortho-quinone/reactive oxygen species mechanism.

Children have been reported to be more susceptible to PAH exposures due to them being more at risk of activating the AhR pathway than adults in the same population⁵⁸. When the AhR binds PAHs once they enter the cell, the receptor then dimerises with the aryl hydrocarbon nuclear translocator (ARNT)^{22,59}. In the cytoplasm, the AhR is complexed with proteins such as p23 and heat shock protein 90 (Hsp90), and upon PAH binding, hepatitis B virus X-associated protein 2 is released from the complex^{9,12}. The complex then enters the nucleus where Hsp90 is released from the complex and it then binds to specific genes containing a xenobiotic response element (XRE)⁹. Two such genes are *CYP1A1* and *CYP1B1* which are part of the cytochrome P450 enzyme family involved in xenobiotic metabolism⁶⁰. These enzymes are substrate-inducible and function to convert inert PAHs into reactive electrophiles which can then be excreted from the cell³⁰. Once these enzymes are induced, their induction may last for long periods of time, even if PAH exposure is stopped⁹. The *CYP1A1* enhancer region is divided into region A and region B which contain one and two XREs respectively⁶¹. Following oxidation by *CYP1A1* or *CYP1B1*, the PAH-epoxide is formed which is then cleaved by epoxide hydroxylase to form the dihydrodiol^{60,62}. Further activation by *CYP1A1* or *CYP1B1* results in the formation of the diol-epoxide which is known to react with DNA to form covalent adducts^{30,60,62}. These reactions are summarised in Figure 1.2 using B[a]P as an example, but Chr, 5-methylchrysene, B[a]A, and DB[a,l]P are also metabolised in this way. There are multiple diol epoxide metabolites due to the asymmetry of PAH molecules^{30,63}. Additionally, these metabolites can have both diastereomeric and enantiomeric stereoisomers which have different biological behaviours due to their different chemistries⁶³. Taking B[a]P as an example, following epoxide metabolism either the 7,8-dihydrodiol-9,10-epoxide or the 9,10-dihydrodiol-7,8-epoxide geometric isomers can be formed⁶³. The diastereomeric isomers could be *syn*- and *anti*- variants of both of the above, and the enantiomers could be (+)- or (-)- of each geometric and diastereometric isomer⁶³. These metabolites covalently bind in a *cis* or *trans* manner to the exocyclic amino group of purine bases, however guanine tends to be the preferred base, particularly when located adjacent to a methylated cytosine^{30,63-65}. DNA adducts form preferentially at the N2 position of guanine (dG-N²-BPDE) and at the N6 of

adenine (dA-N⁶-BPDE) ¹². While many adducts are repaired, those that are not may lead to substitution or deletion mutations. PAH metabolites also form adducts with proteins, including human serum albumin (HSA) and haemoglobin (Hb). These adducts can accumulate in the blood until they are degraded when the proteins themselves are degraded ⁶⁶.

The radical cation mechanism involves the P450 peroxidase enzyme which catalyses the one electron oxidation of the PAH during which one electron from the π electron system of the molecule is lost ¹³.

The resultant radical cation PAH forms a few stable DNA adducts, but also many unstable DNA adducts which lead to depurination and therefore the formation of apurinic sites which are mutagenic and the most common type of DNA damage. These unstable adducts are formed through the binding of the radical cation of the N7 or C8 of purine bases ¹².

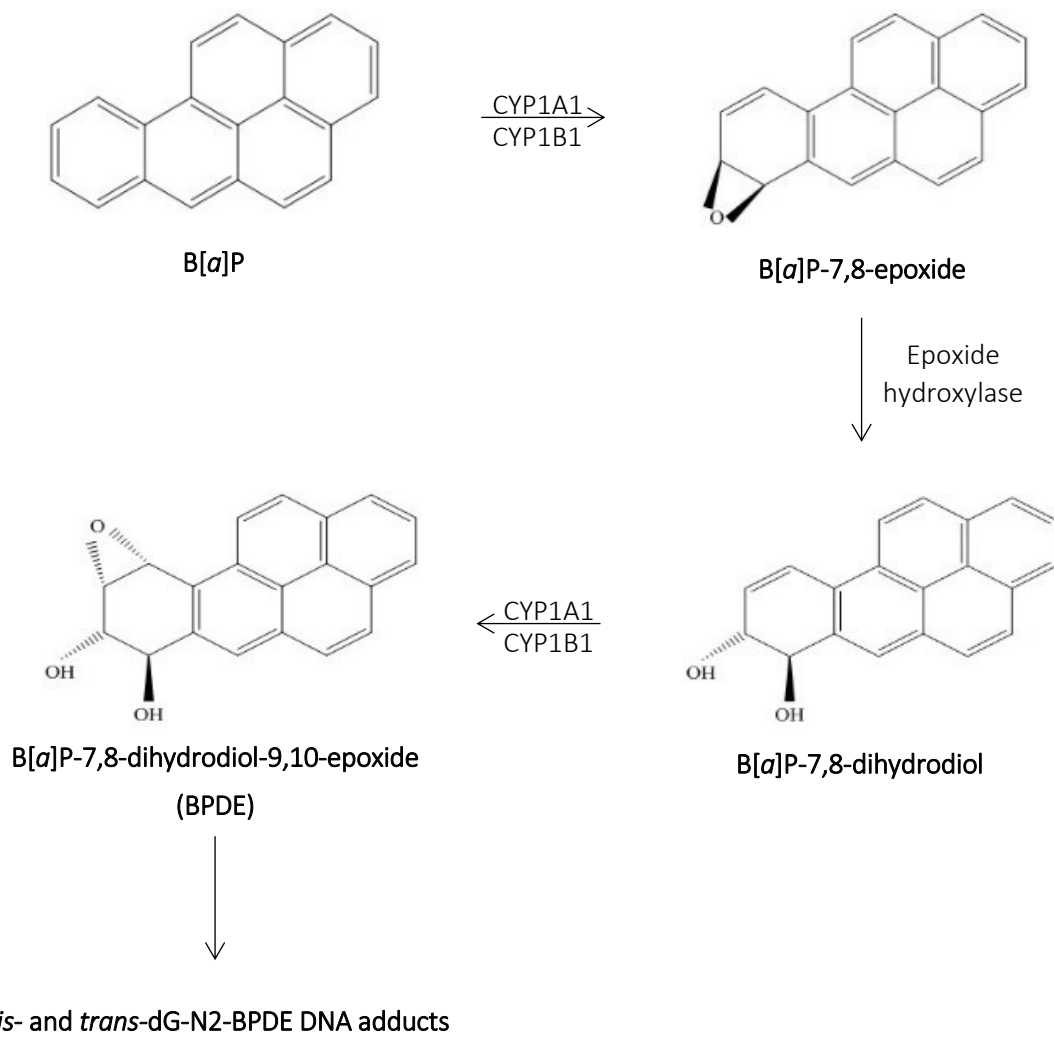


Figure 1.2. Metabolism of B[a]P via the diol epoxide mechanism resulting in the formation of DNA adducts

The ortho-quinone pathway results in the formation of o-quinones which also cause the formation of both stable and unstable DNA adducts as well as reactive oxygen species (ROS). Dihydrodiol dehydrogenase enzymes compete with P450 enzymes to oxidise the PAH into a ketol which spontaneously rearranges itself into a catechol. Following autoxidation, the o-quinone is formed which may bind to DNA to form adducts or enter a redox cycle where it is reduced back into the catechol causing the formation of ROS. High levels of ROS may lead to the formation of over 100 genomic oxidative base lesions as well as 2-deoxyribose modifications⁶⁴. One of the major lesions formed by ROS is the 7,8-dihydro-8-oxoguanine (8-oxo-G) lesion which pairs with adenine instead of cytosine thereby increasing mutational load⁶⁴. Base excision repair (BER) is the mechanism by which depurinated, deaminated, alkylated and oxidative lesions are corrected^{12,64,67}. However, ROS can induce strand breaks at the DNA backbone which are repaired either by the single strand break pathway or the double strand break pathway⁶⁴.

A study carried out in rats exposed to B[a]P showed that by inhibiting the cytochrome P450 enzymes, no ROS were formed⁶⁸ which suggests that all mechanisms are dependent on the presence of these enzymes, even if they do not actually play a direct role. Other enzymes involved in the metabolism and detoxification of PAHs are: CYP1A2, CYP2C9, CYP3A4, UDP-glucuronosyltransferase, glutathione transferase, epoxide hydrolase, methyl transferase, sulfotransferase, NADPH quinone oxidoreductase, and aldo-keto reductase^{30,60}. Exposure of lung epithelial cells to PAHs from air samples collected during haze (168-307 ng PAHs/m³) and non-haze (42-87 ng PAHs/m³) events showed that chronic exposure to lower PAH levels resulted in the preferential formation of ROS as a consequence of increased expression *AKR1C2* gene expression, whereas cells exposed to a single acute exposure of haze PAH samples exhibited increased *EPHX1* expression and formed increased diol epoxides⁵⁴. B[a]P exposure has been reported to induce other genotoxic effects like sister chromatid exchange, formation of micronuclei and formation of 8-oxo-G adducts¹⁰. Mutations in oncogenes and tumour suppressor genes such as *K-RAS* and *p53* have been well-documented both in human and mouse tissues^{10,30}.

Exposure to B[a]P results in G-to-T transversion mutations, for example in *p53* and *k-ras* genes⁶⁹. Over 70% of mutations induced by B[a]P are C:G>A:T transversions⁷⁰. In lung cancer and hepatocellular carcinomas, G-to-T transversions have been shown to occur more frequently in the *p53* gene⁷¹. Similar mutation patterns between smokers and non-smokers were reported by the authors. Codon 273 of the human *p53* gene has been termed a mutational hotspot following B[a]P exposure, particularly for G-to-T transversions⁷². The (+)-*trans*-dG-*N*²-BPDE adduct was shown to have a higher mutational frequency than the (-)-*trans*-dG-*N*²-BPDE adduct and that the mutational frequencies induced by both adducts were reduced when the cytosine 5' to dG-*N*²-BPDE was replaced with a methylated cytosine⁷². Exposure to DB[a,l]P resulted in tumour formation and mutation patterns that matched that of DNA adducts⁷³. Similarly to *p53*, the predominant mutations in the *k-ras* gene were G-to-T transversions but A-to-G transitions and A-to-T transversions were also reported⁷³. Increased numbers of DNA adducts are associated with increased mutation frequency in codons 12 and 14 of the *k-ras* genes⁷⁴. This preferential damage at codon 14 was shown by the authors to be due to the methylation status of the cytosine, where presence of a methylated cytosine resulted in increased adduct formation⁷⁴. The mutation spectrum of B[a]P shows a number of cancer driver mutations occurring at genes related to cell cycle regulation, regulation of cell death and proliferation, chromatin modification and DNA repair, all of which contribute towards the development of immortal cells⁷⁰.

At the inter-individual level, gene-exposure and gene-gene interactions play an important role in determining the extent of damage from genotoxic agents⁷⁵. *CYP1A1* expression levels have been linked to DNA adduct levels where it has been reported that smoking females have increased *CYP1A1* expression and consequently higher DNA adduct levels than their male counterparts⁷⁶. Additionally, polymorphisms at phase I or phase II enzymes have been shown to result in increased DNA adduct levels in humans exposed to urban air pollution, bitumen or environmental tobacco smoke^{75,77}. Specifically polymorphisms at the genes coding for *CYP1A1*, *mEH*, *NAT2*, and *GSTP1* enzymes may have an effect on the metabolism and excretion of PAHs^{60,77-79}. Additionally, polymorphisms in

19q13.3 genes have been associated with lower levels of BPDE-DNA adducts and which suggests altered repair efficiency of these adducts ⁸⁰.

Tumorigenicity, however, cannot be predicted from DNA adduct formation alone. Exposure of mice to a PAH mixture extracted from coal tar resulted in relatively low levels of DNA adducts however, the mixture had high tumorigenic effects ⁸¹. It has been shown that the formation of dG-N²-BPDE adducts causes an increase in *p53* activity ⁸² which is not surprising given that DNA damage activates *p53* and that the protein plays an essential role in the repair of adducts ⁸³. ROS may cause oxidative damage through the formation of 8-oxo-G lesions which may trigger tumour initiation. Also, ROS affect a number of cellular processes required for tumour development such as cell proliferation, inflammation and cell cycle regulation through proteins such as *p53* ⁸⁴. Regulatory T cells (Treg cells) exposed to the PAH phenanthrene had reduced *FOXP3* expression which subsequently caused impaired regulatory function and conversion of the cells to a *CD4⁺CD25^{hi}CD127^{lo}* phenotype clearly indicating immune modulation by a PAH ⁸⁵ which may contribute to the carcinogenicity of these compounds or the pathogenesis of other PAH-related diseases. In humans, urinary PAH metabolites have been shown to be significantly associated with malondialdehyde which is a product of lipid oxidation damage ⁸⁶. These results indicate that PAH exposure may result in metabolism alterations which may play a role in the downstream health consequences of PAH exposure ⁸⁶. Processes related to inflammation have been reported to modulate B[a]P-induced carcinogenesis by resulting in higher concentrations of DNA-reactive metabolites and reducing DNA repair. These mechanisms have been reviewed by Shi *et al.* (2017) ⁸⁷.

Nucleotide excision repair (NER) is the major response to the DNA damage induced by PAHs ^{19,64,67}. DB[*a,l*]P dA-N⁶ DNA adducts are significantly less susceptible to NER dual incisions when compared to the stereochemically identical dG-N² adduct ⁸⁸. Recombinant repair and transcription-coupled repair (TCR) have also been reported to be involved ¹². Global genomic nuclear excision repair (GG-NER) which corrects damage in transcriptionally silent parts of the genome, and transcription-coupled

nuclear excision repair (TC-NER) which occurs in actively transcribed DNA strands are the main sub-pathways for the removal of bulky DNA adducts. GG-NER is highly dependent on p53 status for repair of BPDE-DNA adducts⁸³. DNA adducts may block DNA repair activity by blocking polymerase replication activity¹². Hepatocytes derived from DNA repair deficient mice (*Xpa*^{-/-}*p53*^{+/-}) were more sensitive to B[a]P exposure compared to cells derived from wild-type mice⁸⁹. In the latter, DNA repair and cell cycle control genes were activated, whereas in the *Xpa* deficient cells mitogen-activated protein kinase signalling was found to be deregulated which may have been responsible for the observed down regulation of cancer-related pathways⁸⁹. Furthermore, SNPs in *XPA* and *XPC* genes which play key roles in NER may modulate the levels of DNA damage caused by PAH exposure⁹⁰.

A reported non-genotoxic effect of PAH exposure which may contribute to their carcinogenicity is that Gap junction intercellular communication is inhibited independently of PAH metabolism which may be a consequence of membrane damage or interaction with membrane components^{3,30,32,91-94}. A recent study has shown that while B[a]P exposure alone inhibits Gap junctional intercellular communications, subsequent treatment with a mixture of low molecular weight PAHs results in further inhibition³². The level of inhibition varies for different individual PAHs⁹⁴. PAHs with particular structural features called bay and fjord regions tend to cause higher levels of inhibition. In most literature, a bay region is defined as one shown in Figure 1.3A where B[a]P is used as an example. More recently, a publication has defined these as cove regions⁹⁵. Similarly, fjord regions as shown in Figure 1.3B with the example of DB[a,l]P were re-defined to be bay regions⁹⁵. In accordance with the most recent definition, fjord regions are those which are enclosed on 5 sides, however the majority of publications, even those published after the paper by Ehrenhauser (2015)⁹⁵ use the older nomenclature where a fjord region is closed on 4 sides as in Figure 1.3B. Presence of these regions has been shown to be associated with increased carcinogenic potential along with other features such as the compound's molecular weight, number of benzoid rings, and affinity for the AhR^{22,30,92}.

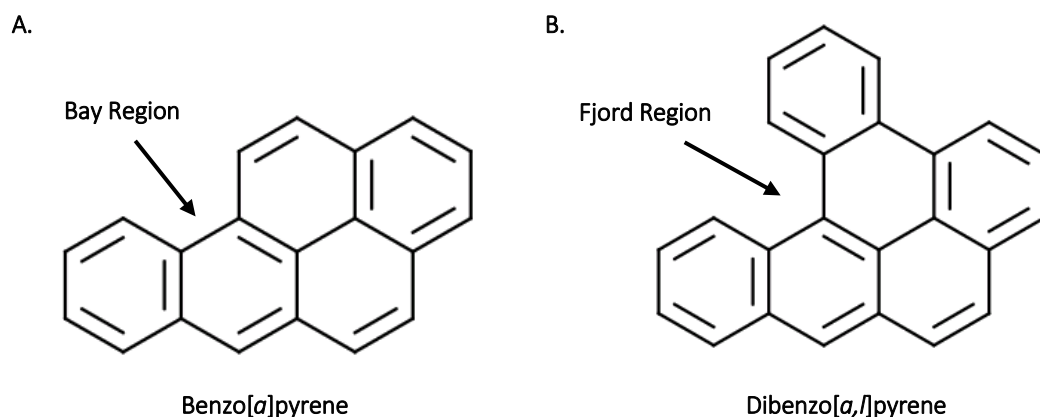


Figure 1.3. A: Chemical structure of B[a]P as an example of a PAH with a bay region (or cove region based on a more recent definition). B: Chemical structure of DB[a,h]P as an example of a PAH with a fjord region.

1.2.4.2 In Vitro Evidence

A number of *in vitro* studies have been carried out to assess the effects of PAH exposure, often represented by B[a]P, on a variety of cell lines. Broadly, the effects of exposure in such studies are DNA damage, the formation of DNA adducts, altered gene expression, aberrant cell cycle regulation, and cytotoxicity.

1.2.4.2.1 DNA Damage and Formation of DNA-adducts

The number of DNA strand breaks measured using the comet assay in hepatocytes exposed to B[a]P, increased linearly with increasing pure B[a]P concentrations with no breaks observed at concentrations lower than 1 μM ¹⁹. A similar trend was observed for DNA adduct levels which exhibited a dose-dependent increase, with measurable adduct levels at 0.025 μM B[a]P and a plateauing trend as concentrations reached 20 μM ¹⁹. Additionally, adduct levels showed a sigmoid time response with a slow initial increase for 6 hours, followed by a rapid formation period for the next 4 hours and a plateau by 16 hours post-exposure¹⁹. Hepatocytes exposed to pure B[a]P, and B[a]P in industrial, urban and atmospheric air samples with the same concentration of B[a]P showed that pure B[a]P does not seem to exhibit a genotoxic response, but industrial and urban samples induce a 140% and 300% increase in DNA strand breaks respectively¹⁹. Formation of BPDE adducts

was also higher in the air samples: 30%, 90% and 600% in the industrial, urban and atmospheric samples respectively ¹⁹. This suggests that the total PAH composition of the samples affects genotoxic outcomes. This is supported by a study in which hepatocytes were exposed to B[a]P alone, a mixture of B[a]P and DB[a,l]P, and an air sample. The mixture showed a more-than-additive genotoxic response, with increased phosphorylation of *CHK1*, *p53*, and *H2AX* genes, with the air samples showing persistent activation of *CHK1* at B[a]P equivalent concentrations that were much lower than those of B[a]P alone that elicited the same activation ⁴.

DNA damage signalling has been shown to be increased in hepatocytes exposed to air PM fractions containing PAHs with a higher molecular weight and increased benzene rings when compared to exposure to smaller compounds ⁴. Chronic PAH exposure using non-haze air samples (42-87 ng/m³), followed by an acute haze sample exposure (168-307 ng/m³) has been shown to cause increased levels of DNA damage and subsequent genomic instability when compared to the high-level exposure on its own in lung epithelial cell lines ⁵⁴. The authors suggested that chronic exposure, even at lower levels, results in saturation of the PAH-metabolising pathways, consequently increasing ROS levels and the associated stress lead to DNA damage and genomic instability which may in turn result in alterations of the DNA damage response networks.

1.2.4.2.2 Gene Expression Changes

Breast and liver cell lines exposed to B[a]P showed time- and concentration-dependent gene expression changes, with the differentially expressed genes being associated with xenobiotic metabolism, cell cycle regulation, apoptotic and anti-apoptotic pathways, chromatin assembly, and the oxidative stress response ⁹⁶. In a subsequent study, the same authors exposed the cell lines to B[a]P, BPDE, or 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) ⁹⁷. TCDD is known to induce xenobiotic metabolism by AhR activation but does not form DNA-reactive metabolites, BPDE does not induce AhR activation but reacts with DNA to form adducts, and B[a]P does both. The authors found that all three exposures induced gene expression changes, TCDD and B[a]P induced changes related to

xenobiotic metabolism, BPDE and B[a]P repressed histone genes which the authors hypothesise may be important with respect to the DNA damage response⁹⁷. Finally, the authors observed an overlap of differentially expressed genes between TCDD and BPDE exposure, and that B[a]P exposure induced some gene expression changes that were not observed in either TCDD or BPDE exposures⁹⁷. This latter observation suggests that B[a]P exposure may affect signalling pathways independent of the activation of AhR or the DNA damage response. Liver cells exposed to different B[a]P metabolites showed differences between the gene networks activated by early and late metabolites⁹⁸. B[a]P-9,10-diol activated networks associated with B[a]P metabolism, cell proliferation and oxidative stress, while BPDE activated genes related to DNA damage repair, apoptosis and cell energetics⁹⁸. Both the mRNA and microRNA profiles of liver cells exposed to B[a]P showed time-dependent effects of exposure⁹⁹. Several of the altered mRNAs were targets of the differentially expressed microRNAs, and several of these microRNAs are involved in many of the pathways mentioned above related to B[a]P genotoxicity⁹⁹.

The transcriptome is also greatly modified in macrophages exposed to B[a]P where 1100 genes have been reported to be differentially expressed after 24 hours of exposure¹⁰⁰. As previously reported, these genes are predominantly within the AhR and p53 signalling pathways. Lung adenocarcinoma progenitor cells exposed to B[a]P or a mixture of low molecular weight PAHs showed increased expression of *Cox-2*, a known inflammation marker, and it was found that expression levels were increased further after combined treatment of B[a]P and the mixture³².

Liver and lung cell lines exposed to soil collected from an area around a coke oven factory were found to have several differentially expressed genes related to functions already discussed above: metabolism, DNA damage repair, oxidative stress and cellular proliferation¹⁰¹. Interestingly, the liver cells had twice as many differentially expressed genes compared to lung cells following exposure, which suggests tissue specific responses¹⁰¹.

1.2.4.2.3 Cell Cycle Arrest

MCF7 and T47-D breast cancer cell lines exposed to varying doses of B[a]P showed dose-dependent increases in p53-mediated cell death as well as cell cycle arrest at S and G2/M phases respectively¹⁰². Interestingly, two triple negative breast cancer cell lines in the same study were not responsive to B[a]P, even at high doses. Two human lung epithelial cell lines have shown cell cycle arrest at S phase following chronic exposure to PAHs from PM2.5 air samples during haze and non-haze events⁵⁴.

1.2.4.2.4 Cell Migration, Viability and Cytotoxicity

Lung epithelial cells chronically exposed to air samples from haze events showed decreased viability, atypical nuclei and cytoplasm, increased necrosis, as well as a lower IC50 on subsequent acute exposure to the haze sample⁵⁴. The combination of dysregulated xenobiotic metabolism and DNA damage responses may be responsible for the increased susceptibility of chronically exposed cells to further PAH exposures.

A549 cells (human lung carcinoma cell line) exposed to roadside PAH samples showed cytotoxic responses dependent on the PAH composition of the sample, more specifically the concentration of high molecular weight compounds⁵. The study found that samples with increased high molecular weight PAHs exhibited an increased cytotoxic response when compared to samples with lower amounts of high molecular weight compounds. The authors postulated that the low molecular weight compounds caused synergistic promotion of this cytotoxic response. MCF-7 breast cancer cells exposed to > 1 µM of B[a]P for 96 hours showed a large decrease in cell viability⁹⁶. A further three breast cancer cell lines exposed to B[a]P demonstrated activation of ERK-MMP9 signalling as a consequence of increased levels of ROS and this resulted in increased cell migration¹⁰³.

1.2.4.3 In Vivo Evidence

Fewer *in vivo* studies assessing the effects of PAH exposure have been carried out compared to *in vitro* studies and some of these are described below. Liver tissue from mice exposed to DB[a,h]A showed a dose-dependent increase in the number of differentially expressed genes following

exposure with many genes associated with cancer, circadian rhythm, cell cycle, apoptosis, and immune response¹⁰⁴. When compared to results from liver tissue of mice exposed to B[a]P, the responses were found to be distinct for each exposure, with DB[a,h]A inducing a greater number of differentially expressed genes¹⁰⁴. DB[a,h]A was also found to have increased potency on the AhR compared to B[a]P¹⁰⁴. Another study looking at the gene expression profiles of liver tissue from B[a]P exposed mice also showed altered profiles in genes related to xenobiotic metabolism, immune responses and other downstream p53 targets¹⁰⁵. In mice exposed to B[a]P, BPDE-DNA adduct formation was reported both in target organs, such as the lung and forestomach, and non-target organs. In these same mice, B[a]P exposure induced tissue-specific gene expression profiles, with only two target organs, the lung and spleen, showing profile similarities¹⁰⁶. Rats exposed to four different PAHs showed unique gene expression signatures for each compound at short time-points¹⁰⁷. Additionally, the expression profiles of these rats were able to correctly and accurately predict exposure¹⁰⁷. Murine lung tissue exposed to exhaust emissions from gasoline direct injection engines exhibited up-regulation of *Cyp1A1* and *Cyp1B1* which are heavily involved in the diol epoxide mechanism of PAH metabolism¹⁰⁸. *Hmox1* was also up-regulated and this gene is considered to be a marker of oxidative stress¹⁰⁸.

1.2.4.4 Epidemiological Evidence

A number of epidemiological studies relating PAH exposure to known downstream carcinogenic outcomes such as DNA adduct formation have been published. Additional studies have also directly linked PAH exposure to breast, lung and upper gastrointestinal cancers. The paragraphs below summarise the findings from these studies which have measured PAH exposure from air pollution and environmental sources, from tobacco smoke, from dietary sources, from occupational sources, and *in utero* exposures.

Single nucleotide polymorphisms (SNPs) have been identified in phase I, phase II, aldo-keto reductase and NADP oxidoreductase enzymes which affect susceptibility to PAHs and cancer^{9,11,109}. In

oesophageal high-grade squamous dysplasia patients, *AhR* expression was found to be more than 9-times higher in those patients with a family history of upper gastrointestinal cancer ¹¹⁰. In a population based case-control study carried out in Long Island, several PAH exposure sources were found to be associated with increased incidence of breast cancer, with combined indoor sources such as smoking, exposure to tobacco smoke, intake of PAH-rich foods, and the use of stoves and fireplaces increasing incidence by 45% ¹¹¹. A subsequent study by the same authors reported that the associations between vehicular traffic and breast cancer risk were increased in women with polymorphisms in the DNA damage repair genes *ERCC2*, *XRCC1* and *OGG1* ¹¹². A recent study in postmenopausal women found that the association between PAH-DNA adducts and postmenopausal breast cancer was modified by BMI, with the incidence in overweight and obese women being raised ¹¹³. When comparing breast cancer incidence between rural and metro areas in the USA with respect to PAH emissions, an increased incidence of breast cancer was observed in women from the metro area ¹¹⁴. A small to moderately increased risk in colorectal adenoma and a moderately increased risk for pancreatic cancer have been reported from dietary B[a]P exposure as estimated from meat intake and cooking methods ¹¹. No association was found between dietary B[a]P intake and colon cancer, prostate cancer, or non-Hodgkin lymphoma. Dietary intake of B[a]P has been shown to be associated with colorectal adenoma in a case-control study of 146 cases and 228 controls ¹¹⁵. Another study reported that every 10 ng of B[a]P consumed per day corresponded to a 6% increased risk of large colorectal adenoma ¹¹⁶. However other studies have reported a null association between dietary B[a]P intake and colorectal cancer ^{117,118}.

A comparison of occupational exposure to PAHs between a few Central and Eastern European (CEE) countries and the UK found that occupational PAH exposure did not lead to an increased burden of lung cancer in the CEE countries, and suggested that the opposite finding in the UK cohort may be a result of higher exposure levels or a cooperative effect between PAHs and asbestos ¹¹⁹. The number of DNA strand breaks has been shown to be associated with dermal exposure to PAHs in trainee firefighters ¹²⁰. A pooled analysis of DNA adducts in 3600 subjects found that adduct levels tended to

follow the same seasonality trends as PAH concentrations in ambient air, and that subjects in Northern Europe have lower adduct levels than those from Southern Europe ¹²¹. Despite these observations, the authors noted high inter-individual variation which was only partly explained by seasonality ¹²¹. Polish coke oven workers with PAH exposure levels above the median had higher mitochondrial DNA copy number, as well as higher levels of BPDE-DNA adducts, increased numbers of micronuclei and shorter telomere length ^{122,123}.

A study of pregnant women from four different populations found that, as expected, the women having the highest ambient exposure to PAHs had the highest levels of BPDE-DNA adducts ¹²⁴. The authors also reported that the levels of BPDE-DNA adducts in cord blood were comparable to that of the mother despite the foetuses being exposed to an approximately ten-fold lower dose ¹²⁴.

While all these studies provide further evidence of the genotoxicity of PAHs in humans and have directly linked PAH exposure to cancer incidence, the body of epidemiological evidence linking exposure to the biological mechanisms like epigenetics and gene expression which in turn lead to tumorigenic outcomes is limited. Additionally, further studies are required to confirm the findings presented above since heterogeneous PAH exposure sources were used, with occupational PAH exposures being much higher than ambient and dietary exposures.

1.2.5 Other Consequences of PAH Exposure

While PAHs are predominantly linked to carcinogenic outcomes, the body of evidence linking PAH exposure to other diseases and adverse outcomes is increasing. Acute exposures to PAHs have been reported to induce skin irritation and inflammation, while chronic exposures may lead to the formation of cataracts, damage to the liver and kidneys, breathing problems, and abnormal lung function ³. Prenatal exposures have been associated with a number of adverse outcomes including reduced birth weight ¹²⁵.

Various cardiovascular outcomes have been associated with PAH exposure. Occupationally-exposed coke oven workers with a *miR-146a rs2910164* CC genotype were reported to be more likely to have

decreased heart rate variability¹²⁶. Mice with differential levels of AhR signalling exposed to B[a]P exhibited altered growth rates of both the body and organs, atherosclerosis, and changes in the gene expression profiles of their aortas⁵⁷. Additionally, the effects of PAH exposure on lymphocytes and associated immunological consequences have been studied and reviewed¹²⁷. Activation of AhR signalling as well as the formation of PAH-metabolites results in the dysregulation of Ca²⁺ levels in both B and T lymphocytes. This in turn alters antigen and mitogen receptor signalling pathways leading to suppressed humoral and cell-mediated immunity. Interestingly, high levels of PAH exposure may activate apoptotic pathways but conversely, low exposures may boost immune responses¹²⁷. Positive associations between urinary PAH metabolite levels and both high blood pressure and atherosclerotic cardiovascular disease have been reported and these may be partially mediated by obesity¹²⁸. Mortality from fatal ischaemic heart disease has been reported to be positively associated in European asphalt workers¹²⁹.

Exposure to PAHs has been associated with worsened asthma symptoms in children, possibly mediated by methylation changes at the *FOXP3* gene²³. Urinary metabolites of PAHs have been reported to be inversely associated with forced expiration volume in the elderly¹³⁰. In 2730 subjects with reduced lung function, PAH exposure measured using urinary PAH metabolites was found to be associated with diabetes¹³¹. Occupational exposures to PAHs have also been linked to increased mortality from obstructive lung diseases, however this study did not account for previous occupational history and smoking status which may have confounded this observation¹³².

PAH exposure, as measured from urinary metabolites, has been shown to be negatively associated with the sex chromosome ratio (Y:X ratio) in the sperm of 197 Polish men, suggesting that PAH exposure may play a role in the recently observed decline in the proportion of male births¹³³. Mouse spermatocyte-derived cells exposed to BPDE exhibited dose-dependent inhibition of cell viability, and both senescence and apoptosis were induced¹³⁴. Additionally, these cells had shorted telomeres, exhibited DNA damage associated with telomeres and decreased telomere activity, all of which the

authors suggested were mediated by *TERT* expression¹³⁴. These mechanisms may explain the mechanism by which PAH exposure reduces male fertility. The effects of PAH exposure on female reproductive health have recently been the topic of 9a scoping review, with the authors concluding that enough evidence exists to carry out systematic reviews about whether PAH exposure has adverse effects on female fertility and pregnancy viability¹³⁵. Zebrafish eggs exposed to B[a]P until 3.3 or 96 hours post-fertilisation showed a dose-dependent increase in the number differentially expressed genes and differential exon usage genes¹³⁶. These genes were found to be associated with many disease pathways such as growth failure, congenital heart disease and abnormal morphology of embryonic tissue, which indicate that early-life exposures may have long-term adverse effects¹³⁶.

Early-life B[a]P exposure in zebrafish has been shown to reduce locomotor and cognitive ability, neurotransmitter levels, loss of dopaminergic neurons, neurodegeneration, and increased levels of amyloid β protein¹³⁷. A separate study on zebrafish reported that the neuro-behavioural deficiencies caused by exposure to B[a]P during development are inherited transgenerationally¹³⁸. Rats exposed to B[a]P showed behavioural changes and neurotransmitter receptor genes were found to be differentially expressed¹³⁹. Rats exposed to B[a]P has increased levels of neuronal damage in their hippocampi¹⁴⁰. This same study also reported protein expression signatures associated with spatial learning and memory deficits, and identified *RARB* and *BDNF* genes as potential biomarkers of this¹⁴⁰.

PAHs have also been reported to be neurotoxic, and this may be caused by PAHs inducing neuronal cell death or dysregulating the expression of the *N*-methyl-D-aspartate receptor (NMDAR). A study of gene expression patterns in the hippocampi of mice exposed to B[a]P showed increased expression of the *Grin1* and *Grin2a* NMDAR subunits¹⁴¹. Interestingly, no changes were observed in DNA damage response genes, even though this has previously been reported in mouse lung and liver tissues^{141–143}.

Prenatal PAH exposure has been shown to be associated with a lower mental development index in three-year olds, and children with high prenatal PAH exposure have significantly higher odds of cognitive developmental delays¹⁴⁴. In a cohort of 5-year old children, prenatal airborne PAH exposure

levels above the median (17.96 ng/m³) was associated with decreased nonverbal reasoning ability as measured using the Raven Coloured Progressive Matrices which corresponded to an estimated average decrease in IQ of 3.8 points¹⁴⁵. Children prenatally exposed to more than 2.26 ng/m³ of airborne PAHs scored 4.31 and 4.67 points lower in full-scale and verbal IQ scores respectively at the age of 5¹⁴⁶. Associations between prenatal PAH exposure and ADHD symptoms in children have been reported and these are particularly strong in children facing material hardship¹⁴⁷. Additionally, prenatal exposure to PAHs has been shown to diminish the self-regulatory capacity of children in early and middle childhood which in turn has consequences on social competence¹⁴⁸.

Taken together, the evidence discussed above strongly suggests that diseases other than cancer also need to be considered when assessing the negative impacts of PAH exposure. Additionally, while the carcinogenic mechanisms of PAH exposure are reasonably well understood, the biological processes underlying the initiation and progression of other diseases need to be investigated.

1.3 Epigenetics

The term epigenetics refers to changes within the genome relating to gene expression and chromatin structure which take place without an alteration to the underlying DNA sequence. These heritable, reversible changes may take place in response to exposure to our environment such as the exposure to PAHs. The most commonly studied epigenetic mechanisms are DNA methylation and histone modifications. Collectively these mechanisms work to regulate gene expression, transcription, genome accessibility, chromatin state, overall DNA integrity, and maintain normal higher-order nuclear organisation. Several reviews comprehensively describe these epigenetic mechanisms in normal and disease phenotypes^{149,150,159,160,151–158} as well as in the context of environmental exposures^{161–165}. The following paragraphs aim to summarise the most common mechanisms.

1.3.1 DNA Methylation

DNA methyltransferases (DNMT) and demethylases control normal gene expression by regulating DNA methylation. There are three DNMTs: DNMT1 which is responsible for the maintenance of

methylation, and DNMT3a and DNMT3b which are responsible for embryonic *de novo* methylation. The most common sites of DNA methylation are cytosine residues found adjacent to guanine nucleotides (CpG sites), and a methyl group is added to the 5' position of the cytosine ring resulting in a methylated cytosine (5-mC). The methylation status of these sites may be associated with gene expression depending on their genomic location. Methylated cytosine residues have been termed the fifth base, with a sixth base also being described¹⁶⁰. This sixth base is 5-hydroxymethyl cytosine (5-hmC), and this is highly expressed in brain and bone marrow tissues, as well as embryonic cells¹⁶⁴. Conversion of methylated cytosines to hydroxymethylated cytosines had been attributed to ten-eleven translocation (TET) proteins¹⁶². The presence of 5-hmC is also thought to interfere with the ability of methyl-binding proteins to bind DNA and can result in demethylation. Additionally, 5-hmC is considered to be an active mark with functions relating to regulation of transcription and chromatin remodelling¹⁶⁰.

The distribution of CpG sites is uneven throughout the genome and they are usually found in clusters called CpG islands defined as portions of DNA more than 200 kb in length with a CG content of over 50%, and a ratio of observed CG dinucleotides to the expected number must be greater than 0.6^{158,166-168}. There are thought to be approximately 29,000 CpG islands within the human genome¹⁵¹. The 2kb regions of DNA flanking both sides of a CpG island are known as CpG shores. Methylation at CpG shores has been shown to have an inverse relationship with gene expression¹⁶⁹. The subsequent 2 kb regions flanking both the north and south shores are known as CpG shelves.

Many gene promoters overlap CpG islands. The addition of methyl groups to CpG sites within gene promoters results in the subsequent silencing of that gene due to transcription factors being unable to bind as a consequence of MECP2 and MBD proteins having already bound to the DNA, while the removal of methyl groups results in the opposite effects¹⁵⁸. MBD2, MBD4, and TET proteins are thought to be involved in the removal of methyl groups from cytosine bases¹⁶⁴. Conversely, the

methylation status of CpG sites with the gene body is believed to be positively correlated with gene expression¹⁷⁰.

Additional functions of DNA methylation are gene imprinting, where one parental allele is silenced by being heavily methylated and the other is expressed due to being generally unmethylated, as well as X chromosome inactivation and transposon control^{151,154,155,160}. During development, the majority of the genome is unmethylated, and many promoter regions maintain this unmethylated state, however some gene promoters, such as those described above, become methylated^{151,159}.

1.3.2 Histone Modifications

DNA methylation does not act in isolation to regulate gene expression. Histone modifications are another important epigenetic mechanism which may have consequences on gene expression when altered due to resultant changes in chromatin structure. DNA is wrapped around histone proteins to form chromatin, and nucleosomes are the basic units of chromatin made up of a histone octamer. There are four core histones: H2A, H2B, H3 and H4 and each octamer is composed of one H3-H4 tetramer and two H2A-H2B dimers. In a single nucleosome, 147 base pairs (bp) of DNA are wrapped around this histone complex in a left-handed superhelix, with each turn at approximately 10.4 bp¹⁷¹.

During post-translational modification, several possible reversible changes may occur at the amino acids found at the *N*-terminal ends of histone proteins which include acetylation and methylation most commonly, in addition to phosphorylation, ubiquitination, sumoylation, ADP ribosylation, deamination, and proline isomerisation among many others^{149,153,156,164}. All of these modifications have a role in transcription regulation, but some have other roles in DNA repair (acetylation, lysine methylation, phosphorylation and ubiquitination), DNA condensation (acetylation and phosphorylation), and DNA replication (acetylation)¹⁵⁶. The *N*-termini play a major role in nucleosome packaging which in turn determines higher-order chromatin structure and therefore accessibility, meaning that addition or removal of histone modifications directly affect gene expression¹⁴⁹.

Chromatin that is accessible for transcription is known as euchromatin, while inaccessible or silent

chromatin regions are known as heterochromatin. There are two subtypes of heterochromatin, one that is permanently silenced called constitutive heterochromatin, and facultative heterochromatin may be reactivated under specific genetic or environmental conditions. In addition to being tightly compacted, the DNA in heterochromatin is also heavily methylated to further suppress expression and perhaps silence genes permanently ¹⁵¹. The classes of enzymes involved in the most common histone modifications are histone acetyltransferases (HATs), histone deacetylases (HDACs), methyltransferases, and demethylases. HDACs cause histone deacetylation which reduces the space between the DNA and a nucleosome, thereby reducing access for transcription factor binding and shifting the chromatin from a euchromatin to a heterochromatin state ¹⁵⁷.

Histone H3 lysine4 di- and trimethylation (H3K4me₂, and H3K4me₃ respectively) as well as H3 lysine36 di- and trimethylation (H3K36me₂, and H3K36me₃ respectively), and H3 lysine79 trimethylation (H3K79me₃) are marks generally associated with active transcription, while H3 lysine27 trimethylation (H3K27me₃) and H4 lysine 20 trimethylation (H4K20me₃) are associated with transcriptionally repressed regions ^{149,156}. Additionally, a higher concentration of acetylated histones is found located at or around the promoters of active genes, particularly H3 lysine9 acetylation (H3K9ac), with deacetylation associated with transcriptional repression. It is important to note however, that the same modifications may have both active and repressive effects depending on the genomic location of the mark ¹⁵⁶. Lysine acetylation has a rapid turnover with half-lives of minutes, while that of lysine methylation is much slower with half-lives ranging from half a day to several days depending on whether the mark is active (faster) or repressive (slower) ¹⁴⁹. Lysine methylation is associated with a number of different functions including both activation and repression of transcription, formation of heterochromatin and chromosome loss ¹⁵⁷.

DNA methylation and histones work hand in hand to regulate gene expression through the recruitment of HDACs and other chromatin-binding proteins to gene promoter regions by DNMTs ¹⁵⁴. DNMTs recruit HDACs to hypermethylated chromatin allowing them to form complexes with other

proteins¹⁵⁷. This subsequently blocks the binding of the transcriptional machinery. The mechanism described would suggest that DNA methylation changes precede histone modifications, however this is not always the case and DNMTs have been reported to be recruited following histone modifications^{157,160}.

1.3.3 Epigenetics and the Environment

Given the fluidity of epigenetic changes, it is unsurprising that changes to both the internal and external environment, as well as various exposures, may alter the epigenetic state. The epigenetic consequences of several environmental exposures have been described, including those caused by endocrine disruptors, pollutants such as particulate matter, heavy metals, infectious pathogens, radiation, indoor allergens and, most notably, tobacco smoke^{163,164}. Many of these exposures are associated with established disease phenotypes, however, other less-obviously dangerous exposures like diet, social influences, and temperature changes may also affect epigenetic patterns^{163,172}.

While epigenetic responses to the environment are often dose-dependent, the relationship between dose and response is not necessarily linear. The duration of the exposure and developmental stage are other important considerations. Acute, low-dose exposures during foetal development may have much greater consequences compared to a high-dose exposure in adults¹⁶⁴. Additionally, the consequences of low-dose chronic exposures may be unpredictable, with reported results for some exposures being equivalent to high-dose acute exposures, and others reporting opposite responses¹⁶⁴. It is important to note that underlying genotypes may result in an increased genetic predisposition to the effects of particular environmental exposures¹⁶³. A good example of this is differences in the methylation levels of the same genes between males and females which may consequently result in different histone and chromatin modifications. However, it is possible that disease progression may occur without any pre-disposing genotypes or genotoxic responses to environmental exposures¹⁷².

The epigenetic landscape is known to be tissue-specific and it follows that any epigenetic changes induced by environmental exposures would occur in a tissue-specific manner, with sensitivity also

differing between organs ¹⁶⁴. Particularly in human studies, epigenetic research is limited to tissues that are easy to obtain such as blood, but these epigenetic alterations are not necessarily representative of those occurring at target organs ¹⁶¹. For many environmental exposures, the triggered epigenetic changes occur in specific patterns related either to the exposure itself, or diseases related to the exposure including pulmonary diseases, cardiovascular diseases, obesity, and cancer ^{161,164}.

1.3.4 Epigenetic Changes and Cancer

The cancer epigenetic landscape has been well-described in literature and dysregulation of epigenetic mechanisms is a recognised hallmark of cancer. The same differentially methylated regions (DMRs) have been shown to be involved in epigenetic changes in both normal cell differentiation and cancer ¹⁶⁹. In cancer, the highly regulated mechanism of DNA methylation becomes dysregulated, very often hypermethylation occurs at CpG islands and shores while hypomethylation frequently takes place further away at distal regulatory regions and repetitive elements as exhibited by global hypomethylation ^{84,154,158,167,173}. Almost all cancers have been shown to have decreased overall methylation when compared to normal tissue, with this loss of methylation occurring largely at repetitive elements which are normally methylated ¹⁵⁸. An overall deficiency in the methylation at repetitive elements in tumour tissue results in genomic instability which aids the progression of tumorigenesis, particularly given that over 50% of the genome is made up of such elements ^{84,158}. This genomic instability stems from the unravelling of heterochromatin caused by the loss of methylation of tandem repeats which normally help keep the DNA tightly packaged. This can lead to translocations, chromosome rearrangements, and copy number variations ¹⁵⁸. Long interspersed nuclear element 1 (LINE-1) are retrotransposons which are silenced by being heavily methylated, and loss of methylation at these elements, along with other retrotransposons such as Alu, has been associated with a number of different cancer types ¹⁵⁸. Hypomethylation is also known to take place at CpG sites that are located outside of CpG islands but are still associated with gene promoters and

such events lead to the expression and activation of genes that are silenced in normal circumstances which may disrupt normal cellular processes^{84,158}.

Despite the global hypomethylation associated with cancer genomes, hypermethylation events also occur which are modulated by the overexpression of DNMTs. Imprinted genes are naturally found to be hypermethylated, with the majority of other CpG islands usually being hypomethylated. In early tumorigenesis however, CpG islands tend to be the targets of hypermethylation events which often results in the silencing of tumour-suppressor genes, DNA repair genes, and transcription factors in a process termed epigenetic silencing¹⁵⁰. In fact, it has been suggested that epigenetic silencing occurs more frequently in cancer development than mutational events¹⁵⁰. Hypermethylation of promoters promotes tumorigenesis and is associated with overexpression of DNMTs which is a common characteristic of many tumours^{84,158}. Methylated CpG sites may be considered mutational hotspots due to the transition of 5-mC to T on deamination¹⁵⁴. These mispairing lesions are not easily recognised and therefore are also not easily repaired¹⁵⁴.

ROS such as hydroxyl radicals produce DNA lesions such as 8-hydroxy-2-deoxyguanosine which can themselves prevent DNMT binding^{84,162}. Associations have been found between the induced gene-silencing of key antioxidant enzymes required to metabolise ROS and tumour development due to gene promoter hypermethylation⁸⁴. Several environmental exposures have been linked to carcinogenic outcomes, with global hypomethylation a consistent observation in cancer genomes which may, in part, be mediated by oxidative stress¹⁶². Many environmental toxins are known to induce oxidative stress, which results in genes like *SIRT3* and *IDH2* becoming overexpressed. These in turn result in the production of α -KG which is known to activate TET proteins which catalyse the conversion of 5-mC to 5-hmC.

Acetylation, deacetylation, arginine methylation, lysine methylation, and demethylation histone modifications have also been implicated in cancer. Acetylation of a lysine residue creates a further surface to allow for the binding of transcription factors and chromatin regulators, and cell

proliferation is highly dependent on the correct acetylation patterns maintained by HATs. Any mutations occurring at HAT genes may therefore promote proliferation¹⁵⁷. Increased expression of *HDAC4*, *HDAC8*, and *HDAC9* particularly have been associated with the silencing of tumour suppressor genes given the function of HDACs in altering the chromatin state from euchromatin to heterochromatin¹⁵⁷. Methylation of histone arginine residues is associated with the transcriptional activation of many tumour suppressor genes, and this histone mark works synergistically with histone acetylation to achieve this. The repressive histone modification H3K27me3 is catalysed by EZH2 which is involved in the maintenance and differentiation of stem cells. Overexpression of this gene is common in many cancers which leads to gene silencing¹⁵⁸. Other genes associated with histone modifications and known to be dysregulated in cancer are *JMJD2C* and *MLL*.

Many of these observed mechanisms do not only occur in cancer, but also during the development of other diseases¹⁵⁴. The progression of many diseases is based on changes in gene expression and chromosome instability both of which may be modulated by the epigenetic mechanisms as described.

1.4 PAH Exposure and Epigenetics

Some work has been carried out to assess the relationship between PAH exposure and epigenetic mechanisms, particularly DNA methylation, and the sections below aim to briefly summarise the overarching trends. In subsequent chapters, the state of the knowledge linking PAH exposure and DNA methylation changes will be discussed in detail. Studies have been published reporting differential global and gene methylation caused by PAH exposure, and *in utero* exposures are considered to have the most harmful effects.

At known mutation hotspots, PAH-DNA adducts form preferentially at guanine bases adjacent to methylated cytosines^{69,74,174}. BPDE-DNA adducts have been shown to inhibit both the maintenance and *de novo* methylation of DNA¹⁷⁵. One likely cause is that adducts tend to form preferentially at guanine bases, and when these are repaired, this can lead to mutations resulting in the CpG site being lost¹⁷⁵. Another reason for this is the reduced binding of DNMTs to DNA due to the presence of

adducts¹⁶⁵. Additionally, BPDE adducts inhibit the transfer of methyl groups from S-adenosylmethionine to cytosine bases¹⁷⁶. Another study has shown that the interaction between BPDE-DNA adducts and DNA methylation is dependent on the stereochemistry and position of the adduct, as well as the methylation status on the complementary strand¹⁷⁷. In addition to reactivity, the methylation status of cytosines adjacent to guanines affect the structural conformation of the (-)-trans-BPDE-*N*²-dG adduct. An unmethylated cytosine results in the formation of an adduct with a minor groove structure external to the DNA duplex, whereas a methylated cytosine leads to an intercalative conformation¹⁷⁸.

The lesions caused by ROS and oxidative stress may interfere with the ability of DNMTs to bind to DNA, resulting in global hypomethylation which has been frequently associated with the progression phase of carcinogenesis⁸⁴. Additionally, the formation of 8-oxo-G and *O*⁶-methylguanine lesions at guanine bases make it difficult for adjacent cytosines to become methylated. The latter may also mispair with thymine which may also contribute to global hypomethylation. As described above, global hypomethylation may result in chromatin condensation and transcriptional inactivation⁸⁴. Oxidative stress has also been associated with levels of hydroxymethylation¹⁶⁴. TET enzymes responsible for the conversion of 5-mC to 5-hmC are sensitive to the intracellular redox environment which suggests that oxidative stress induced by PAH exposure and metabolism may also alter genomic hydroxymethylation¹⁶⁰.

Presence of BPDE-DNA adducts have been shown to stabilise the nucleosomes by various chemical mechanisms including hydrogen bonding between DNA and histone proteins which may explain the lesions' resistance to NER¹⁷⁹. Additionally, BPDE-DNA adducts have been shown to trap histone tails, preventing them from carrying out their various functions and causing altered chromatin structures by impeding the binding of proteins such as HATs and HDACs¹⁸⁰. Several studies spanning the late 70's and 80's have clearly established the interaction between PAH-DNA adducts and histone proteins. Both H3K4me3 and H3K9ac have been shown to be increased at the human LINE1 promoter

subsequent to B[a]P exposure *in vitro* resulting in the decreased ability of DNMT1 to bind to the region and reactivation of the retrotransposon ¹⁸¹. These observations may provide a mechanism for the observations of reduced methylation at repetitive elements as a result of PAH exposure. In a genome-wide study of human cells, B[a]P induced the hyperacetylation of 1456 gene promoters and hypoacetylation of a further 775 ¹⁸². Several of these changes were observed at genes associated with chromatin remodelling, transcription, cancer, and DNA damage ¹⁸². At the expression level, B[a]P exposure has been shown to dysregulate expression of HDAC proteins. In rats it has been shown that *HDAC1* and *HDAC3* become down-regulated and *HDAC5* is overexpressed following gestational exposures ¹⁸³. Finally, chromatin remodelling mechanisms are involved in the activation of *CYP1A1* by the AhR complex which is known to be induced by PAHs as described above. Hyperacetylation of H3K14 and H4K6, trimethylation of H3K4, and phosphorylation of H3S10 are all associated with this process ¹⁸⁴.

1.5 Summary, Hypothesis, Aims and Objectives

The carcinogenicity of PAHs has been well-established through *in vitro*, *in vivo* and epidemiological studies, with links to scrotal cancer going as far back as the 1700's. DNA methylation, the most commonly studied epigenetic mechanism, has been established to be dysregulated in the cancer genome and is considered to be one of the hallmarks of cancer. Additionally, a growing number of studies has shown that environmental exposures modify DNA methylation, with the most well-studied exposure being tobacco smoke which has been shown to have long-lasting effects on DNA methylation in humans.

The number of studies demonstrating that PAH exposures have epigenetic consequences is increasing, however, many of the studies to date tend to focus on high exposures both *in vitro* and *in vivo*. Many of the concentrations used are much higher than those experienced in real-world environmental conditions. Additionally, of the epidemiological studies that have been carried out,

many these have focussed on either occupational or prenatal exposures. Finally, few studies have looked at the epigenetic changes caused by PAHs on a genome-wide scale.

The over-arching hypothesis of this thesis is that PAHs induce genome-wide DNA methylation changes, and that the resulting downstream consequences of these changes, be they gene expression changes or changes in DNA structure, could enable or contribute to the mutagenic and carcinogenic mechanisms of PAHs. In order to test this hypothesis, the thesis has two aims and each aim has their consequent objectives: the first aim is to measure DNA methylation differences in mice exposed to B[a]P to link the large body of genotoxic evidence already available to epigenetic outcomes and mechanisms. This will be achieved by preparing reduced representation bisulphite sequencing (RRBS) libraries from the lung tissues of mice exposed to different doses of B[a]P and intersecting any observed methylation differences to reported gene expression levels in these same samples. The second aim is to identify and understand the DNA methylation changes caused by exposure to environmentally-relevant concentrations of PAHs in humans. This will be achieved by estimating the inhalation and dietary exposures of cohort subjects to the eight most carcinogenic PAHs, with inhalation and dietary exposures considered separately and as a combined exposure. The inhalation exposure will be estimated using land use regression models built in a previous European study. The dietary exposure will be estimated using food frequency questionnaires and data from the literature regarding the concentrations of PAHs in various foods. Finally, the inhalation and dietary exposures will be assessed together in a combined analysis. These estimates will then be used in conjunction with data from the Illumina Infinium HumanMethylation450 platform to detect CpG loci associated with exposure.

2 Chapter 2 – Methods

Figure 2.1 shows a summary of all studies conducted and presented in this thesis.

2.1 RRBS of lung tissue from mice exposed to B[a]P

2.1.1 Mouse Samples

The lung tissue DNA used in this study was obtained from mice that had been treated with various doses of B[a]P for a different study as previously described¹⁸⁵. In this previous study by Lemieux *et al.* 2011, twenty-five-week old male Muta™Mouse animals were treated with B[a]P dissolved in olive oil for 28 days by oral gavage. There were five animals in each of the following dose groups: vehicle treated controls (olive oil only), and 25, 50 and 75 mg B[a]P/kg body weight/day. Three days after the final treatment, isoflurane anaesthesia was administered and the mice were sacrificed by cardiac puncture. Various tissues were collected, including the right lobe of the lung, which were flash frozen in liquid nitrogen and stored at -80 °C until DNA extraction. DNA was extracted as described by Labib *et al.* 2012¹⁴² from randomly selected frozen lung tissue. Throughout the experiment, the mice were caged in plastic film isolators (Harlan Isotec, U.K.) and food and water were available *ad libitum*. All experiments were approved by Health Canada's Animal Care Committee and the mice were bred, maintained and treated in accordance with the Canadian Council for Animal Care Guidelines. These lung DNA samples were obtained for this current study through Dr Volker Arlt from the Department of Analytical, Environmental & Forensic Sciences at King's College London who was involved in the previous studies by Lemieux *et al.* (2011) and Labib *et al.* (2012) for the purposes of quantifying the BPDE-DNA adducts in these tissues. All subsequent work described below and presented in this thesis was carried out by the author of this thesis unless otherwise stated.

2.1.2 Preparation of Reduced Representation Bisulphite Sequencing (RRBS) Libraries

The Premium RRBS Kit developed by Diagenode (Cat. No. C02030032, Diagenode, Belgium) was used to prepare the libraries for 12 of the mice described above, 3 mice from each of the exposure groups

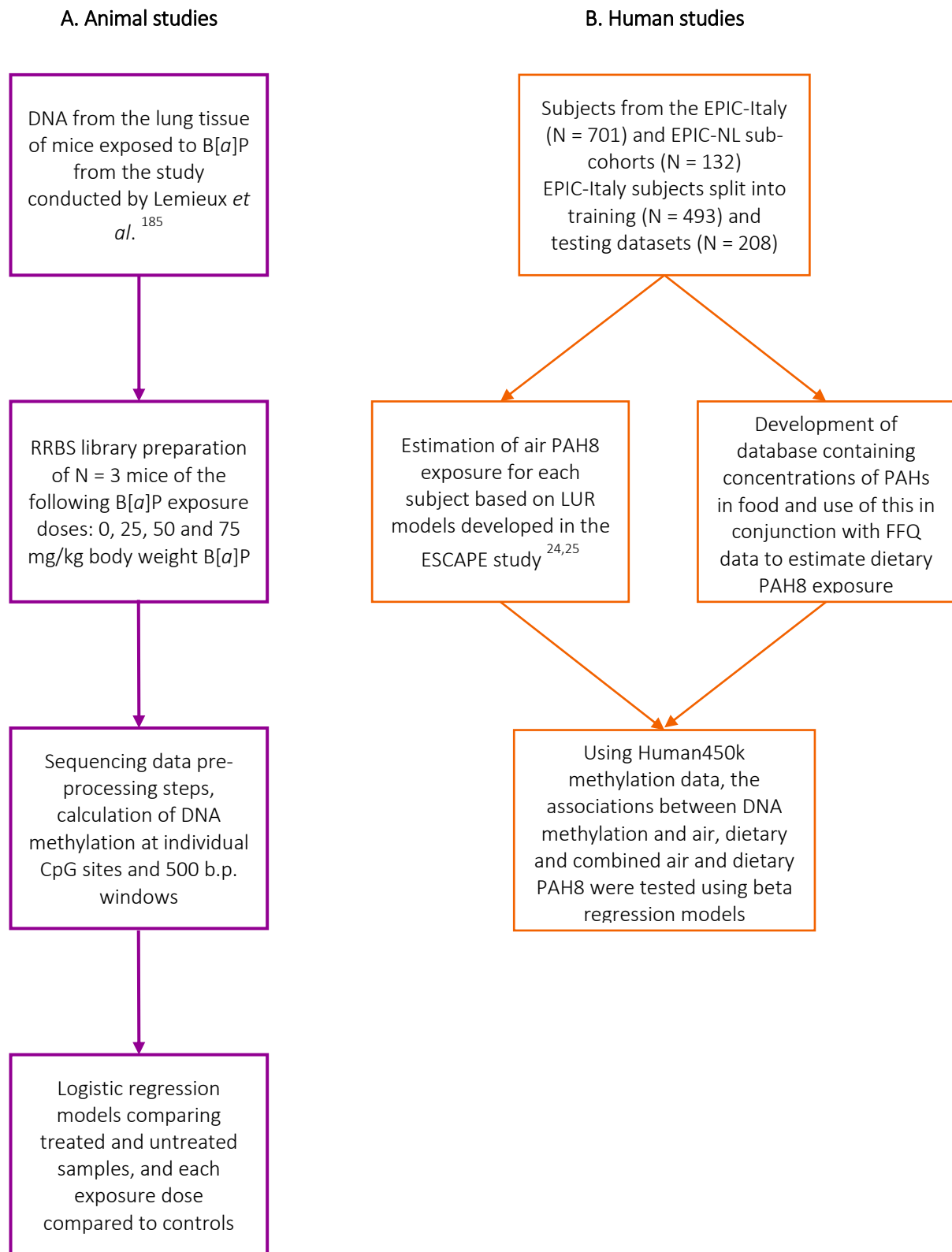


Figure 2.1. Figure showing summary of studies conducted and presented in this thesis

and 3 control mice. The workflow comprised 6 steps which have been previously described¹⁸⁶. The first step involved enzymatic digestion of 100 ng of DNA in 26 µl with *MspI* restriction enzyme which formed sticky ends as shown in Figure 2.2 below. Ends preparation (Figure 2.3) was carried out using dNTPs and an ends preparation enzyme to modify the sticky ends to allow for subsequent adapter ligation (Figure 2.4) which would allow for multiple libraries to be pooled together downstream. The digested DNA samples for each mouse had their own unique 6 b.p. adapter sequence.

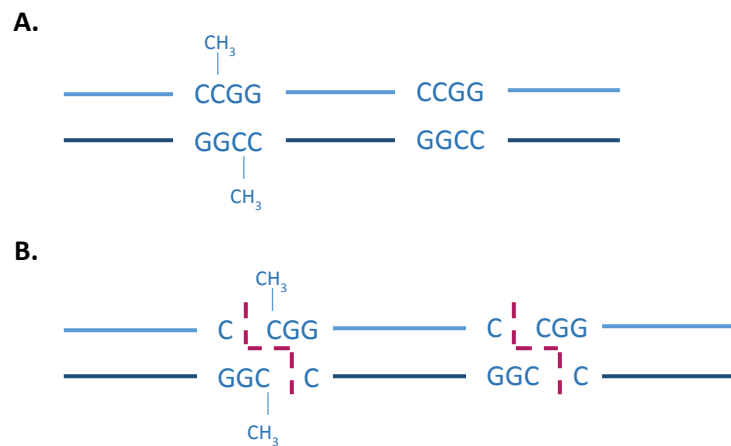


Figure 2.2 Schematic of *MspI* recognition sites (A) and cleavage pattern to form sticky ends (B).



Figure 2.3. Schematic showing ends preparation on *MspI* digested fragments.

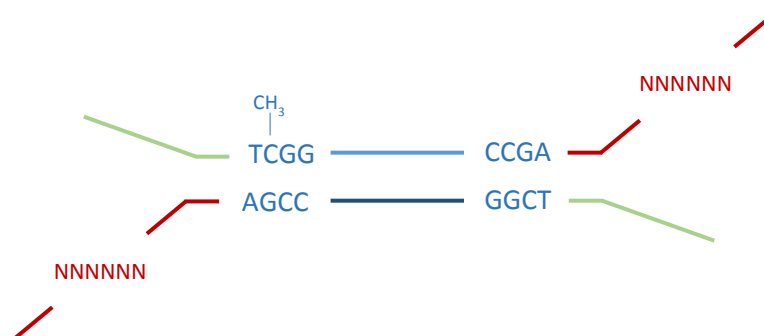


Figure 2.4. Schematic showing ligation of adapter sequences to DNA fragments.

Fragment size selection was carried out using home-made magnetic AMPure XP beads, however the ratio differed to the x0.8 ratio suggested in the protocol. A ratio of 0.5x was used to remove fragments larger than 500 b.p. which were bound to the beads and then discarded. More beads were added to the supernatant in a ratio of 0.2x to bind the larger of the remaining fragments which would have had a size of between 200-400 b.p.. Once the supernatant was discarded, the DNA was eluted from the beads using the resuspension buffer provided in the Premium RRBS kit. Since the libraries from each sample were to be pooled, the volume of each library to be added to the pool was calculated using the Ct values from a qPCR run and the recommended formula given by Diagenode:

$$17 \times 2^{-dCt}$$

Where $-dCt$ refers to the difference in Ct value as follows:

$$dCt = Ct_{max} - Ct_{sample}$$

Once the samples were pooled, bisulphite conversion was carried out. This process converted all unmethylated cytosine bases in the pool to thymines, meaning that any cytosines that remained were methylated (Figure 2.5).

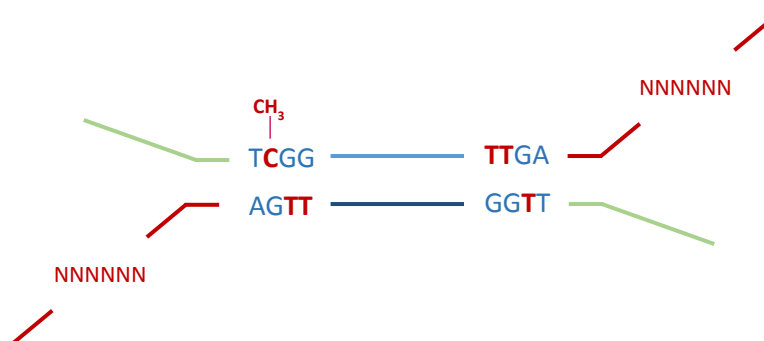


Figure 2.5. Schematic illustrating the effects of bisulphite conversion. Methylated cytosine bases are unaffected while unmethylated cytosines are converted to thymines.

The final step of the library preparation protocol included an enrichment PCR to amplify the library (Figure 2.6), a clean-up step using home-made magnetic AMPure XP beads as recommended in the

protocol, and quality control for which an Agilent High Sensitivity D5000 ScreenTape tape was run on the BioAnalyzer Agilent 2200 TapeStation (Agilent Technologies) to ensure that the library was enriched for fragments at the expected size (≈ 260 b.p). The pooled libraries were sequenced at the Imperial BRC Genomics Facility using Illumina HiSeq2500 100 b.p. paired end sequencing on two lanes of a flow cell.

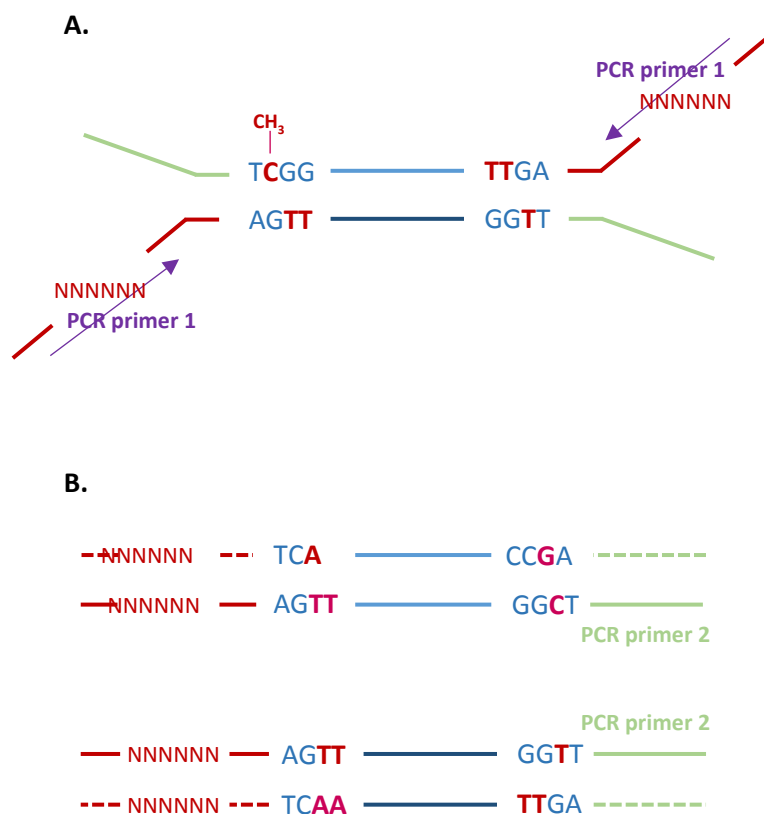


Figure 2.6. Schematic showing where the first enrichment PCR primers bind (A.), the amplicons for each strand and where the second set of PCR primers bind (B.) to produce fragments to be sent for sequencing after clean up and quality control.

2.1.3 Pre-processing of Data

The files of data from each mouse sample were processed individually as described below using the R statistical program and various packages within it, as well as several bioinformatics tools available on the Imperial College High Performance Computing (HPC) servers. Since the data were sequenced on two lanes of a flow cell, the files from each lane were concatenated together, ultimately resulting in

the formation of two files per sample, one for each set of reads (R1 and R2) produced by the paired-end sequencing. The adapter sequences were trimmed from the reads using the Trim Galore! program from Babraham Bioinformatics. Alignment to the *mm10* genome was carried out using the Bismark program called Bowtie2, and methylation calling was done using the Bismark methylation extractor to produce a methylation coverage file which contained the coordinates (chromosome and position) of each sequenced site along with the counts of Cs and Ts for each sequenced site. The data were filtered to exclude all reads which did not map to chromosomes 1-19 and chromosomes X and Y. Where data were available for two sites in consecutive positions on the same chromosome, this indicated that data were available for both strands, and this was “collapsed” into a single data point by merging the data. Sites with a total number of reads less than 10 and more than the number of reads corresponding to the 99.9th percentile were removed. Methylation for each site was then calculated as a percentage as follows:

$$\% \text{ Methylation} = \frac{\text{Number of methylated reads}}{\text{Total number of reads}} \times 100$$

Methylation levels for sites with total number of reads below the median number of reads for that sample were weighted, while those with a number of reads above the median were assigned a weight of 1 to avoid over-weighting. The weights for each site were calculated using the following equation:

$$\text{Weight} = \frac{\text{Number of reads}}{\text{Median number of reads}}$$

The weights were then multiplied by the methylation levels to produce weighted methylation.

2.1.4 Finding Associations Between DNA Methylation and B[a]P Exposure

2.1.4.1 Sliding Window Analysis

The *mm10* genome was tiled into windows of 500 b.p. with a 250 b.p. overlap using the makewindows tool in the bedtools programme, and these windows were intersected with the data from each mouse individually such that all sequenced sites were assigned to a window. The data were

filtered to windows that were common to all 12 samples (N = 152,691 windows), and the methylation for each window was calculated as the average weighted methylation of all sites in that window. All the common windows were annotated using the HOMER (Hypergeometric Optimization of Motif EnRichment, V4.9) software in order to determine the underlying distribution of genomic features such as number of windows in promoter regions, as well as the gene nearest to each window and the distance to the transcription start site (TSS) of that gene.

The ‘methylKit’ R package¹⁸⁷ was used to develop four logistic regression models to find differentially methylated windows (DMW) between control mice and those exposed to B[a]P. The initial model compared the untreated mice to all mice treated with B[a]P irrespective of dose. The subsequent models compared the control mice to the mice treated with each B[a]P dose separately. All models were corrected for overdispersion, and model results were filtered for windows which showed a statistically significant ($p < 0.05$) difference in methylation of at least 25%. Correction for multiple testing was not applied to these models due to the low number of samples and limited statistical power. The distributions of these DMWs were compared to the underlying genomic feature distribution to find enrichment/depletion of particular features. This comparison was carried out using Fisher’s Exact Tests and structuring the data as shown in the matrix in Table 2.1.

Table 2.1. Matrix used in Fisher's Exact Tests to assess enrichment or depletion of methylation changes at genomic regions compared to distribution of all tested sites/windows

Number of differentially methylated windows with genomic annotation	Total number of differentially methylated windows - Number of differentially methylated windows with genomic annotation
Number of 500 b.p. windows tested with genomic annotation	Total number of 500 b.p. windows tested – Number of 500 b.p. windows tested with genomic annotation

Comparisons were also made between the direction of the difference i.e. hypo- and hypermethylation at each genomic annotation. The ratio of hypomethylation to hypermethylation events at each genomic region was compared to the ratio of all hypomethylation to hypermethylation events using Fisher's Exact Tests, with the matrix structure shown in Table 2.2 .

Table 2.2. Matrix used in Fisher's Exact Tests to compare hypomethylation and hypermethylation events for each genomic annotation to the overall ratio.

Number of hypomethylated windows with genomic annotation	Total number of hypomethylated windows - Number of hypomethylated windows with genomic annotation
Number of hypermethylated windows with genomic annotation	Total number of hypermethylated windows – Number of hypermethylated windows with genomic annotation

Results were compared to those of previous B[a]P exposure studies as listed on the Comparative Toxicogenomics Database (<http://ctdbase.org/>)¹⁸⁸. Associations between gene expression and methylation were analysed for DMWs annotated to genic regions (5'UTR, promoter, exon, 3' UTR, and transcription termination sites (TTS)). The gene expression data were generated in the study carried out by Labib *et al* (2012)¹⁴² by synthesising cDNA from RNA extracted from the lungs of the mice and running that on the Agilent 4x44K oligonucleotide microarray. Spearman's correlation was used to test associations between DNA methylation and gene expression. Gene expression data were not available for one of the mice in the medium dose (50 mg B[a]P/kg bw) and so this sample was excluded in the analysis comparing gene expression and DNA methylation.

2.1.4.2 CpG Level Analysis

A second analysis looked at the data at the individual CpG level, without grouping into genomic windows. The analysis pipeline was the same as described in the preceding section: The data were filtered to CpG sites common to all mice (N = 38,874), and the methylation for each site was taken to

be the weighted methylation. All sites were annotated using HOMER (Hypergeometric Optimization of Motif EnRichment, V4.9) to ascertain the distribution of the sites and distance to the nearest TSS. The subsequent analyses were identical to those described in the sliding window section above.

2.2 Air and Dietary Exposure to PAHs in EPIC Subjects

2.2.1 EPIC Cohort

The European Prospective Investigation into Cancer and Nutrition (EPIC) cohort is a prospective cohort with over half a million participants from 23 centres located in 10 countries in Western Europe. The rationale and study design of the EPIC study are described in the manuscript of Riboli and Kaaks ¹⁸⁹. Recruitment took place between 1992 and 1999 and data were collected about diet, nutritional and metabolic characteristics, lifestyle factors, medical history, and cancer risk. Additionally, blood samples were collected at baseline for 387,889 subjects. The EPIC study protocol was approved by the ethical review boards of the International Agency for Research and Cancer (IARC) and by the local participating centres.

For the study presented in this thesis, data from two EPIC sub-cohorts were used: EPIC-Italy and EPIC-Netherlands (EPIC-NL) which are represented in Figure 2.7.. The eligibility criteria for inclusion were that for a participant subject the following data must have been available: anthropometric data, methylation data from the Illumina Infinium HumanMethylation450 platform, road and traffic variables required to estimate air PAH8 exposure which are described in more detail in section 2.2.3, and food frequency questionnaire data. The Dutch subjects were recruited from two cohorts called Prospect and the Monitoring Project on Risk Factors for Chronic Diseases (MORGEN) ¹⁹⁰. Prospect is a prospective cohort study in 17,357 women aged 49-70. Between 1993 and 1997 these women participated in the nationwide Dutch breast cancer screening programme. 115 women from Prospect were included in the current study. MORGEN is a general population sample of 22,654, 20-59 year old men and women. Between the years of 1993 and 1997 approximately 5000 new, random subjects were examined. Only 17 women from MORGEN were included in the current study. This is due to the

requirement of the availability of subjects to have both DNA methylation data and the variables required to estimate air and dietary PAH8 exposures, all of which are discussed in the next section. The EPIC-Italy subjects come from two centres: Turin and Varese. For both, recruitment took place between 1993 and 1998, with the former having 10,604 volunteers and the latter 12,083. In the current study, 550 subjects came from the Turin centre and 151 from the Varese centre, with subjects originating from two case-control studies on breast and colorectal cancer. In Epic-Italy blood samples were drawn at least one year before cancer diagnosis for cancer cases. Given that only a small number of subjects have been included in this thesis due to eligibility criteria outlined above, it is likely that these subjects are not representative of the wider study populations from which they originated, but rather these subjects are a sample selected for exploratory analyses based on availability of all the required data.

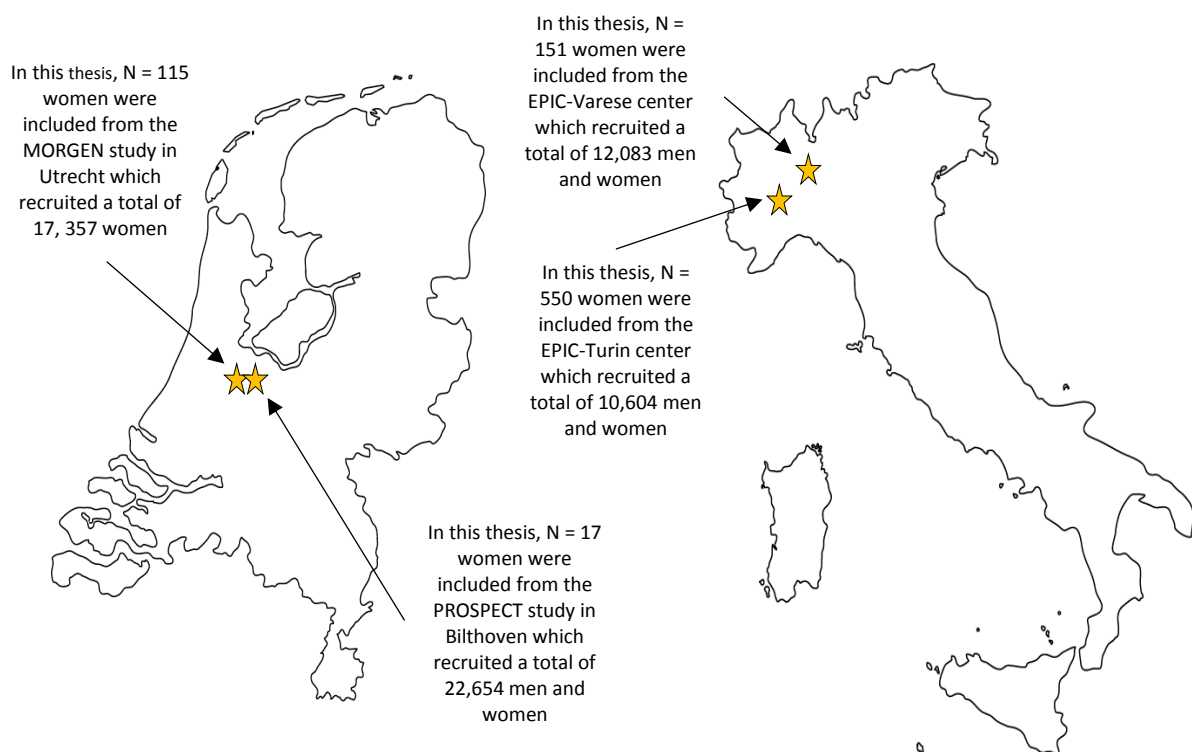


Figure 2.7. Maps showing where EPIC-Italy and EPIC-NL subjects were recruited, and total study populations of each recruitment center. N.B. Maps are not drawn to the same scale

2.2.2 DNA Methylation Data

Generation and pre-processing of methylation data using the the Illumina Infinium

HumanMethylation450 platform was carried out as described previously^{191,192} not as part of the studies presented in this thesis but previously by other group members involved in the EPIC study.

The general outline of the steps undertaken has been included here: Briefly, using the Illumina Infinium HumanMethylation450 platform and blood from the participating subjects, genome-wide methylation analyses were carried out. For the Dutch samples, laboratory procedures were carried out at ServiceXS BV in Leiden, Netherlands following manufacturers' protocols. The Italian samples were processed and analysed at the Human Genetics Foundation in Turin. Using the QIAGEN QIASymphony DNA Midi Kit, extraction of genomic DNA from thawed buffy coat layers was carried out. Bishulphite-conversion was carried out on 500 ng of DNA per sample using the Zymo Research EZ-96 DNA Methylation-Gold™ Kit and the samples were then hybridised to Illumina Infinium HumanMethylation450 BeadChips. The arrays were subsequently scanned and control probes were used to evaluate sample quality using the Illumina HiScanSQ system. Exportation of raw intensity data were carried out using Illumina GenomeStudio (version2011.1) followed by data pre-processing using in-house software written for the R statistical computing environment. Missing measurements were defined as those obtained by averaging over less than three beads or if average intensities were below those obtained from negative control probes. Background subtraction and dye bias correction (for probes using the Infinium II design) were also performed. Methylation levels were expressed as beta values i.e. the ratios of intensities from methylated cytosines over the summed intensities of methylated and unmethylated cytosines.

Some probes on the Illumina Infinium HumanMethylation450array are known to be polymorphic CpGs located on single nucleotide polymorphisms (SNP) in addition to a number of probes known to cross-hybridise to other probes^{193,194}. Consequently, these probes were removed prior to further analysis along with probes for sites on the sex chromosomes. Probes were also excluded if <80% of

subjects had complete data for that probe. As a result, in this study 365,714 and 337,993 probes were interrogated within the EPIC-Italy and EPIC-NL cohorts respectively.

2.2.3 Air PAH8 Exposure Estimation

The European Study of Cohorts for Air Pollution Effects project (ESCAPE) is a study that contains several air pollution measurements, with the aim to investigate the long-term effects of air pollution. Estimates of air PAH8 exposure were calculated from ESCAPE and all subjects were included in this thesis due to having both the methylation array data described above and having been included in ESCAPE. As part of the ESCAPE study, a three-step procedure was used to build land-use regression models for the study participants as previously described^{24,25}. Briefly, air samples were collected over different seasons between October 2008 and April 2011 and for these, air pollution components such as PM10 (particulate matter with an aerodynamic diameter less than 10 µm in size), PM2.5 (particulate matter with an aerodynamic diameter less than 2.5 µm in size), NO₂ and NO_x were measured and characterised. The organic components of PM2.5 were further analysed resulting in measurements for PAHs, elemental carbon, organic carbon, and hopanes/steranes. A land-use regression model was built for each of these pollutants in each study area in a total of 10 European countries. These land-use regression models used the yearly mean concentration as the dependent variable and an extensive list of geographical attributes as possible predictors. The geographical attributes included building density, road traffic, and population.

In the present study, the LUR models for ΣPAHs from Jedynska *et al.*²⁵ were used, where ΣPAHs was made up of the 8 most carcinogenic PAHs (PAH8): B[a]A, B[b]F, B[k]F, B[ghi]P, B[a]P, Chr, D[a,h]A and I[cd]P. In the Netherlands, the model was generated for data from 16 monitoring sites in Rotterdam, Amsterdam, Groningen, and Amersfoort. The final model had an R² of 58%, an R² of 31% from leave-one-out cross validation (LOOCV) and a root mean squared error (RMSE) of 0.511 ng/m³. The measured concentration of PAH8 was reported to be 1.40 ng/m³ with a range from 0.44-3.24 ng/m³. The final model included road length of major roads in a 50m buffer and urban green in a 5000m

buffer as covariates. The model for Italy was developed from data from 20 monitoring sites in Rome. The final model had an R^2 of 84%, an R^2 of 74% from LOOCV, and an RMSE of 0.389 ng/m³. The measured concentration of PAH8 was 2.03 ng/m³ with a range from 0.86-4.58 ng/m³. The covariates included in the model were road length of major roads in a 50m buffer, the inverse distance to the nearest major road, and industry in a 5000m buffer. Using these models, the estimated air PAH8 exposure was calculated in ng/m³ for the chosen EPIC-Italy and EPIC-NL subjects. Truncation was carried out meaning that subjects whose air PAH8 exposure fell outside the range measured in the ESCAPE study were excluded.

2.2.4 Dietary PAH8 Exposure Estimation

To estimate dietary PAH8 exposure, first a dataset using measurements available in the literature of PAH concentrations in various food was created. This dataset was then used in conjunction with food frequency questionnaire data available for the same subjects described above in order to estimate dietary PAH8 exposure.

2.2.4.1 Dataset of PAH concentrations in food

A literature search was carried out on the NCBI database PubMed using the search terms “Polycyclic aromatic hydrocarbons diet” and “Polycyclic aromatic hydrocarbons food”. The studies had to meet two inclusion criteria:

1. The studies had to have been carried out from 1980 onwards.
2. The studies had to be carried out in Europe, using food samples obtained within Europe.

While the first criterion was decided upon arbitrarily, for the purpose of having a cut-off point, the second was chosen because the subjects included in the EPIC-Italy and EPIC-NL cohorts were based in Europe at least at the time of recruitment. The literature search resulted in 88 studies (for which access was available) which met the criteria and were subsequently included in the dataset ^{195,196,205–214,197,215–224,198,225–234,199,235–244,200,245–254,201,255–264,202,265–274,203,275–282,204}. All the data from the studies were included in the final dataset, with the exception of a study by Rose *et al.* ²⁴⁹ who carried out a vast

number of measurements of different meats, various cooking methods and many cooking times. In this case, the measurements included in the dataset were those where the authors described the cooking time as 'normal', which implies representation of the average cooking times for those food items and methods. The dataset was separated into two main parts: studies which measured the PAH content of food items (N = 77) ^{199,200,210–219,202,220–229,203,230–239,204,240–244,247–251,205,252–261,206,262–269,273,274,207,275–282,208,209}, and studies which gave PAH concentrations for food types (N = 10) ^{195–198,201,245,246,270–272}. Food items throughout this thesis refer to individual foods e.g. butter, olive oil etc., while food types refer to general groups of foods e.g. oils and fats. Data for all the different PAHs measured in the various studies were included in the final datasets, even if only one study measured a particular compound. Also, where the study measured the groups of PAHs recommended by various bodies, such as the European Commission (EC) and the Scientific Committee on Food (SCF) which are described in the introduction of Chapter 5, these were also included. For all studies, the year and country in which the study took place were added to the datasets. The end result was two datasets: one which contained PAH concentrations for individual food items, and a second which contained PAH concentrations for food types. The PAH concentrations included were all measured in µg PAH/kg of food or an equivalent unit.

2.2.4.2 Estimating Dietary PAH8 Exposure in the EPIC Cohort

The data collected about PAHs in food were then used to estimate the daily intake of subjects from the EPIC cohort described in the previous chapter as all subjects had food frequency questionnaire (FFQ) data available. The granularity of the FFQ data differed slightly between the two sub-cohorts. In order to simplify the calculation of the exposure estimates, the FFQ variables in EPIC-Italy were merged into seventeen variables found in EPIC-NI:

- Potatoes and other tubers
- Vegetables
- Legumes

- Fruits, nuts and seeds
- Dairy and dairy products
- Cereal and cereal products
- Meat and meat products
- Fish and shellfish
- Egg and egg products
- Fat
- Sugar and confectionary
- Cakes and biscuits
- Non-alcoholic beverages
- Alcoholic beverages
- Condiments and sauces
- Soups and bouillons
- Miscellaneous

Throughout the rest of this thesis, these variables will be referred to as food classes, to distinguish between them and the datasets of food items and food types. The intake of each of these food classes per subject was measured in g/day.

The exposure estimate for each subject was calculated based on the content of PAH8 as described above and by CONTAM²⁸³ for two reasons: the air pollution exposures were estimated based on concentrations of the same eight PAHs, and these PAHs are believed to have the highest carcinogenic potential. The PAH concentrations dataset was reduced to include only the eight PAHs, and since not all studies measured all eight PAHs, the following imputation steps were carried out separately for the food item and food types datasets. Each item and food type was annotated with one of the food classes in the list above. The data were subset by food class so that any imputed values would be based on other values most closely related to the missing data. Where more than 4 PAHs for a single

food item or type were missing, the median value for each PAH within the food class was assigned to the missing values of that PAH. Subsequently, k-nearest neighbour imputation was used for instances where the data were more sparse. Following imputation, the sum of the eight PAHs was taken resulting in a PAH8 estimate for each food item and food type in the datasets.

To obtain the estimates of dietary PAH8 intake per subject per day, the median PAH8 of each food class was multiplied by the intake of each food class for each subject, resulting in a dietary exposure estimate measured in ng PAH8/g intake of food in food class/ day. This calculation was carried out only using the medians from the food types dataset. The reason for this is that the data sources for the food items dataset were more heterogeneous than the food types studies, resulting in large variations within food classes and skewed data due to outliers.

2.2.5 Combined Air and Dietary PAH8 Exposure

Air PAH8 exposure and dietary PAH8 exposure as estimated above resulted in exposure measurements on very different scales and different units (ng/m³ in the former, and ng/day in the latter). In order to combine the two estimates to obtain a single estimate for the combined exposures, the air and dietary PAH8 estimates were scaled and converted to Z-scores independently. The Z-scores for each exposure were then summed together to obtain a combined air and dietary PAH8 score of exposure.

2.2.6 Data and Models

Large underlying differences which are further described in Chapter 4 were observed between the EPIC-Italy and EPIC-NL subjects, and it was hypothesised that these differences would impede replication or validation of results between the two datasets. For this reason, the EPIC-NL subjects were kept together as a single dataset, but the EPIC-Italy subjects were split into a Training dataset (70%, N = 493) and a Testing dataset (30%, N = 208) to establish a dataset in which observed results were more likely to validate given the similarities. These datasets were used for all analyses described below.

Models were built *a priori* for each dataset as shown below where PAH8 exposure represents air PAH8 exposure, dietary PAH8 exposure or combined air and dietary PAH8 exposure, depending on the analysis. Confounders were chosen *a priori* to ensure that the most well-established technical and anthropometric confounders in epigenetics studies were included, while also ensuring that the number of included confounders was minimised to prevent model over-fitting. WBC distributions, smoking status, sex, and age have all been well-established to be associated with DNA methylation in the literature and were included as confounders in the models. A publication by colleagues who also worked with the EPIC-Italy cohort also recommended that cancer case status should be adjusted for in this cohort ¹⁹¹. Finally, the common technical covariates of position on array chip and array chip position were included in the models to account for any potential batch effects.

Model 1. Model used on the Training dataset.

$$\text{Methylation} = \text{PAH8 exposure} + \text{Array chip} + \text{Position on array chip} + \text{WBC proportions} + \text{Center} \\ + \text{Age} + \text{Sex} + \text{Cancer case status} + \text{Smoking status}$$

Model 2. Model used on the Testing dataset.

$$\text{Methylation} = \text{PAH8 exposure} + \text{Position on array chip} + \text{WBC proportions} + \text{Center} + \text{Age} + \text{Sex} \\ + \text{Cancer case status} + \text{Smoking status}$$

Model 3. Model used on the EPIC-NL dataset.

$$\text{Methylation} = \text{PAH8 exposure} + \text{Array Chip} + \text{Position on array chip} + \text{WBC proportions} + \text{Center} \\ + \text{Age} + \text{Smoking status}$$

The difference between models 1 and 2 is that model 2 does not adjust for potential batch effects between the different Illumina Infinium HumanMethylation450 microarray chips. While in general, it is good and common practise to adjust for this ²⁸⁴, in the present case there were over 90 unique chip IDs in the Testing dataset which, along with the other covariates, would have led to significant over-fitting of the model given that the dataset included only 208 subjects. Given that the Training dataset had many more subjects, it was more robust to this adjustment. Model 3 used on the EPIC-NL dataset

also did not adjust for array chip for the same reasons. This model also did not adjust for sex or cancer case status as all subjects were female and had a control status. WBC proportions were estimated using the Houseman method which is well-established and commonly employed in such studies ²⁸⁵. The Houseman method uses the methylation levels of a specific set of probes from the Illumina Infinium HumanMethylation450 microarray to estimate the proportion of the different WBCs in each subject.

These models were fitted using generalised linear models with parameters recommended by Ferrari and Cribari-Neto ²⁸⁶ who developed the method for response variables restricted to an interval. Since methylation can only take on values between 0 and 1 (or 0 to 100%), it is from the beta distribution and fits the criteria for the use of beta regression. The beta regression method has previously been used in studies involving DNA methylation ^{191,192,287,288}. Often, methylation studies will convert the beta methylation values to M values which involves a Logit transformation of the beta values. This method is more statistically robust when linear models are employed since beta values tend to be heteroscedastic for values at either end of the distribution ^{289,290}. One of the main drawbacks of this method is that it tends to inflate very small differences in methylation at the extremes, for example, a difference between 0.001 and 0.0001 beta values is minimal as these sites would generally be considered to be unmethylated, but the corresponding M values are -9.96 and -13.29 respectively. Another drawback is that the resulting model coefficients are not interpretable as percentage change in methylation per unit change in exposure. Finally, despite the transformations, M-values do not always fit the assumptions made by linear models. Although the beta regression method used in the current study also does not allow for the interpretation of model coefficients as percentage change in methylation per unit change in exposure, it does overcome the inflation of beta values close to 0 or 1. While a number of R packages exist to carry out beta regression, in this thesis the *vglm* function which is part of the 'VGAM' package was used ²⁹¹. In order to obtain model estimates of methylation change per unit change of PAH8 exposure, mixed effect linear models were run on the beta values, adjusting for the technical covariates array chip and position on chip as random effects. However, these mixed

effects models were only used for the purposes of obtaining an interpretable coefficient. Finally, inflation of the test statistics generated from the beta regression models was tested for using the R package 'bacon'²⁹².

2.2.7 Global Methylation, Methylation at Genomic Regions and EWAS

2.2.7.1 Global Methylation

To interrogate any effect of air PAH8 exposure, dietary PAH8 exposure or combined air and dietary PAH8 exposure on global methylation levels, the arithmetic mean of all 365,714 probes in the Training and Testing datasets, and 337,993 probes in the EPIC-NL dataset was calculated for each subject. Each PAH8 exposure variable was also divided into quartiles, with each subject being assigned to a quartile based on their exposure. These quartiles are shown in the cohort description tables in the results section of Chapters 4, 5, and 6 for air, dietary and combined PAH8 exposures respectively, and these were then used in each of the above models along with the mean methylation per subject. The trend was analysed by running the regression using exposure as a continuous variable.

2.2.7.2 Methylation at Genomic Regions

Similar to the global methylation analysis described above, PAH8 exposure quartiles for each PAH8 exposure variable were used to assess any influence of exposure on methylation at genomic regions. The arithmetic mean of the methylation of all probes annotated to the following genomic regions was calculated: CpG island, CpG island shore, CpG island shelf, promoter region (all probes annotated to TSS200, TSS1500, 5' UTR and the 1st exon), gene body, 3'UTR and intergenic regions. Probes in intergenic regions were defined as all probes which were annotated to the promoter, gene body or 3'UTR regions. The number of probes in each genomic region for the EPIC-Italy and EPIC-NL cohorts are summarised in Table 2.3 . For each genomic region, the PAH8 exposure quartiles, continuous PAH8 exposure, and mean methylation were used in all three models described previously.

Table 2.3. Table of number of CpG probes at each genomic region analysed in the EPIC-Italy derived Training and Testing Datasets, and the EPIC-NL dataset

	Training and Testing Datasets	EPIC-NL Dataset
CpG islands	107,472	100,239
Shores	91,001	84,481
Shelves	35,402	32,580
Promoters	138,193	127,406
Gene body	125,842	116,378
3' UTR	13,226	12,046
Intergenic	88,453	82,163

2.2.7.3 EWAS

EWAS were carried out on all probes (N = 365,714) in the Training dataset using Model 1 described above for each PAH8 exposure separately. Two levels of multiple testing correction were carried out: the more stringent Bonferroni correction, as well as FDR correction. The reason for this is that correction for multiple testing is widely accepted to be a required adjustment in EWAS studies, but there is currently debate on the level of stringency required. Both sets of results are presented in the results chapters of this thesis where the cut-off values for Bonferroni correction are outlined, as these thresholds were dependent on the number of CpG sites tested in the relevant analysis. For FDR correction, $q < 0.05$ was used as the cut-off. In order to estimate effect sizes, a mixed effects model as described in the previous section was carried out on the significantly differentially methylated probes. All probes were then annotated using the HOMER (Hypergeometric Optimization of Motif EnRichment, V4.9) suite of tools in order to obtain a finer resolution of the genomic location of each probe. For each genomic annotation, Fisher's Exact Tests were carried out in order to determine whether more or less methylation differences were observed in the results than would be expected based on the distribution of all tested probes. In order to do this, the number of probes annotated to each genomic feature was calculated for both the significantly differentially methylated probes and all the probes tested (N = 365,714). Matrices as shown in Table 2.4 were then constructed with the following structure to determine any differences between the observed and expected distributions for each genomic annotation by feeding these into the Fisher's Exact test function in R:

Table 2.4 Matrix used in Fisher's Exact Tests to assess enrichment or depletion of methylation changes at genomic regions compared to distribution of all tested sites/windows

Number of differentially methylated probes with genomic annotation	Total number of differentially methylated probes - Number of differentially methylated probes with genomic annotation
Number of probes tested with genomic annotation	Total number of probes tested – Number of probes tested with genomic annotation

In order to further interrogate whether the differentially methylated probes in each genomic annotation were more hypo-/hypermethylated than expected, the observed ratio of hypomethylated to hypermethylated probes in each genomic annotation was compared to the overall ratio of hypomethylated to hypermethylated probes using Fisher's Exact Tests. For this analysis, matrices as shown in

Table 2.5 were constructed for each genomic annotation. These analyses were carried out for each set of EWAS results, i.e. differentially methylated probes with respect to air, dietary, and combined air and dietary PAH8 exposure.

2.2.8 Building a Methylation Index of PAH8 Exposure

Using the differentially methylated probes identified in the Training dataset, general linear models were built that could predict PAH8 exposure based on methylation at the probes in order to validate the EWAS results.

Table 2.5 Matrix used in Fisher's Exact Tests to compare hypomethylation and hypermethylation events for each genomic annotation to the overall ratio.

<p>Number of hypomethylated probes with genomic annotation</p>	<p>Total number of hypomethylated probes - Number of hypomethylated probes with genomic annotation</p>
<p>Number of hypermethylated probes with genomic annotation</p>	<p>Total number of hypermethylated probes - Number of hypermethylated probes with genomic annotation</p>

This was done for each set of results individually. First, all probes along with sex, age, cancer case status and smoking status were added to the model. In order to determine whether all probes and covariates were required in the model, a LOOCV method was used to estimate the α and λ model parameters. The α parameter is known as the shrinkage parameter and determines whether all variables in the model are included in the final model, i.e. no shrinkage and therefore $\alpha=0$ which is known as ridge regression, or whether some variable shrinkage will occur i.e. some variables will be given a coefficient of 0; if $\alpha > 0$ but < 1 this is called elastic net regression, and if $\alpha = 1$ this is called lasso regression. The second parameter estimated is called the λ parameter and this determines by how much each of the model coefficients will be penalised. The optimal parameters were selected from the LOOCV model which gave the smallest RMSE based on a number of models tested using different combinations of α and λ . The model was trained on the Training dataset, and tested on the

Testing dataset only. Model performance was evaluated by the correlation between the PAH8 exposure levels predicted by the model and the actual PAH8 exposure levels for each subject in each dataset.

3 Chapter 3 – RRBS of Lung Tissue from Mice Exposed to B[a]P

3.1 Introduction

3.1.1 Measuring DNA Methylation

Numerous techniques have been developed to measure DNA methylation and these have been reviewed many times^{293–296}. These techniques may be broadly divided into groups: immune-based methods such as ELISA, immunoprecipitation, and immunohistochemistry, chromatography-based techniques like high performance liquid chromatography (HPLC), mass spectrometry-based methods, microarray methods, and sequencing-based methods. The following sections will predominantly aim to summarise the latter two techniques due to them being the preferred techniques employed by studies in recent years. Both sequencing and microarray methods require the starting DNA material to be bisulphite converted. Bisulphite conversion is a method which converts unmethylated cytosines to thymines using sodium bisulphite, meaning that any cytosines remaining in the converted, amplified DNA would have been originally methylated thereby conserving the methylation pattern^{166,297}. In this way, an epigenetic mechanism is converted into a genetic mark which can be observed through sequencing or microarray hybridisation. Methylation levels for both microarray and sequencing methods are calculated by calculating the proportion of methylated cytosines to unmethylated cytosines, meaning that complete bisulphite conversion is a pre-requisite to obtaining reliable data. However, bisulphite conversion reduces the complexity of the genome and may cause fragmentation of DNA, both of which make amplification difficult²⁹⁴.

3.1.1.1 Microarray Methods

The Illumina GoldenGate DNA Methylation BeadArray was one of the first microarrays developed to measure DNA methylation. However this array was only able to measure at 1,505 CpG sites in 371 genes²⁹⁸ and so was succeeded by the Illumina Infinium HumanMethylation27 BeadChip array which interrogated over 27,000 CpG sites spread across the entire human genome. Following on from this, Illumina developed an expanded platform called the Infinium HumanMethylation450 BeadChip which

was able to interrogate the methylation at more than 485,000 human CpG sites²⁹⁹. Almost three quarters of all CpG sites on the array are associated with coding RNA transcripts. A significant proportion ($\approx 41\%$) of the CpG sites on the 450K array are located within gene promoters, followed by 31% within gene bodies, 25% intergenic CpG sites, and 3% at the 3' UTR²⁹⁹. The majority of CpG sites on this platform are located in open sea regions (intergenic), with CpG island regions having the second-highest representation, followed by CpG shores and shelves²⁹⁹. CpG shores are the 2kb regions flanking CpG islands, and shelves are the subsequent 2 kb flanks of the shores. Some of the drawbacks of this platform are the inclusion of CpG positions located on known SNPs and probes which have been shown to cross-hybridise, as well as low coverage of distal regulatory elements^{193,194}. The most recently developed platform is called the MethylationEPIC BeadChip which covers more than 850,000 genome-wide human CpG sites. This latest array includes more than 90% of the probes from the 450k array, including some of the problematic cross-hybridising probes. The aim of the EPIC platform was to improve coverage at regulatory elements, which included enhancers, however, it still only includes 7% of distal and 27% of proximal ENCODE regulatory elements³⁰⁰. The methylation across the distal regions has also been found to be highly variable³⁰⁰. Overall, each subsequent array development has provided significant improvements over the preceding technology. Advantages of using EPIC platform include the ability to use DNA extracted from formalin-fixed paraffin-embedded samples and that only low amounts of starting material (≈ 250 ng) are required, but this array is limited to human samples. It is widely accepted that microarrays provide highly reliable and reproducible results ($\pm 1-2\%$) in addition to being high-throughput and therefore ideal for the analysis of large numbers of human samples. While other microarray-based technologies do exist for the measurement of DNA methylation, such as those from Agilent and Affymetrix, the Illumina platforms are the most extensively used.

3.1.1.2 Sequencing-based methods

3.1.1.2.1 Whole Genome Bisulphite Sequencing

Whole genome bisulphite sequencing (WGBS) is a technique that combines bisulphite conversion with next generation sequencing (NGS) methods. This method allows for the interrogation of the entire methylome by providing the methylation status of almost every C base in the genome at single base resolution. While the advantages of using this technique are immediately apparent, there are significant downsides. The first main drawback is the cost of sequencing which, in many cases, is prohibitively expensive. This is followed by the significant bioinformatics challenges presented by such detailed and vast data. Lastly, the amount of starting DNA material required for this approach is quite large (μg scale compared to the ng scale required by many other techniques). In order to overcome at least some of these problems, targeted bisulphite sequencing techniques have been developed however these are still dependent on bisulphite conversion of the DNA. Recently, a review assessing biases of WGBS suggested that bisulphite conversion itself may induce significant sequencing biases which are further exacerbated when the prepared library is PCR amplified prior to sequencing, a step required by many WGBS library preparation protocols ³⁰¹.

3.1.1.2.2 Reduced Representation Bisulphite Sequencing

Reduced representation bisulphite sequencing (RRBS) is a method that allows for the measurement of genome-wide DNA methylation at the single nucleotide level. This is achieved by enriching CG-rich parts of the genome using a combination of restriction enzyme digestion and a bisulphite based technique which still captures most of the relevant regions such as CpG islands and promoters, while reducing the amount and cost of sequencing required ³⁰²⁻³⁰⁴. In this way the method is significantly more cost-effective than whole-genome bisulphite sequencing because at least one useful methylation measurement is covered at each end-sequencing read, requiring that only about 1 % of the genome be sequenced, and that high coverage may be obtained with a reduced amount of reads ^{303,304}. Use of the restriction enzyme *MspI* is often recommended due to its insensitivity to DNA methylation. The recognition site of this enzyme is CCGG which guarantees that a CpG will be present

at every read since each digested fragment will contain two terminal CpGs^{303,304}. The protocol may also be applied to any species since the method is not dependent on hybridisation in the same way that microarray-based technologies are. The method may also be applied to formalin- and paraffin-fixed samples^{304,305}.

3.1.1.2.3 Pyrosequencing

The most common applications of pyrosequencing are validation of differentially methylated CpG loci identified using one of the methods described above and interrogation of the methylation status of genes of interest. The method involves PCR amplification of bisulphite converted DNA using primers designed on a bisulphite converted genome. The amplicon, usually ≈100 bp, is then sequenced in a short-read sequencing-by-synthesis reaction. Whenever the correct complementary nucleotide is incorporated into the sequence, a reaction occurs where light is emitted, the intensity of which is then measured. The methylation level of each CpG site within the amplicon is estimated depending on the signal intensities of the incorporated dGTP (indicating cytosine base on amplicon) and dATP (indicating thymine base on amplicon) nucleotides. While this technique is cost-effective once the specialised equipment has been purchased, it can be labour-intensive depending on the number of CpG loci being interrogated and CpG sites more than 100 bp apart would require multiple assays. Designing successful primers for bisulphite converted DNA may also prove challenging due to reduced complexity, as is assaying repetitive elements which are prone to sequences of repeated bases. This technique allows for the measurement of small differences in methylation between samples making it possible to use on samples which are heterogeneous.

3.1.1.3 Comparing Methods

As mentioned previously, there are several other techniques that could be used to measure DNA methylation, with only the most popular/current ones mentioned here. While all the methods mentioned are highly sensitive providing data at the single nucleotide level, pyrosequencing is by far the lowest throughput of them all. The Illumina BeadChip arrays are the most cost-effective per

sample, particularly compared to WGBS, and require only small input quantities of DNA which, like RRBS, need to be of high-quality. The BeadChips however are limited with respect to them being applicable to only human samples, and to the CpG sites included in the array design. WGBS has the potential to interrogate the methylation status of the entire genome and is the most sensitive method with respect to genomic regions with low CpG density. Out of all the techniques, WGBS requires the largest amount of starting material and, along with RRBS, requires specialised bioinformatic knowledge to interpret. RRBS provides a middle-road between BeadChip arrays and WGBS in that it is both time- and cost-effective, covering more of the genome than BeadChips but reducing the redundancy and volume of data produced by WGBS. However it is important to highlight that this method has limited coverage in genomic regions with low CpG density, and shows a bias for regions with high CpG density.

3.1.2 The Effects of PAH Exposure on DNA Methylation

The effects of PAH exposure on gene expression have been studied extensively, but fewer studies have investigated the effects of exposure on DNA methylation. Most often, a representative compound such as B[a]P is used, particularly in cell line and animal model experiments. Also, the vast majority of the studies summarised below assessed methylation differences either at a global level or at specific gene loci, and only a few looked at genome-wide changes.

3.1.2.1 *In vitro*

It has been well-established that DNA adducts, specifically PAH-DNA adducts, form preferentially at guanines 3' to methylated cytosines in a CpG context^{174,306,307}. This binding is believed to lead to the spontaneous deamination of the 5-mC base resulting in mutations¹⁷⁴ and increased mutation frequency when CpG sites were methylated has been reported³⁰⁷. The authors reported a 71% G to T mutation rate when cytosines were methylated compared to 48% when they were not, and 11% G to A mutations in a methylated context compared to 4% in an unmethylated context³⁰⁷. Additionally, the various isomeric forms of the BPDE-DNA adducts discussed in section 1.2.4.1 have been reported to

exhibit conformational differences dependent both on the chemistry of the isomer in question and the CpG context ³⁰⁶.

Exposure of human bronchial epithelial cells (HBE cells) to B[a]P resulted in global hypomethylation in a dose-dependent manner, however this effect was not observed in cells which were PARG-deficient (shPARG) ³⁰⁸. This finding is supported by reports from the same authors and others of decreased methylation following B[a]P exposure as well as changes in the expression and protein levels of DNMTs ^{308–310}. HBE cells showed decreased *DNMT1* mRNA and protein expression following B[a]P exposure, but shPARG cells did not suggesting a possible role for PARG silencing of *DNMT1* ³⁰⁸. *DNMT3b* and *MBD2* expression was increased in both cell lines, but the difference compared to controls was much greater in the HBE cells ³⁰⁸. Formation of both BPDE-N²-dG and BPDE-N⁶-dA adducts results in dysregulated methylation rates when found in or adjacent to CpG sites, sites where BPDE-N²-dG adducts bind preferentially ³⁰⁹. These adducts prevented the binding of murine *Dnmt3a* which may explain the reduced methylation rates ³⁰⁹. Similar observations have been reported with prokaryotic DNMTs ^{310,311}.

In addition to decreased global methylation, observations have been made *in vitro* of altered DNA methylation at gene-specific loci. The *RAR-β2* promoter was found to be hypermethylated in oesophageal cancer cells exposed to BPDE ³¹². This observation was coupled with a decrease in gene expression levels and a time-dependent recruitment of *DNMT3a* to the promoter ³¹². More recently, a genome-wide study of methylation changes caused by B[a]P exposure in HBE cells was published ³¹³. The HBE cells were modified to express an oncogenic allele of *H-Ras* (HBER cells), and these were treated with B[a]P for 7 weeks to create a pre-transformed cell line (HBERNT cells) and for 14 weeks to create a transformed cell line (HBERT cells) ³¹³. The authors reported hypermethylation of 83 genes in both HBERNT and HBERT cell lines compared to untransformed HBER cells, but only 7 of these were associated with decreased gene expression (*CNGA4*, *FLT1*, *FZD10*, *GAREM1*, *NKX6-2*, *SFTMBT2*, and *TRIM36*) ³¹³. These genes represent early but sustained responses to B[a]P exposure, making them

useful as potential biomarkers. The methylation levels of 5 of these 7 genes were restored following treatment with the drug 5-azacytidine, a known DNA methylation inhibitor³¹³.

3.1.2.2 *In vivo*

Some animal model studies assessing the effects of PAH exposure on DNA methylation support the *in vitro* findings summarised above, however others do not. Global hypomethylation has been reported in two zebrafish studies where embryos were exposed to B[a]P^{138,314}. One of these studies also reported reduced expression of DNMTs in wild-type fish¹³⁸. *Ahr2*-null fish did not show any changes in global methylation or DNMT expression which suggests that the epigenetic observations are induced by downstream metabolites of B[a]P which require *Ahr2* expression¹³⁸ which is zebrafish equivalent of the AHR in humans. In addition to global hypomethylation, differential methylation at several disease-related genes has also been reported following B[a]P exposure³¹⁴. These genes included *c-fos*, *sox3*, and *dazl*³¹⁴. In an additional study, zebrafish embryos exposed to B[a]P were found to have 235 differentially methylated gene promoters, and some of these genes were related to adverse neurological outcomes¹³⁷. In this study, promoter hypermethylation events were found in promoters of genes that were down-regulated, while hypomethylation events were associated with up-regulation of the gene¹³⁷.

Gene-specific and genome-wide methylation analyses have been carried out in mice exposed to B[a]P^{315–318}. Mice exposed to B[a]P have shown decreased expression of the *Nr2b* gene which was associated with hypermethylation at the promoter³¹⁵. The mice also exhibited decreased short-term memory and anxiety-like behaviour which the authors linked to the dysregulation of *Nr2b*³¹⁵. Lung^{317,318} and seminal vesicle³¹⁶ tumours induced by B[a]P were profiled for genome-wide methylation changes compared to non-exposed and/or normal-adjacent tissues. In these studies, the different tissue types had distinct methylation profiles, with a number of genes identified as being differentially methylated. In the lung tumour study, the authors mentioned ten genes which stood out based on their methylation status between different tissues and that had links to tumorigenesis: *Bcl2l11*, *Bmp1*,

*Fgfr1op, Hob1, Pdc4, Casp7, Il11, Pten, Maea, and Tpd52l1*³¹⁷. Several CpG islands in the seminal vesicle study were found to be differentially methylated both in tissues from B[a]P-treated asymptomatic mice and exposed mice that developed tumours, which the authors postulate could be persistent early markers of disease³¹⁶. The authors also noted that a number of the differentially methylated genes were part of the Nanog pathway responsible for establishing and maintaining pluripotency in embryonic stem cells and which may play a role in tumorigenesis³¹⁶. Another important observation of this study, which was also made in several *in vitro* studies, is that DNMT expression was found to be reduced³¹⁶.

Only a few other studies exist from other animal models, and their findings are somewhat contradictory. Liver tissues from rats exposed to four different PAHs showed differential expression of *Dnmt3a* and *Dnmt3b* genes compared to control rats which supports other findings and may point to a global decrease in methylation¹⁰⁷. However, PAH exposure of loggerhead sea turtles in the Adriatic sea was found to be positively correlated with global methylation³¹⁹. Finally, *in ovo* exposure of chicken embryos to B[k]F1 resulted in hypermethylation of *Cyp1A4* and *Cyp1A5* genes, and decreased expression levels of *Cyp1A5* mRNA³²⁰.

Taken together, the current state of the knowledge with respect to DNA methylation changes caused by PAH exposure in cell lines and animal models is somewhat limited, particularly at the genome-wide level. However, some observations appear to be consistent: B[a]P exposure induces global hypomethylation, probably due to the down-regulation of DNMT enzymes, and gene-specific changes in DNA methylation do occur but there is no consensus as yet between studies.

3.1.3 Aims

This chapter aimed to determine the effects of B[a]P exposure on DNA methylation in the non-cancerous lung tissue of mice. A methylome-wide analysis was carried out on the results from the RRBS libraries developed from 12 mice exposed to different doses of B[a]P. The analyses were carried

out two genomic resolutions: at the individual CpG site level and at the window level where the methylation of CpG probes in 500 b.p. windows were averaged and weighted.

3.2 Results

3.2.1 Sequencing Statistics

The mouse genome contains 20,383,623 CpG sites and the twelve RRBS libraries sequenced covered between 5.5% and 7.3% of these. The sequencing statistics are summarised in Table 3.1. High amounts of inter-mouse variation were observed between the different mouse samples, with respect to the number of sequencing reads, the number of CpG sites covered, the number of 500 bp windows the sites fit into, and the average methylation levels of all sites (Figure 3.1 and Figure 3.2). The reason for this variation was not clear, since all mice were isogenic, male, and of the same age. The overall pattern of samples with respect to number of CpG sites covered and number of 500 bp windows covered was highly similar, and this pattern reflected that of the total number of reads per sample (Figure 3.1). Samples 1, 3, 9 and 10 had the lowest coverage with respect to number of CpG sites sequenced and these samples also had the lowest number of sequencing reads. No association between overall methylation levels and B[a]P exposure was observed and this was highly variable across samples within the same exposure group (Figure 3.2). The mice treated with the low dose (25 mg/kg b.w.) of B[a]P were the most consistent group, while the control group and high dose mice (75 mg/kg b.w.) showed the most variability (Figure 3.1 and Figure 3.2).

Table 3.1. Table of sequencing statistics for each mouse RRBS library.

Mouse	B[a]P Exposure Dose	CpG Site Level			500 b.p. Window Level
		Total Number of Reads	Median Number of Reads	Number of CpGs Covered	Number of Windows Covered
1	Control	33,698,309	23	1,130,837	487,284
2	Control	65,222,993	32	1,565,539	607,902
3	Control	37,210,689	23	1,197,250	475,764
4	25 mg/kg	86,419,464	38	1,692,682	646,770
5	25 mg/kg	49,341,105	27	1,400,018	572,244
6	25 mg/kg	58,858,478	30	1,495,993	569,700
7	50 mg/kg	56,864,957	30	1,474,974	550,806
8	50 mg/kg	80,587,480	37	1,619,932	556,470
9	50 mg/kg	41,526,583	26	1,180,527	429,396
10	75 mg/kg	43,365,066	27	1,199,905	397,836
11	75 mg/kg	68,403,890	33	1,592,852	607,710
12	75 mg/kg	54,103,941	28	1,423,868	597,084

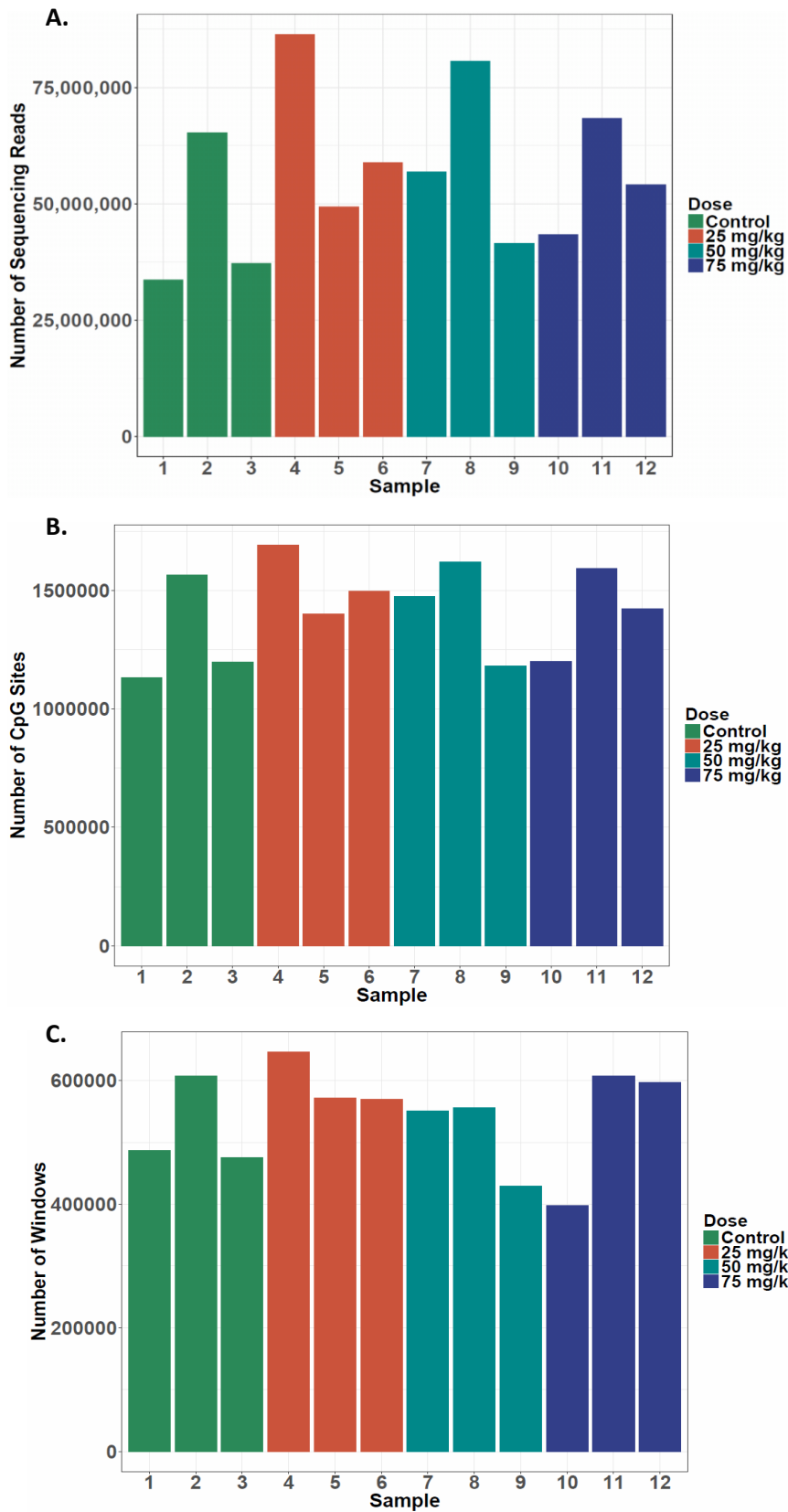


Figure 3.1 Bar charts showing the number of sequencing reads (A), the number of CpG sites (B), and the number of 500 bp windows (C) for each mouse sample coloured by B[a]P dose.

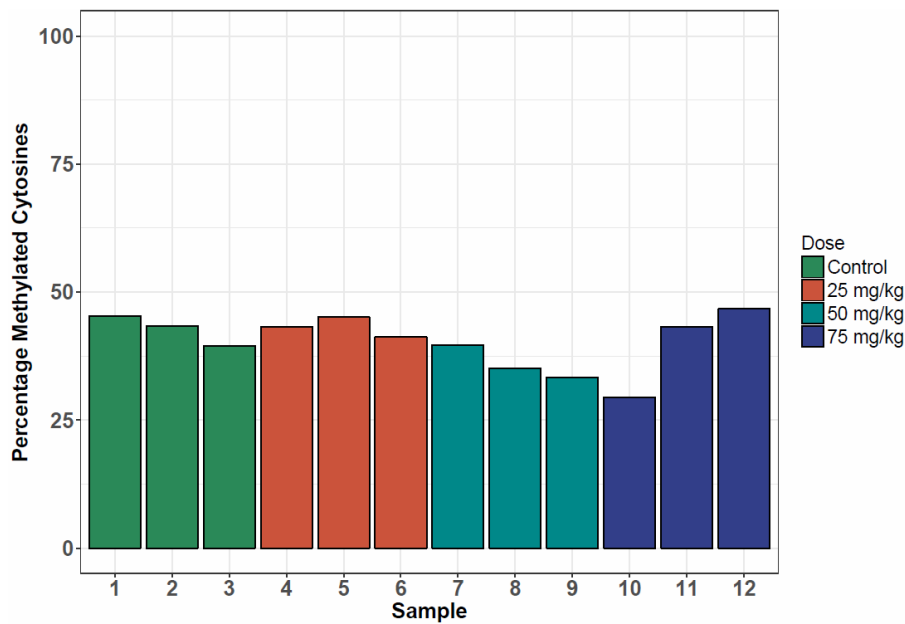


Figure 3.2. Bar chart of percentage methylated cytosine bases for each mouse sample coloured by B[a]P dose.

3.2.2 Genomic Distribution of RRBS Libraries

Each CpG site sequenced was annotated with respect to its genomic location in order to determine the genomic distribution of each sample sequenced and these were compared to the distribution of all CpG sites in the mm10 genome. Figure 3.3A shows the distribution of CpG sites in the mm10 genome, and as expected intronic, intergenic, and repeat regions had the highest proportions of CpG sites. The “other” group refers to sites which do not fit into the other genomic regions and predominantly includes short repeat sequences. Figure 3.3B shows the genomic distribution of all CpG sites sequenced from the library of sample 1 as a representative example. The genomic distribution of CpG sites was very different to that of the *mm10* genome, particularly at regulatory regions such as CpG islands, promoters and exons and this was the case for all twelve samples. This observation was expected since the restriction enzyme digestion step in the RRBS library preparation results in enrichment of genomic regions with a high density of CpG sites. This confirmed that the RRBS methodology worked as expected.

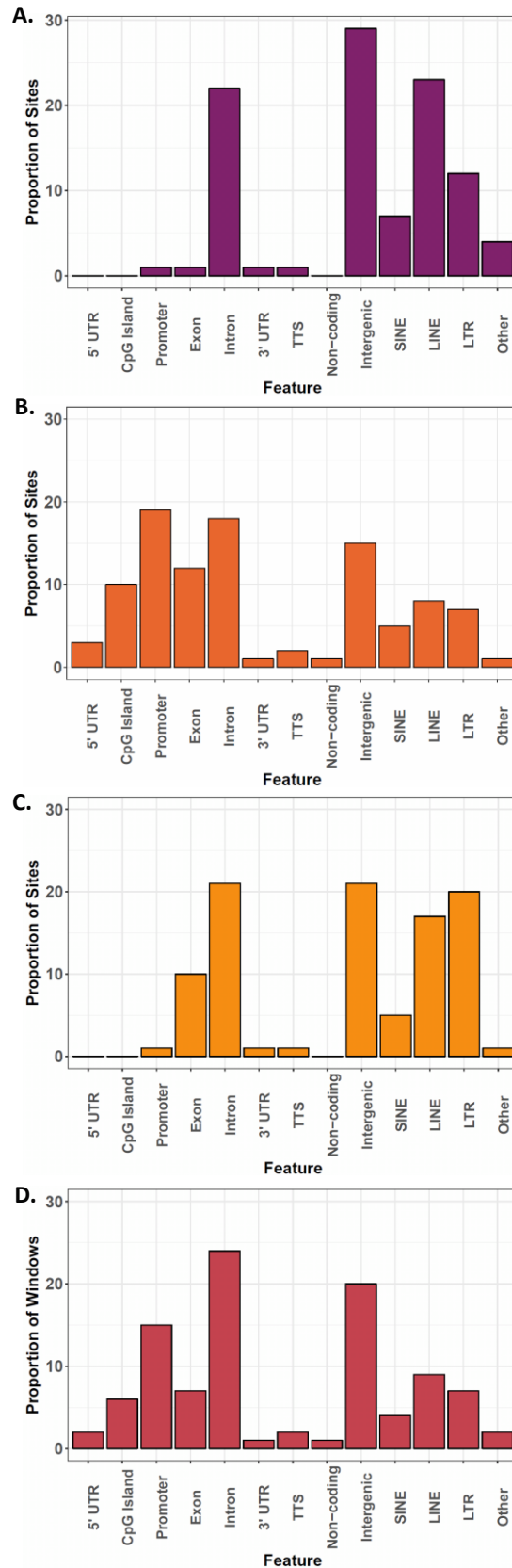


Figure 3.3 A: Bar chart showing the genomic distribution of CpG sites in the mm10 genome. B: Bar chart of genomic distribution of CpG sites from one of the RRBS libraries. C: Bar chart of the genomic distribution of the CpG sites common to all samples (N = 38,874). D: Bar chart of the genomic distribution of the 500 bp windows common to all samples (N = 152,691).

3.2.3 Inter-Sample Variation

The degree of overlap across the twelve libraries was quite low. Although each library sequenced between 1,130,837 (sample 1, control) and 1,692,682 (sample 4, 25 mg B[a]P/kg b.w.) CpG sites, only 38,874 were common to all samples. The genomic distribution of these common sites (Figure 3.3C) was more similar to that of all CpG sites in the mm10 genome than that of the RRBS libraries i.e. the majority of sites were located within introns, intergenically or within repeat regions, however the common sites distribution had a higher proportion of sites located within exons. Assigning the CpG sites to 500 bp windows did result in a higher degree of overlap across samples, however this was still quite low. The individual libraries had data for at least one CpG site within 397,837 to 646,774 windows, but only 152,691 windows were common to all samples. The genomic distribution of these common windows was more similar to that of the RRBS libraries (Figure 3.3D), however introns and intergenic regions had the highest proportions of CpG sites.

3.2.4 Principal Components Analysis

Given the high inter-sample variation, principal components analysis was carried out on the methylation data, separately for the common CpG sites and the common 500 b.p. windows. Figure 3.4A and Figure 3.4C show that both for the common windows and the common sites, the difference in percentage of variance explained by each PC was not very large with the exception of principal component (PC) 12. The maximum percentage variance explained by PC1 was the same for both methylation datasets. Separating samples by the largest PCs and colouring by B[a]P exposure dose showed that exposure did not explain the differences between the different samples (Figure 3.4B and Figure 3.4D). The scatterplot of the first 2 PCs of the common window data (Figure 3.4B) showed that the controls showed the largest variances between samples within the same dose, followed by the samples from the high exposure group (75 mg B[a]P/kg b.w.). The samples from the low exposure group (25 mg B[a]P/kg b.w.) were the most similar to each other. This pattern is similar to that observed for the overall methylation levels per sample. Figure 3.4D however, shows that when comparing the variance between samples based on the methylation of only the common CpG sites,

the samples from the medium exposure group (50 mg B[a]P/kg b.w.) had the largest variance followed by the samples from the high exposure group. Again, the samples from the low exposure group (25 mg B[a]P/ kg b.w.) had the least inter-sample variance. None of the PCs from the windows methylation dataset were significantly associated with B[a]P exposure dose, and only PC10 from the sites methylation dataset was significantly associated with B[a]P exposure dose ($p = 0.029$) however no clear separation of samples by dose was observed when separating samples by this PC (not shown), and thus this is likely a chance finding.

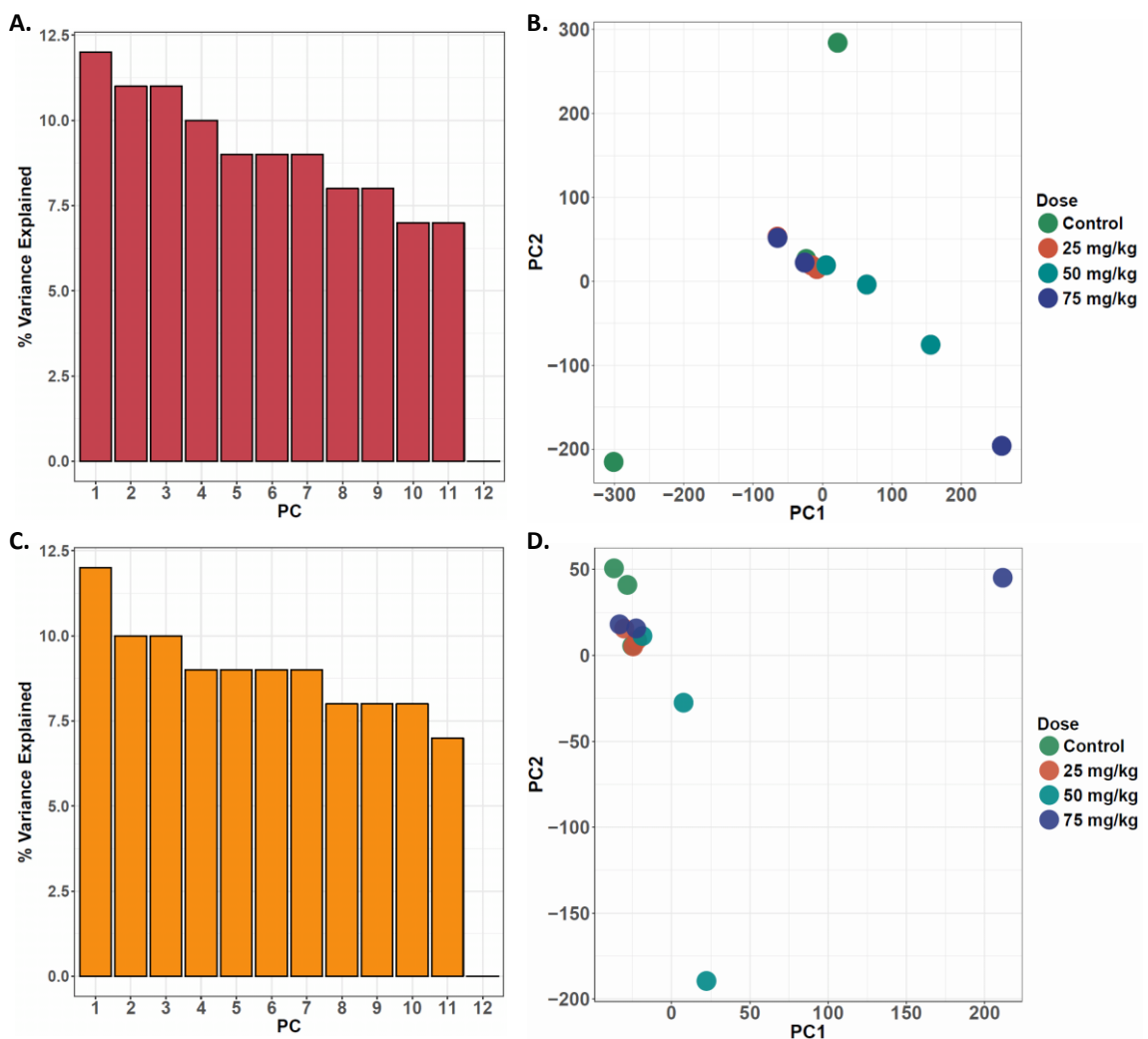


Figure 3.4 A&C: Bar charts showing the percentage variance explained by each PC for the 500 bp window (A) and the CpG site (C) PCA analyses. **B&D:** Scatterplot of the first and second PCs for the 500 bp (B) and CpG sites (D) PCA analyses coloured by B[a]P exposure dose.

3.2.5 EWAS

In order to identify differentially methylated 500 bp windows and differentially methylated CpG sites, for each of these datasets four models were run: a treated vs untreated model (3 controls vs 9 B[a]P-exposed samples), and three models comparing the controls to each of the B[a]P doses (3 controls vs 3 mice exposed to the low, medium or high B[a]P doses). The results presented here are for the treated vs untreated model from the 500 bp window analysis and the CpG site analysis, with the results for the other models briefly discussed later. Figures and tables for these models are shown in Appendix 1.

3.2.5.1 Treated vs Untreated Model EWAS Results

The CpG site analysis identified 430 differentially methylated CpG sites (DMCs) between the controls and the B[a]P-exposed samples, while the window analysis identified 1780 differentially methylated windows (DMWs) (Figure 3.5). A differentially methylated site or window was defined as one which had a p value < 0.05 and a methylation change of at least 25%. This threshold was selected since sequencing depth and sensitivity of detection are correlated, and since sequencing depth was lower than expected, a higher threshold was applied. Of the significant DMCs, 70 overlapped with 95 of the significant DMWs, with the direction of methylation change being the same in both analyses. In both sets of results, more windows and sites were hypomethylated in the B[a]P-treated samples compared to the unexposed controls (66.3% of sites in sites analysis and 62.9% in the windows analysis).

Although the DMCs and DMWs from the models allowed for successful separation of samples based on their B[a]P treatment status, the treated samples did not cluster by B[a]P dose (Figure 3.5).

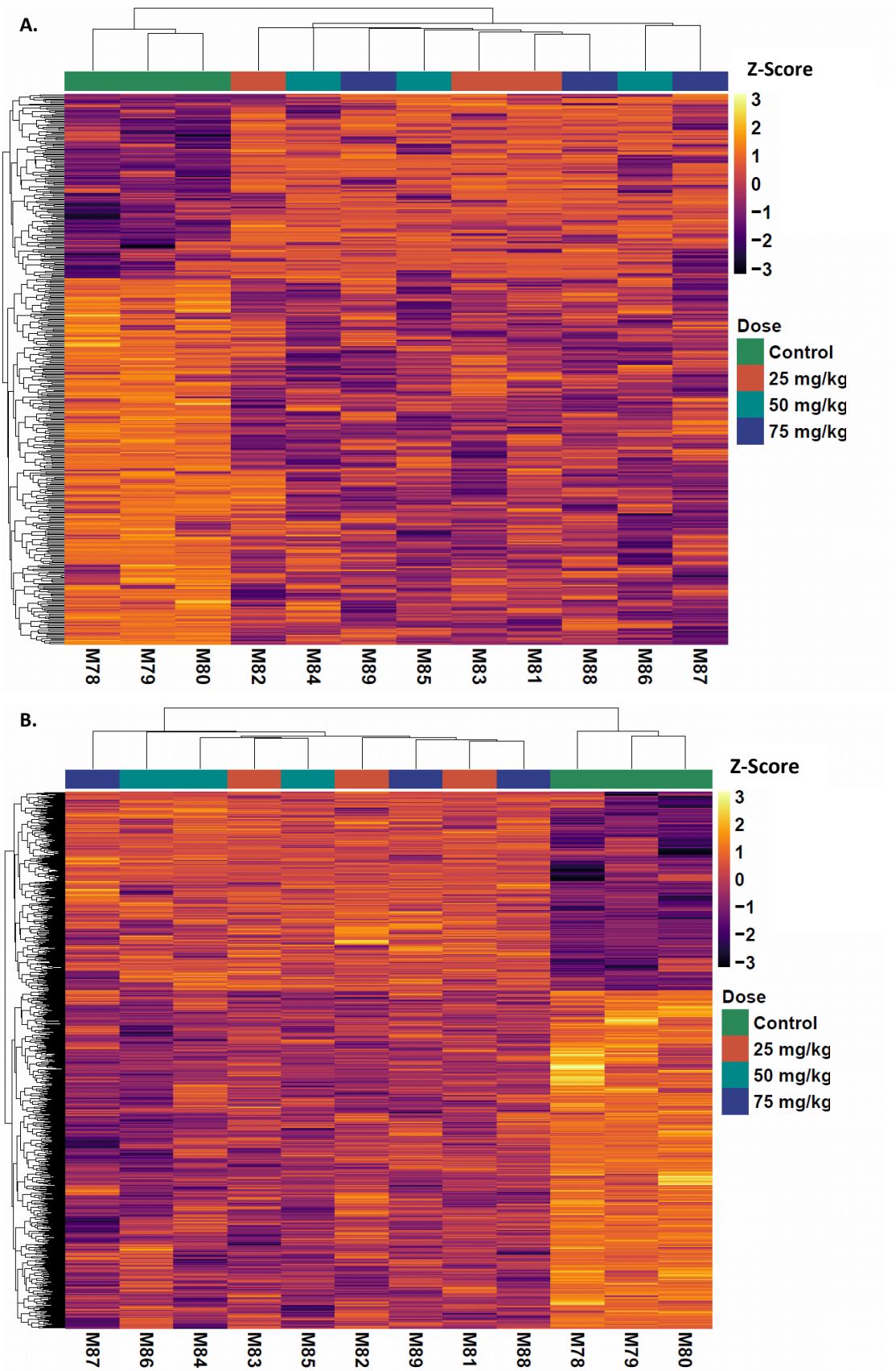


Figure 3.5 Heatmaps of the 430 differentially methylated CpG sites (A) and the 1780 differentially methylated 500 bp windows (B) from the treated vs untreated models.

3.2.5.2 Genomic Distribution of Treated vs Untreated Differences

All the DMCs and DMWs were annotated in order to identify their genomic location. Using Fisher's Exact Test, the genomic distributions of these sites and windows were compared to the genomic distribution of all sites and windows interrogated in the model (38,874 common CpG sites, and 152,720 500 bp windows). Table 3.2 shows the results from the Fisher's Exact Tests and Figure 3.6 show the proportion of sites or windows respectively expected for each annotation and the observed proportions of the significant DMCs and DMWs.. The DMCs were depleted for changes at long interspersed nuclear elements (LINE) (OR = 0.53; $p = 4.03 \times 10^{-5}$) and long terminal repeat (LTR) (OR = 0.72; $p = 0.014$) regions, and enriched for changes at 3'UTR (OR = 2.04; $p = 0.022$), intergenic (OR = 1.44; $p = 0.0010$), intronic (OR = 1.37; $p = 0.0043$), and CpG island (OR = 3.77; $p = 0.013$) regions (Table 3.2 and Figure 3.6A). The significant DMWs showed less changes than expected in 5'UTR (OR = 0.2; $p = 4.9 \times 10^{-12}$), exon (OR = 0.62; $p = 1.8 \times 10^{-5}$), promoter (OR = 0.16; $p = 5.9 \times 10^{-68}$), and CpG island (OR = 0.044; $p = 3.91 \times 10^{-40}$) regions, and more changes than expected at intergenic (OR = 1.77; $p = 2.8 \times 10^{-26}$), intronic (OR = 1.48; $p = 6.7 \times 10^{-14}$), short interspersed nuclear elements (SINE) (OR = 1.38, $p = 0.0031$), and other (OR = 1.42, $p = 0.024$) regions (Table 3.2 and Figure 3.6C). The ratio between hypo- and hypermethylation changes at each genomic region was compared to the overall ratio for all DMCs and DMWs. The results of these comparisons are summarised in Table 3.3 and Figure 3.7. At intergenic (OR = 2.21; $p = 0.0014$) and intronic (OR = 1.86; $p = 0.011$) regions, more hypomethylation events were observed than expected, while at LINEs (OR = 0.19; $p = 1.17 \times 10^{-6}$) and LTRs (OR = 0.52; $p = 0.01$) more hypermethylation events were observed for the DMCs (Table 3.3 and Figure 3.7A). Transcription terminating sites (TTS) (OR = 7.13; $p = 0.039$) had more hypomethylation changes and LINE regions (OR = 0.29, $p = 4.9 \times 10^{-7}$) had more hypermethylation changes than expected in the DMWs (Table 3.3 and Figure 3.7B).

Table 3.2 Table of Fisher’s Exact Test results comparing the number of DMWs (N = 1780) and DMCs (N = 430) to all tested windows (N = 152,720) and CpG sites (N = 38,874) at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred.

Genomic Region	Differentially Methylated Windows			Differentially Methylated CpG Sites		
	Odds Ratio	Confidence Interval	P Value	Odds Ratio	Confidence Interval	P Value
3' UTR	0.99	0.60 – 1.64	1	2.04	1.04 – 3.63	0.022
5' UTR	0.12	0.039 – 0.28	4.9×10^{-12}	0	0 – 13.27	1
Exon	0.62	0.49 – 0.78	1.8×10^{-5}	0.81	0.56 -1.14	0.27
Intergenic	1.77	1.59 – 1.96	2.8×10^{-26}	1.44	1.15 – 1.79	0.0010
Intron	1.48	1.34 – 1.64	6.7×10^{-14}	1.37	1.10 – 1.71	0.0043
Non-coding	0.78	0.33 – 1.54	0.63	1.21	0.15 – 4.49	0.68
Promoter	0.16	0.12 – 0.21	5.9×10^{-68}	1.74	0.78 – 3.35	0.12
TTS	1.34	0.94 – 1.86	0.083	0.64	0.13 – 1.91	0.64
CpG Island	0.044	0.014 – 0.10	3.91×10^{-40}	3.77	1.20 – 9.11	0.013
LINE	1.05	0.89 – 1.23	0.53	0.53	0.38 – 0.74	4.03×10^{-5}
SINE	1.38	1.11 – 1.69	0.0031	0.78	0.46 – 1.26	0.38
LTR	1.19	0.99 – 1.41	0.050	0.72	0.54 – 0.94	0.014
Other	1.42	1.03 – 1.92	0.024	0.55	0.11 – 1.62	0.39

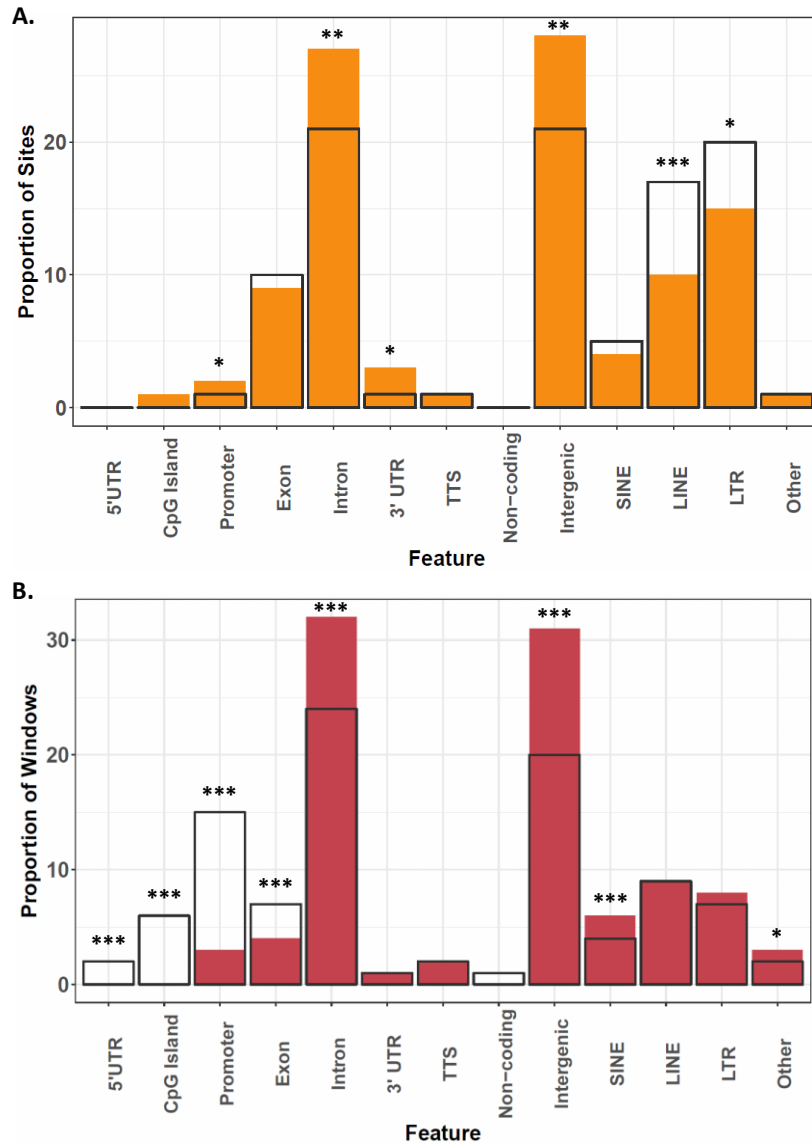


Figure 3.6 Comparison of the genomic distribution of differentially methylated sites (N = 430) and all tested sites (N = 38,874) (A) and of differentially methylated windows (N = 1780) and all tested windows (N = 152,691) (B). The coloured bars show the proportion of significant sites/windows, the grey outline bars show the proportion of all sites/windows tested, i.e. the expected distribution. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.005$ following Fisher's Exact test.

Table 3.3 Table of Fisher’s Exact Test results comparing the number of hypermethylation changes (DMWs: N = 660; DMCs: N = 145) to hypomethylation changes (DMWS: N = 1120; DMCs: N = 285) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected.

Genomic Region	Differentially Methylated Windows			Differentially Methylated CpGs		
	Odds Ratio	Confidence Interval	P Value	Odds Ratio	Confidence Interval	P Value
3' UTR	0.59	0.0075 – 46.30	1	1.54	0.38 – 8.99	0.76
5' UTR	0.59	0.0075 – 46.30	1	0	0 – Inf	1
Exon	1.31	0.57 – 3.30	0.57	1.07	0.50 – 2.41	1
Intergenic	1.13	0.53 – 1.57	0.44	2.21	1.33 – 3.78	0.0014
Intron	1.09	0.80 – 1.49	0.65	1.86	1.13 – 3.15	0.011
Non-coding	0.88	0.10 – 10.61	1	0.51	0.0064 – 40.08	1
Promoter	1.90	0.66 – 6.66	0.26	1.02	0.21 – 6.38	1
TTS	7.13	1.05 – 305.17	0.039	0.25	0.0043 – 4.89	0.26
CpG Island	Inf	0.11 – Inf	0.53	0.76	0.086 – 9.21	1
LINE	0.29	0.17 – 0.48	4.9×10^{-7}	0.19	0.088 – 0.40	1.17×10^{-6}
SINE	0.93	0.43 – 2.12	0.85	0.49	0.17 – 1.44	0.20
LTR	1.12	0.61 – 2.12	0.77	0.52	0.29 – 0.93	0.021
Other	1.03	0.84 – 1.26	0.80	Inf	0.21 - Inf	0.55

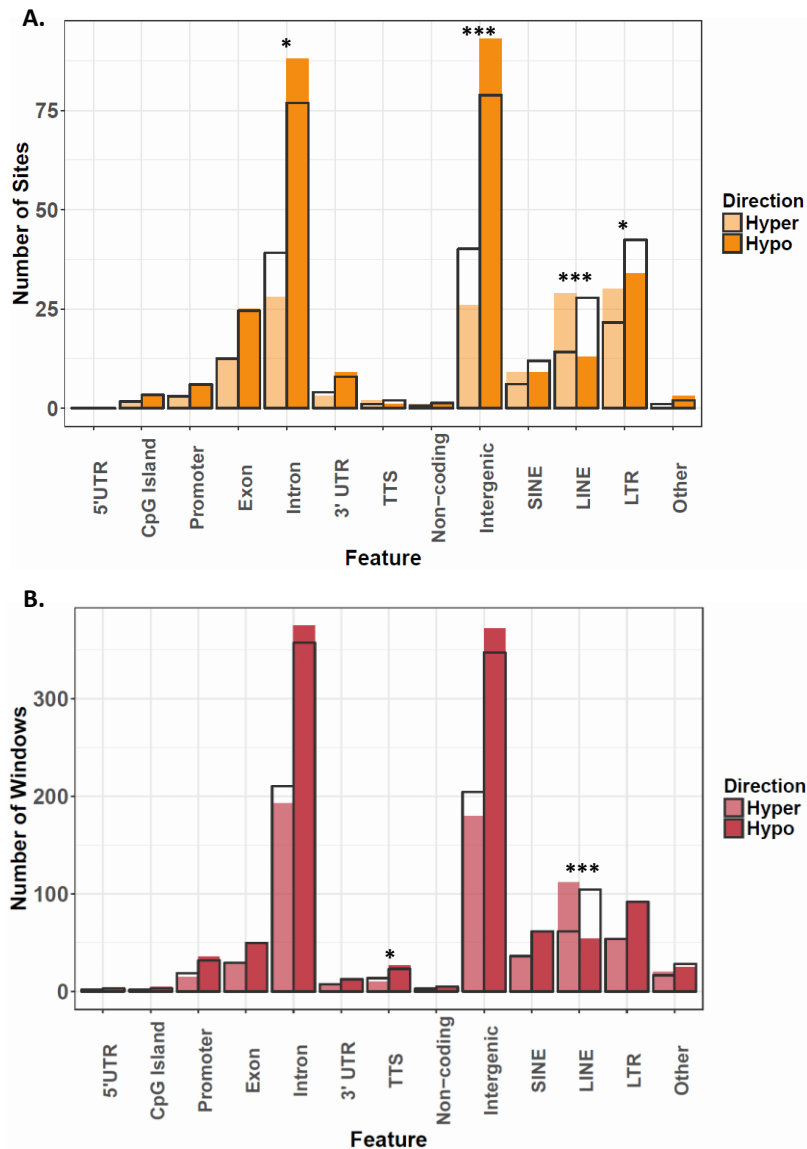


Figure 3.7 Comparison of the genomic distribution of hypermethylated (N = 145) and hypomethylated (N = 285) sites (A), and hypermethylated (N = 660) and hypomethylated (N = 1120) windows (B). The coloured bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.005$ following Fisher's Exact test.

3.2.5.3 Other EWAS Model Results Comparison

The models comparing the controls to each dose separately gave more DMCs and DMWs than the treated vs untreated models and the results are summarised in Table 3.4. Interestingly, the controls vs medium dose models had the largest number of DMCs and DMWs. The number of observed results overlaps across the models were quite low in both sets of analyses (Figure 3.8). The models that compared the controls to each dose individually tended to identify sites or windows which were unique to that level of B[a]P exposure, whereas the treated vs untreated model results were more representative of each of the individual comparisons. The site analysis identified 47 CpG sites that were consistently differentially methylated across all models, and the 500 bp windows analysis found 140 windows that were significant in all models. The degree of overlap was not improved when looking at changes at sites or windows within the same genes instead of matching the exact locus or window of change.

Table 3.4 Table summarising the number of differentially methylated CpG sites and differentially methylated windows for each model.

		Treated vs Untreated	Controls vs Low dose	Controls vs Medium dose	Controls vs High dose
DMCs	Total	430	699	768	664
	Hypomethylated	285	478	505	416
	Hypermethylated	145	221	263	248
DMWs	Total	1780	1910	2671	1952
	Hypomethylated	1120	993	2087	1139
	Hypermethylated	660	917	584	813

3.2.5.3.1 Overlapping DMCs and Gene Expression

The 47 CpG sites that were found to be significantly differentially methylated across all models are located primarily in intergenic, intronic and repeat element regions (Table 3.5). For all of these sites, the direction of change was always consistent for all models, however, in some cases the magnitude of difference differed slightly, and the significance level was highly variable between models (Table 3.5). Eight of these CpG sites were located in genic regions, one in the 3' UTR and the rest within

exons (Table 3.6). Gene expression data generated by Labib *et al.* (2012)¹⁴² from an Agilent 4x44K oligonucleotide microarray were available for six of the eight genes associated with these CpG sites. Spearman's correlation between the gene expression and DNA methylation data were carried out for all available transcripts of these genes however no significant correlations were observed (Table 3.6).

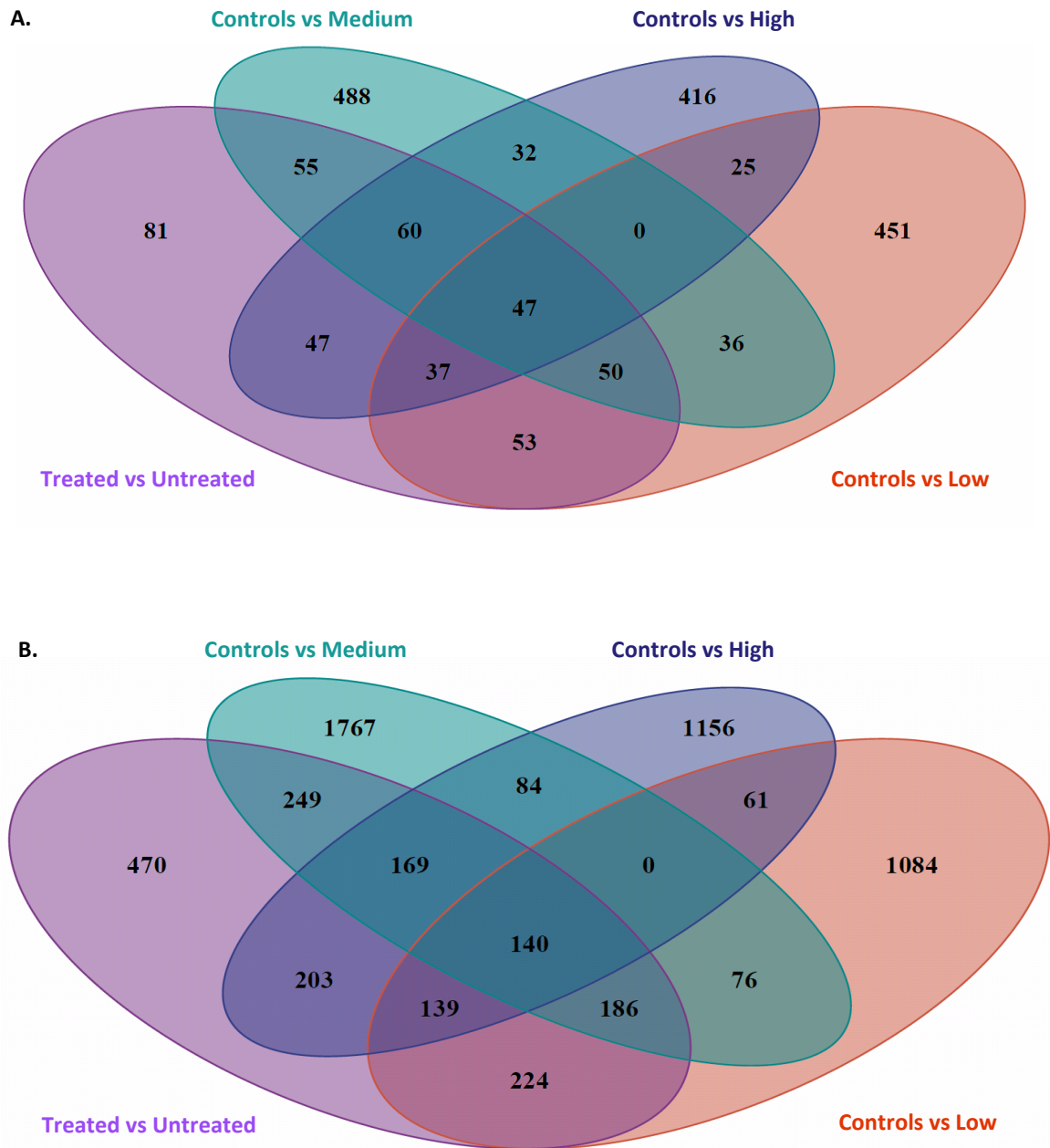


Figure 3.8 Venn diagrams showing the overlaps between the four models run for the sites analysis (A) and the 500 bp windows analysis (B).

Table 3.5 Summary of characteristics and results of significant DMCs across all models.

<u>Chromosome</u>	<u>Position</u>	<u>Genomic Location</u>	<u>Gene Name</u>	<u>Treated vs Untreated</u>		<u>Controls vs Low</u>		<u>Controls vs Medium</u>		<u>Controls vs High</u>	
				<u>Δ Methylation (%)</u>	<u>P Value</u>	<u>Δ Methylation (%)</u>	<u>P Value</u>	<u>Δ Methylation (%)</u>	<u>P Value</u>	<u>Δ Methylation (%)</u>	<u>P Value</u>
chr19	48034775	3' UTR	<i>Cfap58</i>	-46.21	9.0E-04	-53.74	1.9E-03	-55.36	4.5E-03	-29.53	2.0E-02
chr18	31947192	Exon	<i>Lims2</i>	30.23	1.3E-03	28.29	3.6E-02	30.25	9.6E-03	32.16	6.3E-12
chr15	92234452	Exon	<i>Cnfn1</i>	-36.76	9.5E-04	-27.04	2.8E-02	-41.69	5.3E-03	-41.56	2.6E-03
chr9	37417751	Exon	<i>Robo3</i>	32.18	2.8E-03	33.73	8.0E-17	29.09	4.2E-02	33.73	5.7E-19
chr12	102403463	Exon	<i>Lgmn</i>	26.27	2.1E-06	26.43	4.6E-03	26.43	3.3E-03	25.95	4.6E-03
chr14	49298867	Exon	<i>Slc35f4</i>	-36.78	1.8E-03	-43.97	6.8E-15	-29.59	3.3E-02	-36.77	4.8E-02
chr9	14703138	Exon	<i>Piwil4</i>	-49.73	1.5E-05	-51.70	1.7E-31	-61.67	1.3E-03	-35.82	2.2E-19
chr15	76191065	Exon	<i>Plec</i>	43.71	7.4E-25	43.71	1.8E-13	43.71	1.5E-12	43.71	5.3E-14
chr16	90308599	Intergenic		31.89	3.1E-03	33.50	1.3E-11	28.68	4.9E-02	33.50	6.3E-17
chr4	125248927	Intergenic		49.04	2.3E-05	49.04	2.4E-02	49.04	4.2E-02	49.04	2.8E-02
chr12	95628973	Intergenic		-39.74	2.4E-04	-39.81	8.5E-03	-44.83	2.2E-03	-34.59	5.0E-03
chr13	48431703	Intergenic		-39.14	5.1E-04	-25.66	9.3E-03	-56.25	2.6E-34	-35.50	1.7E-02
chr18	57080283	Intergenic		-42.03	4.4E-05	-31.91	8.0E-12	-43.55	9.0E-17	-50.62	3.4E-03
chr3	143797069	Intergenic	<i>Gm42705</i>	-35.40	3.4E-05	-33.60	7.4E-03	-43.56	3.7E-25	-29.05	1.6E-15
chr14	118663419	Intron	<i>Abcc4</i>	27.75	5.3E-05	27.75	3.8E-02	27.75	4.5E-02	27.75	2.8E-02
chr10	80888831	Intron	<i>Tmprss9</i>	37.74	1.0E-26	37.74	1.4E-17	37.74	2.6E-12	37.74	1.5E-16
chr1	78132122	Intron	<i>Pax3</i>	-38.85	3.5E-04	-28.57	2.0E-11	-33.61	8.5E-15	-54.38	4.1E-23
chr14	77707363	Intron	<i>Enox1</i>	-41.08	6.5E-04	-41.14	2.6E-02	-50.72	2.6E-03	-31.38	6.9E-15

chr5	64827327	Intron	<i>Klf3</i>	32.62	2.8E-05	32.80	9.8E-03	32.27	1.3E-02	32.80	8.7E-03
chr19	44818920	Intron	<i>Pax2</i>	-40.00	2.2E-04	-42.91	1.1E-02	-46.26	1.2E-02	-30.84	3.4E-02
chr19	25566174	Intron	<i>Dmrt1</i>	34.14	4.2E-03	34.91	2.7E-02	39.08	8.7E-14	28.42	1.3E-02
chr13	46615383	Intron	<i>Cap2</i>	36.70	8.0E-06	37.16	9.7E-13	36.08	4.0E-03	36.85	1.8E-09
chr5	116519253	Intron	<i>Srrm4</i>	-37.37	9.7E-04	-39.69	1.9E-02	-42.10	1.2E-02	-30.33	4.8E-03
chr5	98002765	Intron	<i>Antxr2</i>	-47.07	1.7E-04	-44.70	1.6E-02	-44.90	4.0E-02	-51.62	2.0E-02
chr13	47916746	Intron	<i>G630093K05Rik</i>	-32.02	4.9E-03	-27.75	2.6E-02	-38.91	2.7E-02	-29.39	1.1E-02
chr12	49486435	LINE		40.68	9.4E-06	40.68	1.3E-02	40.68	2.6E-02	40.68	1.7E-02
chr12	60278368	LINE		36.52	2.7E-06	36.52	8.3E-03	36.52	1.0E-02	36.52	2.2E-02
chr13	50403278	LINE	<i>Gm8739</i>	56.28	3.0E-03	58.55	3.4E-36	51.74	2.6E-02	58.55	8.9E-30
chr14	51190583	LINE		47.96	1.5E-03	46.96	4.1E-02	51.55	9.7E-03	45.36	2.4E-03
chr17	22175541	LINE		29.72	5.4E-05	29.72	3.7E-02	29.72	3.9E-02	29.72	2.2E-02
chr3	29557233	LINE	<i>Egfem1</i>	27.93	1.3E-04	28.03	2.4E-02	27.72	3.4E-02	28.03	2.0E-02
chr3	70046864	LINE		32.74	1.5E-04	32.75	1.1E-02	33.59	4.5E-03	31.87	1.5E-02
chr4	15020242	LINE	<i>Necab1</i>	29.69	1.3E-05	29.38	1.0E-02	29.84	5.5E-03	29.84	9.5E-03
chr4	144762711	LINE		26.81	2.7E-05	26.81	2.6E-02	26.81	2.9E-02	26.81	3.1E-02
chr9	90285341	LINE		27.28	2.9E-06	27.28	1.2E-02	27.28	8.8E-03	27.28	1.2E-02
chrX	104628878	LINE	<i>Zdhhc15</i>	27.93	3.8E-04	29.89	8.9E-03	27.12	2.5E-02	26.78	2.8E-02
chr6	51012286	LINE		-39.73	5.1E-03	-36.79	3.8E-02	-39.31	5.7E-03	-43.09	1.8E-02
chr1	95894667	LTR		40.23	6.9E-04	33.64	1.4E-02	43.52	2.6E-28	43.52	1.6E-25
chr1	100783796	LTR		27.31	7.8E-07	27.31	7.8E-03	27.31	7.7E-03	27.31	5.6E-03
chr11	116720787	LTR		39.12	3.1E-05	40.24	2.2E-03	38.41	2.4E-03	38.71	1.0E-02
chr13	65982935	LTR		29.73	7.4E-06	29.73	1.1E-02	29.73	1.5E-02	29.73	1.7E-02
chr16	46267988	LTR		26.23	3.6E-06	26.23	9.0E-03	26.23	1.5E-02	26.23	1.2E-02

chr2	33389625	LTR		28.03	3.5E-16	28.03	1.5E-02	28.03	9.9E-03	28.03	8.5E-03
chr8	10345321	Other	<i>Myo16</i>	-40.08	4.0E-04	-36.35	1.3E-03	-35.70	5.0E-02	-48.18	7.3E-17
chr7	14950525	SINE		32.45	5.8E-03	28.92	1.3E-05	29.85	2.9E-05	38.57	4.8E-02
chr7	45218386	SINE	<i>Tead2</i>	33.80	9.3E-05	33.54	2.5E-02	34.00	2.2E-02	33.87	2.2E-02
chr1	131987623	SINE	<i>Gm29103</i>	-50.75	1.8E-03	-32.14	3.5E-02	-52.58	1.9E-02	-67.51	2.2E-26

Table 3.6. Table showing summary of correlation between gene expression and methylation levels of DMCs. The gene expression data were generated from Agilent 4x44K oligonucleotide microarrays as part of the study carried out by Labib *et al.*¹⁴². The results for different microarray transcripts for the same gene are separated by a semicolon.

Chromosome	Position	Genomic Location	Gene Name	Gene Expression Data	Number of Transcripts	Correlation Coefficients	P values
chr19	48034775	3' UTR	<i>Cfap58</i>				
chr18	31947192	Exon	<i>Lims2</i>	✓	1	0.35	0.29
chr15	92234452	Exon	<i>Cntn1</i>	✓	2	-0.45; -0.48	0.16; 0.14
chr9	37417751	Exon	<i>Robo3</i>	✓	1	0.07; 0.11	0.85; 0.74
chr12	102403463	Exon	<i>Lgmn</i>	✓	2	0.27	0.43
chr14	49298867	Exon	<i>Slc35f4</i>	✓	3	0.06; -0.49; -0.07	0.85; 0.12; 0.83
chr9	14703138	Exon	<i>Piwil4</i>	✓	2	-0.15; -0.19	0.67; 0.57
chr15	76191065	Exon	<i>Plec</i>				

3.2.5.3.2 Overlapping DMWs and Gene Expression

Similar observations were made for the 140 differentially methylated 500 bp windows (Table 3.7). The vast majority of these windows are located intergenically, intronically, or within repeat elements. The direction of methylation change was the same for all four models, but magnitude and significance were found to vary between models (Table 3.7). Only eleven windows were associated with genic regions: one in the 5' UTR, three in exons, two in promoters, and five in transcription termination sites (TTS) (

Table 3.8). Gene expression data were available for seven of the genes associated with these DMWs and correlation between methylation and gene expression was analysed (

Table 3.8). Only one transcript of the D5Erttd579e gene was found to be significantly positively correlated with methylation levels in the window located in the promoter of this gene (

Table [3.8](#)). Visualisation of this relationship shows that, in general, the controls had higher methylation and higher expression than the B[a]P-exposed samples (Figure 3.9).

3.2.5.3.3 Genomic Distribution

Despite the lack of overlap across models of DMCs and DMWs, particular genomic regions were shown to have consistently more (enrichment) or less (depletion) methylation changes than expected by chance (Figure 3.10). From the CpG sites models, LINEs and LTRs had significantly less changes than expected across all models (Figure 3.10). The results of the 500 bp window models show that 5'UTR, CpG island, promoter and exon regions all had significantly less methylation changes than expected, and "other" regions had significantly more changes than expected across all models (Figure 3.10). Significantly more methylation changes occurred in intronic and intergenic regions across all models, both for the CpG site and the 500 bp window analysis (Figure 3.10).

Table 3.7 Summary of characteristics and results of significant DMWs across all models. The position column indicates the position of the first base in each 500 bp window.

<u>Chromosome</u>	<u>Position</u>	<u>Genomic Location</u>	<u>Gene Name</u>	<u>Treated vs Untreated</u>		<u>Controls vs Low</u>		<u>Controls vs Medium</u>		<u>Controls vs High</u>	
				<u>Δ Methylation (%)</u>	<u>P Value</u>	<u>Δ Methylation (%)</u>	<u>P Value</u>	<u>Δ Methylation (%)</u>	<u>P Value</u>	<u>Δ Methylation (%)</u>	<u>P Value</u>
chr19	7295001	5' UTR	<i>Mark2</i>	-50.00	6.1E-03	-32.46	3.7E-02	-53.56	3.7E-03	-63.99	1.1E-02
chr7	101504501	CpG Island	<i>Pde2a</i>	-38.79	3.8E-04	-39.32	1.2E-02	-35.02	1.5E-02	-42.02	8.6E-03
chr2	131953001	Exon	<i>Prn</i>	-34.22	5.5E-04	-29.21	3.2E-03	-36.80	2.3E-03	-36.65	1.0E-02
chr17	56098751	Exon	<i>Hdgfl2</i>	38.52	6.2E-04	35.57	2.9E-03	38.96	1.1E-02	41.01	2.6E-02
chr4	43519251	Exon	<i>Tpm2</i>	-32.45	2.7E-03	-39.15	2.2E-02	-25.77	2.7E-02	-32.44	3.8E-02
chr11	57617251	Intergenic		38.72	1.3E-03	38.84	3.4E-02	39.97	1.2E-02	37.34	5.0E-02
chr16	90308501	Intergenic		30.96	1.8E-03	32.57	2.8E-12	27.76	4.1E-02	32.57	1.4E-18
chr17	5817751	Intergenic	<i>Gm26595</i>	47.35	2.0E-03	72.71	1.7E-02	30.57	2.8E-02	38.76	1.4E-02
chr17	88102751	Intergenic		33.43	5.2E-04	30.44	1.2E-27	27.01	3.9E-02	42.86	1.1E-02
chr4	136134251	Intergenic		29.47	3.7E-03	35.22	2.0E-02	26.55	3.0E-02	26.65	3.9E-02
chr5	119066501	Intergenic		59.77	8.4E-04	54.38	2.9E-02	57.60	9.9E-03	67.32	7.3E-04
chr6	93672001	Intergenic		33.09	2.1E-04	34.88	1.9E-03	26.88	1.1E-02	37.49	2.6E-02
chr7	142764251	Intergenic		73.08	3.7E-05	68.63	4.9E-40	70.09	4.1E-03	80.54	3.7E-03
chr8	34703251	Intergenic		40.10	5.7E-06	43.59	2.4E-04	41.46	4.9E-37	35.24	2.6E-03
chr8	57142251	Intergenic		38.45	3.1E-03	43.31	3.5E-02	36.32	4.3E-02	35.70	4.5E-02
chr8	86757501	Intergenic		27.13	6.7E-04	25.94	2.0E-02	30.08	6.3E-03	25.38	7.3E-03
chr8	92374751	Intergenic	<i>Irx5</i>	42.32	6.1E-03	30.97	4.9E-02	44.71	1.0E-02	51.28	8.9E-03
chr8	122238001	Intergenic	<i>Gm20388</i>	40.56	5.7E-04	45.52	2.0E-02	36.00	1.3E-02	40.18	1.6E-02
chr9	31583751	Intergenic		31.02	7.9E-04	28.80	2.0E-02	36.19	3.6E-03	28.05	2.1E-02
chr10	72099001	Intergenic	<i>Gm34609</i>	-56.56	9.2E-04	-67.54	3.4E-03	-53.21	4.9E-02	-48.92	8.4E-03

chr10	72099251	Intergenic	<i>Gm34609</i>	-56.56	9.2E-04	-67.54	3.4E-03	-53.21	4.9E-02	-48.92	8.4E-03
chr12	13100751	Intergenic		-53.00	7.8E-03	-41.20	9.5E-03	-63.19	3.9E-02	-54.61	3.1E-02
chr12	95628501	Intergenic		-39.43	9.9E-04	-39.26	1.1E-02	-44.75	5.8E-03	-34.27	5.7E-03
chr13	57945501	Intergenic		-63.16	2.6E-03	-65.49	2.2E-02	-49.41	4.7E-02	-74.57	1.4E-02
chr13	97523751	Intergenic	<i>Gm41030</i>	-53.13	3.1E-03	-46.40	3.2E-03	-42.03	5.0E-03	-70.97	1.4E-02
chr13	98492501	Intergenic		-51.75	2.4E-03	-56.77	4.5E-02	-55.90	2.7E-02	-42.57	4.4E-02
chr18	13349251	Intergenic		-41.66	1.7E-03	-47.18	3.0E-02	-36.76	3.9E-02	-41.03	3.7E-02
chr18	26294251	Intergenic		-31.69	1.7E-03	-33.01	2.6E-02	-29.25	2.1E-02	-32.81	2.2E-03
chr3	149057251	Intergenic		-44.96	1.5E-03	-59.12	8.2E-03	-35.52	1.6E-02	-40.23	1.8E-03
chr3	149057501	Intergenic		-48.28	2.7E-03	-68.26	5.0E-03	-36.50	4.9E-03	-40.07	9.7E-03
chr4	7538751	Intergenic		-34.91	1.9E-03	-31.48	1.1E-02	-46.75	4.1E-02	-26.51	2.7E-02
chr4	7539001	Intergenic		-34.91	1.9E-03	-31.48	1.1E-02	-46.75	4.1E-02	-26.51	2.7E-02
chr4	133348001	Intergenic	<i>Wdtd1</i>	-40.82	2.4E-04	-36.78	2.4E-02	-44.60	5.8E-04	-41.10	3.7E-04
chr5	113516251	Intergenic	<i>Wscd2</i>	-47.84	7.1E-03	-32.98	4.6E-02	-44.67	2.9E-03	-65.87	9.7E-03
chr5	148684001	Intergenic		-65.25	1.1E-03	-63.81	8.6E-03	-63.87	4.9E-02	-68.06	1.5E-02
chr7	65424001	Intergenic	<i>Tjp1</i>	-41.24	8.9E-04	-35.12	3.8E-02	-48.80	4.0E-02	-39.81	2.7E-02
chr7	123836251	Intergenic		-35.69	4.3E-03	-25.67	3.4E-02	-29.89	5.8E-03	-51.52	1.9E-02
chr7	131772501	Intergenic	<i>Fgfr2</i>	-70.02	6.7E-03	-55.32	2.4E-02	-79.55	1.4E-02	-75.20	6.7E-03
chr7	131772751	Intergenic	<i>Fgfr2</i>	-70.02	6.7E-03	-55.32	2.4E-02	-79.55	1.4E-02	-75.20	6.7E-03
chr1	89557251	Intron	<i>Agap1</i>	-43.25	2.3E-03	-37.94	5.3E-03	-45.25	1.9E-27	-46.55	4.4E-02
chr11	120949751	Intron	<i>Slc16a3</i>	45.12	1.3E-02	61.64	1.9E-03	29.44	1.6E-02	44.29	1.6E-02
chr9	114550001	Intron	<i>Trim71</i>	-59.71	1.5E-03	-56.83	3.3E-02	-59.14	4.4E-02	-63.15	1.4E-02
chr1	62720751	Intron	<i>Nrp2</i>	-45.20	6.7E-04	-46.53	1.7E-02	-48.61	4.7E-02	-40.46	2.6E-02
chr9	105693001	Intron	<i>Col6a6</i>	49.81	6.0E-03	35.78	3.6E-02	59.89	1.4E-02	53.76	2.7E-02

chr4	154072751	Intron	<i>Trp73</i>	-27.46	1.5E-06	-27.34	7.2E-03	-27.52	4.5E-03	-27.52	7.5E-03
chr2	57108751	Intron	<i>Nr4a2</i>	-36.55	1.3E-03	-33.09	1.4E-02	-41.84	1.1E-02	-34.73	6.9E-03
chr5	114346751	Intron	<i>Myo1h</i>	-51.82	7.3E-04	-38.33	1.6E-02	-59.24	1.4E-04	-57.90	1.2E-03
chr6	113379001	Intron	<i>Arpc4</i>	38.54	1.1E-02	42.16	8.5E-03	26.51	4.0E-03	46.94	3.6E-02
chr9	116596251	Intron	<i>Rbms3</i>	47.92	2.2E-04	49.18	1.7E-02	49.18	2.4E-02	45.40	4.1E-02
chr7	101504251	Intron	<i>Pde2a</i>	-39.57	6.7E-04	-37.70	1.1E-02	-39.56	1.1E-02	-41.45	2.0E-02
chr3	103619751	Intron	<i>Syt6</i>	54.46	1.0E-02	52.13	3.8E-03	85.80	6.1E-03	25.44	9.3E-03
chr4	133873001	Intron	<i>Rps6ka1</i>	-63.84	1.7E-04	-57.19	4.4E-03	-57.62	1.5E-02	-76.71	1.6E-03
chr5	123603251	Intron	<i>Clip1</i>	-43.64	5.7E-04	-37.91	4.7E-02	-49.45	4.4E-02	-43.55	4.1E-02
chr19	7294751	Intron	<i>Mark2</i>	-46.32	1.0E-02	-29.76	4.8E-02	-50.29	6.5E-03	-58.92	2.6E-02
chrX	69747751	Intron	<i>Aff2</i>	29.18	7.9E-06	29.18	1.9E-02	29.18	1.3E-02	29.18	4.3E-02
chr3	131236751	Intron	<i>Hadh</i>	47.99	1.3E-03	57.49	1.0E-02	41.44	1.4E-02	45.05	4.0E-02
chr15	86027751	Intron	<i>Celsr1</i>	-43.29	4.3E-03	-44.66	4.6E-03	-51.30	1.6E-03	-33.91	4.4E-02
chr11	100393751	Intron	<i>Jup</i>	-37.73	3.1E-04	-37.41	2.9E-02	-39.36	3.9E-03	-36.42	3.2E-02
chr19	8943501	Intron	<i>Mta2</i>	-30.80	1.0E-03	-28.95	4.4E-02	-31.72	1.9E-02	-31.72	1.4E-02
chr7	96643251	Intron	<i>Tenm4</i>	-50.26	2.1E-03	-40.89	7.2E-03	-50.18	2.7E-02	-59.69	3.0E-03
chr7	96643501	Intron	<i>Tenm4</i>	-35.73	5.1E-03	-26.88	4.7E-02	-38.12	3.1E-02	-42.19	8.0E-03
chr16	8643001	Intron	<i>Pmm2</i>	32.05	1.3E-04	35.15	4.7E-03	26.35	5.1E-03	34.64	1.5E-14
chr15	100762501	Intron	<i>Slc4a8</i>	-32.91	1.7E-03	-33.55	2.9E-02	-35.54	2.9E-02	-29.64	2.2E-02
chr15	100762751	Intron	<i>Slc4a8</i>	-34.45	1.9E-03	-34.22	2.0E-02	-40.08	1.6E-02	-29.05	2.3E-02

chr6	142366001	Intron	<i>Recql</i>	-38.50	1.0E-02	-48.56	3.9E-02	-31.61	9.2E-03	-35.33	1.0E-02
chr17	56164501	Intron	<i>Tnfaip8l1</i>	45.04	5.0E-04	57.17	4.3E-04	34.03	2.3E-02	43.92	8.9E-03
chr8	13061251	Intron	<i>Proz</i>	-33.45	6.7E-03	-36.36	1.7E-02	-31.28	2.7E-02	-32.72	5.5E-03
chr1	73957001	Intron	<i>Tns1</i>	-53.46	7.9E-04	-50.24	6.1E-03	-70.25	7.1E-03	-39.87	4.7E-03
chr1	73957251	Intron	<i>Tns1</i>	-55.44	2.0E-04	-52.77	1.9E-03	-68.98	9.5E-03	-44.58	2.6E-03
chr5	149581251	Intron	<i>Wdr95</i>	-30.08	9.0E-05	-32.85	1.2E-03	-31.56	3.9E-17	-25.82	6.6E-14
chr10	60826501	Intron	<i>Unc5b</i>	-30.79	2.6E-04	-32.73	1.1E-02	-32.56	3.9E-04	-27.08	1.2E-02
chr11	34518751	Intron	<i>Dock2</i>	46.94	6.0E-04	43.82	1.6E-02	54.30	9.9E-03	42.69	1.9E-02
chr6	113479751	Intron	<i>Il17rc</i>	-32.26	3.0E-04	-27.07	1.6E-02	-39.80	1.6E-02	-29.93	1.1E-02
chr6	113480001	Intron	<i>Il17rc</i>	-32.79	3.1E-04	-25.92	2.1E-02	-42.46	2.1E-02	-29.99	1.5E-02
chr7	127591001	Intron	<i>Gm44759</i>	38.62	4.7E-04	31.08	1.8E-02	42.26	2.4E-02	42.53	1.0E-02
chr12	56897251	Intron	<i>Slc25a21</i>	55.94	3.8E-03	52.90	1.7E-02	62.26	4.6E-02	52.66	4.4E-02
chr1	106983751	Intron	<i>Serpinb13</i>	57.07	3.3E-03	49.20	2.0E-29	90.94	3.9E-03	31.07	2.1E-02
chr11	79649001	Intron	<i>Rab11fip4</i>	-40.99	4.8E-03	-58.14	3.6E-02	-25.76	3.7E-03	-39.06	7.8E-03
chr1	25679751	Intron	<i>Adgrb3</i>	49.99	3.1E-03	32.50	8.8E-11	71.03	1.3E-02	46.45	2.0E-03
chr7	142354251	Intron	<i>lfitm10</i>	79.06	2.4E-04	86.18	1.1E-04	86.18	2.4E-04	64.84	1.0E-02
chr7	6202751	Intron	<i>Galp</i>	35.29	1.3E-03	30.06	1.6E-03	36.16	5.8E-04	39.63	4.7E-02
chr15	82637751	Intron		34.34	6.3E-05	35.11	3.9E-04	32.46	7.7E-03	35.46	8.8E-05
chr3	146375001	Intron	<i>Gm10636</i>	-42.08	8.9E-04	-34.34	4.1E-02	-46.75	5.3E-03	-45.15	5.0E-52
chr3	146375251	Intron	<i>Gm10636</i>	-34.54	5.8E-04	-28.68	3.9E-02	-43.98	2.4E-44	-30.97	6.0E-04
chr5	148684251	LINE		-65.38	7.4E-04	-64.80	1.2E-02	-64.75	2.9E-02	-66.60	1.8E-02

chr8	26617501	LINE		-36.75	2.8E-03	-49.13	2.4E-03	-29.10	3.5E-02	-32.01	2.6E-02
chr9	25528501	LINE	<i>Eepd1</i>	74.77	6.9E-08	74.77	3.6E-03	74.77	1.7E-03	74.77	4.8E-03
chr4	27119751	LINE		40.26	1.3E-04	41.84	1.9E-02	40.91	1.9E-02	38.04	2.8E-02
chr5	10059001	LINE		32.14	3.6E-05	32.14	4.7E-02	32.14	2.1E-02	32.14	2.8E-02
chr9	25528751	LINE	<i>Eepd1</i>	41.09	1.3E-39	41.09	1.1E-03	41.09	5.5E-04	41.09	2.8E-03
chr2	96221251	LINE		-29.61	4.2E-04	-27.16	6.8E-03	-36.51	5.0E-03	-25.16	3.2E-02
chr2	96221501	LINE		-29.61	4.2E-04	-27.16	6.8E-03	-36.51	5.0E-03	-25.16	3.2E-02
chr8	67191751	LINE		60.69	3.7E-03	68.25	3.3E-03	52.28	3.7E-02	61.56	4.4E-02
chr8	67192001	LINE		60.69	3.7E-03	68.25	3.3E-03	52.28	3.7E-02	61.56	4.4E-02
chr10	106430251	LINE		-39.60	3.9E-03	-46.27	1.3E-02	-29.54	2.3E-02	-42.98	3.8E-02
chr12	49486001	LINE		39.78	2.0E-05	39.78	1.8E-02	39.78	3.3E-02	39.78	2.9E-02
chr2	172241501	LINE		42.80	4.3E-03	40.29	2.0E-02	44.28	4.8E-02	43.82	2.3E-14
chr3	99611751	LINE		38.80	5.1E-05	41.87	5.1E-03	37.73	1.1E-02	36.81	6.5E-03
chr4	144762251	LINE		27.88	2.0E-05	27.88	2.5E-02	27.88	2.8E-02	27.88	3.0E-02
chr4	144762501	LINE		27.88	2.0E-05	27.88	2.5E-02	27.88	2.8E-02	27.88	3.0E-02
chr8	57142501	LINE		38.45	3.1E-03	43.31	3.5E-02	36.32	4.3E-02	35.70	4.5E-02
chr9	76743001	LINE		28.88	4.3E-05	28.88	2.8E-02	28.88	4.8E-02	28.88	3.4E-02
chr1	81443751	LINE		37.70	4.5E-03	31.58	3.9E-02	36.20	4.0E-02	45.32	7.7E-03
chr12	60278001	LINE		36.52	1.4E-06	36.52	7.0E-03	36.52	9.1E-03	36.52	2.0E-02
chr6	101937251	LINE		54.69	1.5E-03	40.85	1.7E-02	62.05	1.7E-02	61.17	2.3E-02
chr1	135341751	LINE	<i>Lmod1</i>	48.68	1.3E-03	55.12	1.3E-02	38.87	1.4E-02	52.05	2.4E-02
chr17	56164751	LINE	<i>Tnfaip8l1</i>	45.04	5.0E-04	57.17	4.3E-04	34.03	2.3E-02	43.92	8.9E-03
chr4	79154751	LINE		58.44	1.9E-04	57.85	4.1E-02	58.73	4.9E-02	58.73	4.4E-02
chr3	133533751	LINE	<i>Tet2</i>	-31.61	3.4E-03	-30.40	2.1E-02	-32.84	5.4E-03	-31.59	2.2E-02
chr15	90784251	LTR		-36.80	2.5E-03	-26.29	1.5E-02	-40.11	2.1E-02	-44.02	3.2E-02
chr14	29814751	LTR	<i>Gm35281</i>	-58.06	1.1E-02	-30.21	1.8E-02	-74.03	4.4E-03	-69.94	2.0E-02

chr15	82957751	LTR	<i>Tcf20</i>	-50.60	2.4E-03	-54.83	1.9E-02	-47.31	2.7E-02	-49.67	3.7E-25
chr1	120210001	LTR	<i>Steap3</i>	32.91	1.1E-03	39.90	1.0E-02	32.37	3.9E-03	26.47	6.1E-03
chr4	11182001	LTR		70.77	5.2E-03	67.42	2.9E-02	65.91	4.5E-02	78.97	1.5E-26
chr18	40117751	LTR		-39.03	4.6E-04	-36.04	5.7E-03	-42.34	1.9E-03	-38.69	2.0E-02
chr15	36729501	LTR		-51.60	1.8E-03	-46.10	2.4E-02	-75.54	3.1E-03	-33.18	5.0E-03
chr11	66408751	LTR	<i>Shisa6</i>	37.00	2.4E-03	28.66	4.2E-02	35.53	4.3E-02	46.79	1.7E-02
chr13	23723001	LTR		47.35	2.2E-03	44.32	4.5E-02	51.51	1.1E-02	46.20	4.1E-02
chr7	142764001	LTR		73.08	3.7E-05	68.63	4.9E-04	70.09	4.1E-03	80.54	3.7E-03
chr19	38056751	LTR	<i>Cep55</i>	-30.81	2.6E-04	-27.51	2.4E-03	-38.04	3.8E-03	-26.87	1.4E-02
chr12	105048001	LTR		-55.56	3.8E-03	-52.42	2.9E-02	-42.37	1.2E-02	-71.89	2.5E-02
chr5	102145251	LTR		73.34	2.2E-04	87.93	6.8E-03	70.28	6.8E-03	61.80	4.4E-28
chr5	102145501	LTR		73.34	2.2E-04	87.93	6.8E-03	70.28	6.8E-03	61.80	4.4E-28
chr3	33610251	LTR		-28.31	2.3E-03	-26.92	3.7E-02	-31.30	1.0E-02	-26.71	1.0E-02
chr3	33610501	LTR		-28.31	2.3E-03	-26.92	3.7E-02	-31.30	1.0E-02	-26.71	1.0E-02
chr14	66076501	Other	<i>Adam2</i>	28.23	2.7E-04	27.77	5.4E-03	30.36	3.9E-20	26.56	7.6E-03
chr7	123046501	Other	<i>Gm45847</i>	-54.32	1.1E-02	-50.74	5.6E-03	-45.88	3.4E-02	-66.34	5.0E-02
chr5	36696501	Promoter	<i>D5Ertd579e</i>	-49.77	5.3E-03	-46.95	4.5E-02	-44.58	3.6E-02	-57.76	2.6E-02
chr6	134887501	Promoter	<i>Gpr19</i>	-59.64	2.7E-03	-50.66	3.1E-02	-63.73	1.4E-02	-64.54	9.5E-03
chr4	132180501	SINE		54.72	2.6E-03	44.83	6.6E-03	59.08	2.9E-02	60.23	7.7E-03
chr19	5357251	SINE		47.91	2.5E-04	44.71	7.6E-03	50.06	7.3E-03	48.97	1.6E-02
chr4	132180751	SINE		33.08	2.7E-03	25.97	1.5E-02	32.98	3.2E-02	40.30	2.3E-02
chr9	114550251	SINE	<i>Trim71</i>	-59.71	1.5E-03	-56.83	3.3E-02	-59.14	4.4E-02	-63.15	1.4E-02
chr3	36819001	SINE		-41.77	3.5E-03	-44.24	1.8E-02	-32.69	2.7E-02	-48.40	3.5E-02
chr1	89441751	SINE		-46.02	2.8E-03	-31.39	4.6E-03	-71.94	1.7E-02	-34.72	3.3E-02
chr5	103647501	TTS	<i>170016H13Rik</i>	-39.51	4.1E-03	-27.30	2.4E-02	-46.31	1.3E-03	-44.92	4.9E-02

chr7	142354501	TTS	<i>lfitm10</i>	54.64	2.9E-06	59.05	2.7E-04	54.00	1.8E-03	50.88	6.4E-03
chr15	75564501	TTS	<i>Ly6h</i>	-29.21	2.6E-03	-25.25	2.4E-02	-32.31	4.7E-02	-30.08	1.1E-03
chr19	42592001	TTS	<i>R3hcc1l</i>	-56.17	5.2E-04	-54.35	1.9E-02	-50.58	5.1E-03	-63.58	4.8E-03
chr12	37582251	TTS	<i>Dgkb</i>	-45.29	7.0E-04	-29.08	1.1E-02	-40.31	4.8E-03	-66.48	5.5E-03

Table 3.8. Table showing summary of correlation between gene expression and methylation levels of DMWs. The gene expression data were generated from Agilent 4x44K oligonucleotide microarrays as part of the study carried out by Labib *et al.*¹⁴². The results for different microarray transcripts for the same gene are separated by a semicolon.

Chromosome	Position	Genomic Location	Gene Name	Gene Expression Data	Number of Transcripts	Correlation Coefficients	P Values
chr19	7295001	5' UTR	<i>Mark2</i>	✓	1	0.51	0.11
chr2	131953001	Exon	<i>Prn</i>				
chr17	56098751	Exon	<i>Hdgfl2</i>				
chr4	43519251	Exon	<i>Tpm2</i>	✓	2	0.2; 0.36	0.42; 0.27
chr5	36696501	Promoter	<i>D5Erttd579e</i>	✓	5	0.48; 0.69 ; -0.09; 0.14; 0.50	0.14; 0.02 ; 0.80; 0.69; 0.12
chr6	134887501	Promoter	<i>Gpr19</i>	✓	1	0.12	0.73
chr5	103647501	TTS	<i>170016H13Rik</i>	✓	1	-0.61	0.052
chr7	142354501	TTS	<i>Ifitm10</i>				
chr15	75564501	TTS	<i>Ly6h</i>	✓	2	0.02; -0.45	0.97; 0.17
chr19	42592001	TTS	<i>R3hcc1l</i>				
chr12	37582251	TTS	<i>Dgkb</i>	✓	4	-0.067; -0.40; 0.36; 0.067	0.85; 0.67; 0.27; 0.85

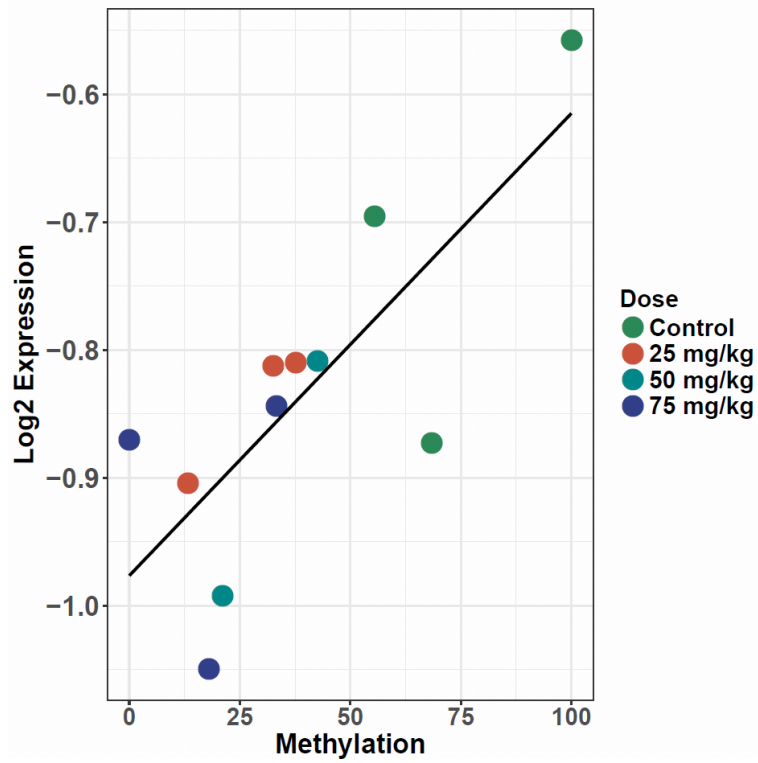


Figure 3.9 Scatterplot of the Log₂ expression of one of the transcripts of the *D5Erttd579e* gene against the methylation levels of the 500 bp window located at position 36696501 on chromosome 5. Expression and methylation were found to be significantly positively correlated using Spearman's correlation.

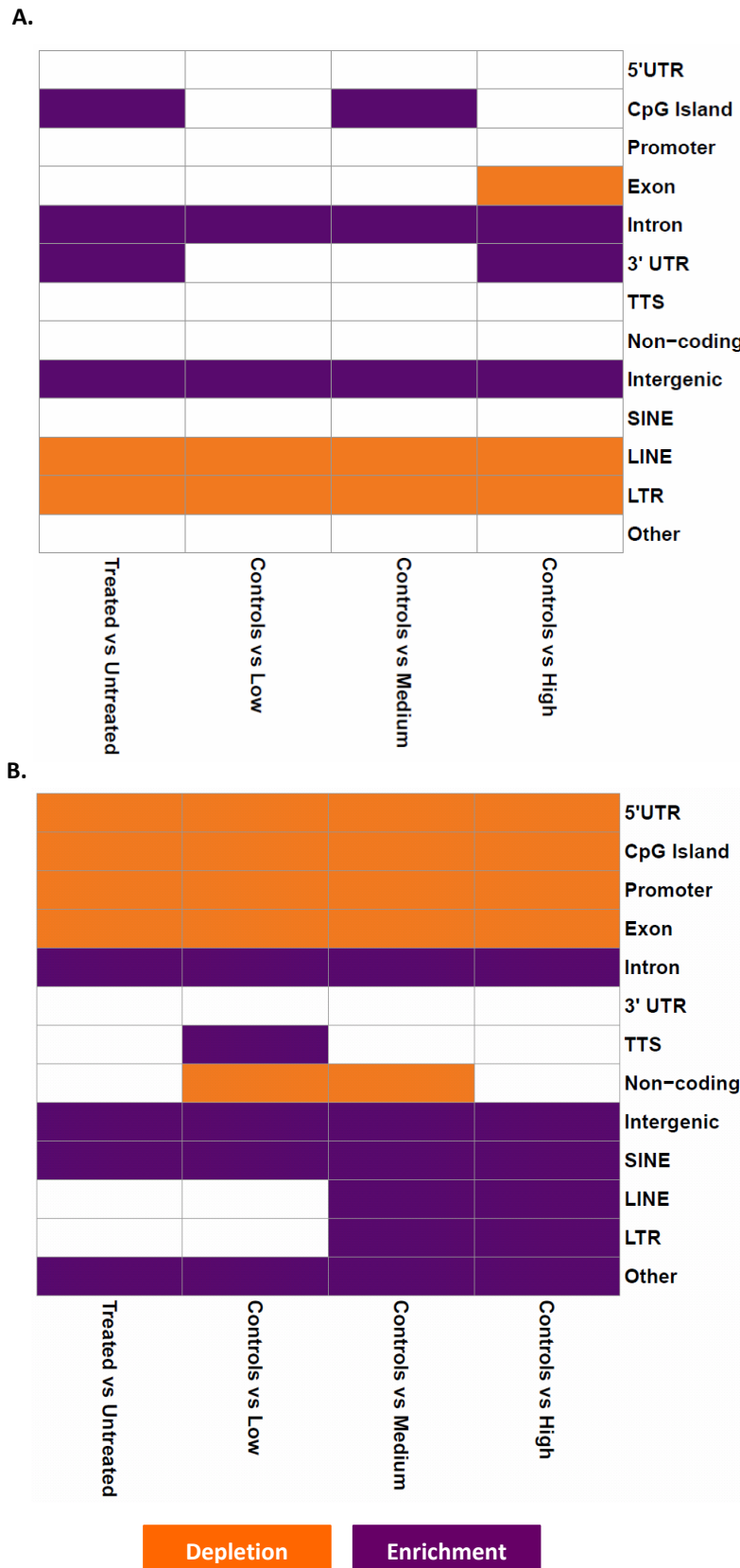


Figure 3.10 Summary of the analysis looking for enrichment or depletion of differential methylation in different genomic regions across the four models for the sites analysis (A) and windows analysis (B).

3.3 Discussion

This is the first study to employ sequencing-based methods to assess the genome-wide effects of B[a]P exposure on DNA methylation in an animal model. RRBS libraries were prepared from the lung tissues of 9 mice exposed to three doses of B[a]P and 3 controls. The analysis was carried out both at CpG level and using 500 b.p. windows with a 250 b.p. slide. Inter-sample variation was high, even between the control mice, and the number of sites and windows overlapping across samples resulted in only a fraction of the data generated being used. Despite this, a number of DMCs and DMWs were identified, however these were predominantly located in non-genic regions (intronic and intergenic) and there was little to no relationship between DNA methylation and gene expression.

3.3.1 Inter-Sample Variation

3.3.1.1 Technical Explanations

Large amounts of variation were observed in the sequencing data generated. The number of sequencing reads varied widely between samples which was consequently reflected in the number of CpG sites and 500 b.p windows covered. The number of sequencing reads obtained per sample was compared to the volume of each sample library added to the pool in order to determine whether any correlation between the two could account for the observed differences. This was found not to be the case with the exception of the samples with the lowest and highest number of reads for which the lowest and second highest volumes were added to the pool respectively. Other steps in the preparation of the libraries may have introduced this variation which have been described in previous studies. Restriction enzyme digestion has been shown to be inhibited *in vitro* due to the presence of BPDE-DNA adducts³¹¹ which may account for some of the differences observed in the samples from B[a]P-treated mice, but variation was also observed in the controls. The bisulphite conversion step of RRBS library preparation is fundamental to the process, however, it may cause issues such as incomplete conversion, reduced DNA complexity, and fragment degradation²⁹⁴. Any of these issues could result in inefficient or curbed PCR amplification which is essential to ensure that sufficient

material is sequenced. Though not presented above, the libraries were checked on the BioAnalyzer Agilent 2200 TapeStation using the Agilent High Sensitivity D5000 ScreenTape and reagents (Agilent Technologies) to ensure that this was not the case. Additionally, the impact of the presence of DNA adducts on these technical steps is as yet not understood. Since all sample libraries were pooled together and sequenced on the same lanes, it is highly unlikely that the variation was introduced during the sequencing process. Therefore, taken together, the results suggest that further work needs to be carried out in order to determine the exact cause of the variation in read number between samples and the potential impact of DNA lesions on the RRBS library preparation steps.

These technical steps may also partly explain the low number of overlapping CpG sites across all samples, specifically MspI digestion and bisulphite conversion. Read depth may also account for this with the median read depth ranging between 23 and 38 for the 12 libraries. Therefore, by sequencing more, on a further lane for example, the number of reads and overlapping CpG loci may have been improved. Additionally, relaxing the threshold for overlapping loci i.e. from all samples to two mice per dose for example, would also allow for more data to be used however this may pose problems analytically. Low overlap has been reported in a previous RRBS study³²¹ and this study, along with others^{322–324}, have analysed RRBS data tiled into windows containing multiple CpG sites in order to maximise the amount of data included in analyses. In order to increase coverage as much as possible a 500 b.p. window with a 250 b.p. slide was used in the current analysis. However, due to the limited number of studies having analysed RRBS data to date, a standard analytical method has not yet been established, and the studies cited above used different window lengths (100 b.p., 200 b.p. and 2kb).

3.3.1.2 *Biological Explanations*

When taking into account only CpG sites or windows common to all mice, PCA identified high inter-individual variation between samples from the same exposure group and this was particularly evident in the control samples suggesting that despite the genetic similarities of the mice, there was inherent epigenetic variation between them. This was also the case in another mouse RRBS study looking at

sex differences between isogenic mice, however the intra-sex differences were slightly smaller and the samples clustered by sex³²² whereas the B[a]P treated mice presented here which were all male did not cluster by exposure dose. High variation at CpG loci or windows between the same B[a]P exposure group gave rise to results that were skewed by a single mouse sample with methylation levels vastly different to all others. In these instances, the observed differences in methylation were not associated with B[a]P exposure, but rather some other underlying biological difference or an artefact introduced during library preparation. To overcome this, a filtering step could have been applied to only include CpG sites or windows with low variation between samples from the same B[a]P dose. The main drawback of this approach is that the methylation dataset is further reduced.

3.3.2 Genomic Distribution of B[a]P-induced DNA Methylation Changes

Of the CpG sites and windows that were found to be differentially methylated in B[a]P exposed mice compared to controls, a significantly higher number than expected were located within intronic and intergenic regions in both analyses. This is also the case in a recent study which carried out RRBS on a human liver cell line exposed to B[a]P³²⁵. The authors noted that nearly 40% of their observed methylation differences occurred in intronic and intergenic regions (22.5% and 17.4% respectively) and this was surpassed only by changes at SINE regions (34.5%)³²⁵. In the results presented above, more than 40% of the observed DMCs and more than 60% of the observed DMWs occurred at intronic and intergenic regions, with only 5-7% occurring at SINE regions. An early RRBS study reported that during cell differentiation, methylation changes occur at regulatory regions outside of promoters³⁰². Additionally, similar observations were made when comparing methylation differences between male and female isogenic mice³²². Taken together, these findings suggest that DNA methylation has a role to play in these genomic regions which may not be directly linked to gene expression unless these regions are as-yet unidentified enhancers.

3.3.3 Methylation and Gene Expression

However, even the few methylation differences found within gene promoters and gene bodies were not correlated with gene expression with a single exception. The promoter of the D5Erttd579e gene was hypomethylated in B[a]P-exposed mice compared to controls and this was associated with reduced expression levels in the exposed mice. The positive correlation was unexpected as the methylation status of gene promoters is widely accepted to be inversely correlated with gene expression. The observed lack of overall correlation between methylation and expression has been reported in other RRBS studies^{325,326}, but one study has reported the expected correlations³²⁷ and another reported mixed associations³²⁴.

In addition to the lack of correlation between gene expression and methylation, no overlaps were found between the differentially methylated genes identified here and gene expression changes reported by Labib *et al.* (2012)¹⁴² in the same tissues of the same mice. A B[a]P-dose-dependent increase in the number of differentially expressed genes was observed by Labib *et al.* (2012)¹⁴² however the number of differentially methylated CpG sites and windows was fairly consistent across all four models, with the controls vs medium dose model having the most differences. Additionally, samples clustered together by dose based on the expression of the genes identified in the published transcriptomic analysis¹⁴², however it was only the controls that clustered separately based on the methylation changes identified above. These observations suggest that the gene expression changes identified by Labib *et al.* (2012)¹⁴² are not a direct consequence of methylation changes at the gene promoters induced by B[a]P exposure.

The Comparative Toxicogenomics Database (CTD) contains reports of 11,904 unique gene interactions associated with B[a]P exposure¹⁸⁸. Based on an estimated 19,000 – 20,000 protein-coding genes in the mouse and human genomes, this means that approximately 59% of the mouse and human genomes have been reported to be affected by B[a]P. The gene interactions reported in the CTD were compared to the results from the analyses carried out in this chapter, and the results of these

overlaps are summarised in Table 3.9. The degree of overlap varied by model and ranged from 45.5 – 62.5% which may suggest that the overlaps are due to chance. This also supports the theory that B[a]P does not induce changes in a gene-specific manner and this is explored further in section 3.3.5 below.

Table 3.9 Table summarising overlaps between chapter results and B[a]P gene interactions reported in the

CTD¹⁸⁸

<u>Resolution</u>	<u>Model</u>	<u>Number of Differentially Methylated Genes</u>	<u>Number of Overlaps with CTD Genes</u>	<u>Percentage of Overlaps to Differentially Methylated Genes</u>	<u>Total Number of Interactions reported in CTD</u>
CpG Sites	Treated vs Untreated	161	86	53.4%	134
	Control vs Low Dose	174	90	51.7%	143
	Control vs Medium Dose	224	122	54.5%	184
	Control vs High Dose	165	83	50.3%	128
	Overlaps from all models	11	5	45.5%	7
500 b.p. Windows	Treated vs Untreated	57	35	61.4%	46
	Control vs Low Dose	89	54	60.7%	76
	Control vs Medium Dose	95	50	52.6%	73
	Control vs High Dose	78	43	55.1%	67
	Overlaps from all models	8	5	62.5%	9

3.3.4 Comparison of Differentially Methylated Genes to Previously Published Results

More overlap was observed between the genes found to be differentially methylated in the mice exposed to B[a]P and other B[a]P exposure studies included in the Comparative Toxicogenomics Database. However, this overlap is still not completely convincing for the following reasons: many of the published studies assessing B[a]P exposure have small overlaps between themselves, and this is apparent since some of the differentially methylated genes identified here have only previously been reported by a single study. Moreover, where multiple studies have published differential expression of the same gene, these are sometimes in conflict with each other thereby undermining a possible relationship between methylation and expression. Consistency across multiple studies is rare, however *Tpm2* expression was found to be down-regulated in two mouse studies^{57,328} and one rat study¹³⁹ which was consistent with the observed hypomethylation of an exon within this gene reported in this thesis.

Recently, Tryndyak *et al.* (2018)³²⁵ exposed human liver cells (HepaRG cell line) to B[a]P and carried out RRBS. The authors noted over 6500 differentially methylated regions in these cells compared to controls. As mentioned above, these authors also identified a significant number of changes within intronic and intergenic regions, as well as SINE elements. Another similarity is that Tryndyak *et al.* (2018)³²⁵ reported that more hypermethylation events than hypomethylation events occurred at Alu repetitive elements (a family of SINE elements), and this was also the case at LINE elements in mouse lung tissue as shown above. However, this is where the similarities between the two studies ended. No overlap between differentially methylated genes identified by Tryndyak *et al.* (2018)³²⁵ and those identified in this thesis were found.

3.3.5 Potential Explanations for the Effects of B[a]P on DNA Methylation

The inconsistencies between studies may be explained due to observations being species- or tissue-specific, DNA methylation particularly is known to exhibit tissue-specific patterns. Additionally, B[a]P has previously been reported to alter gene expression patterns in a tissue-specific manner^{106,142,143,329}.

The lack of correlation between DNA methylation and gene expression may suggest that the observed gene expression differences are not directly regulated by DNA methylation. One possible mechanism that may explain the observations made here is that DNA methylation changes are dependent on the location of BPDE-adduct formation. As outlined earlier, it is well-established that BPDE-DNA adducts form preferentially at guanine bases adjacent to 5-mC which may result in G to T transversion mutations and therefore loss of the CpG locus and its associated methylation status. Based on this mechanism, it follows that BPDE adduct formation is not gene specific, but rather is driven by the underlying DNA landscape. This would explain the low overlap of differentially methylated regions identified by the models comparing controls to mice exposed to each dose of B[a]P, as well as the relatively low intersection of differentially expressed genes between published studies. It is possible that instead of acting in a gene-specific manner, DNA methylation of particular genomic locations, such as introns and intergenic regions, is more susceptible to changes induced by B[a]P. This may be due to less repair of DNA adducts in these regions since they are probably less tightly regulated compared to promoters or exons. Alteration of DNA methylation at intron and intergenic regions, may induce chromatin modifications, which would then be responsible for the observed dysregulation of gene expression. Down-regulation of histone genes has been reported in human cell lines exposed to B[a]P or BPDE^{96,97} and in the presence of excess ROS, one of the ways the cell protects against the damage is to complex the DNA with histones⁶⁴.

3.3.5.1 *Future Work*

This proposed mechanism, however, is merely speculative and much more work would need to be done to support it. Firstly, the RRBS experiment carried out here would need to be repeated in order to sequence more deeply to maximise genome coverage and improve overlap across samples. Additionally, more mice per group should be included to improve statistical power as much as possible. By using lower doses of B[a]P, the potential extrapolation to humans might be improved as the doses employed in this study are not representative of real environmental levels of exposure. In order to determine whether the observations made above are specific to lung tissue, the same

process could be repeated for different tissues extracted from the same mouse samples.

Furthermore, B[a]P is only a single compound, but humans are usually exposed mixtures in which the effects of individual compounds may be more- or less-than-additive. Therefore further experiments should include exposure to environmental mixtures to replicate human exposure as closely as possible.

To better understand the genomic landscape, chromatin immunoprecipitation with sequencing (ChIP-seq) could be used to map various histone modifications, along with assay for transposase-accessible chromatin using sequencing (ATAC-seq) to identify regions of open and closed chromatin. Finally, techniques are now available that allow for the genomic mapping of DNA adducts, including BPDE-DNA adducts. Availability of all these datasets would allow for the complete identification of any modifications to the epigenetic landscape induced by B[a]P exposure and the interpretation of these in the context of BPDE adduct formation. This would then be able to support the mechanism proposed above, or suggest an alternative which will explain the role of epigenetics in the carcinogenicity of B[a]P.

3.3.6 Conclusions

In summary, RRBS analysis of mouse lung tissue exposed to B[a]P identified a number of differentially methylated CpG sites and 500 b.p. windows. In all comparisons (treated vs untreated, and controls vs each individual dose) the majority of changes occurred at intergenic and intronic regions, with more hypermethylation events occurring at LINE elements than expected, and similar observations have been made in an RRBS study of B[a]P-exposed human liver cells³²⁵. Of the changes occurring at coding regions, these were not correlated with the expression levels of the associated genes in the same mouse samples. Additionally, the differentially methylated genes were not identified as differentially expressed in the analysis carried out by Labib *et al.* (2012)¹⁴² using the same gene expression data. Some cross-over between the differentially methylated genes identified here and other studies has been found but this is limited and should be interpreted cautiously.

4 Chapter 4 - DNA Methylation and Air PAH8 Exposure

4.1 Introduction

Human exposure to PAHs occurs through three main routes: air inhalation, dietary intake, and tobacco smoke exposure. An overview of the presence of PAHs in air has already been given in Chapter 1, but in this chapter policy, compliance, and the methods for measuring exposure are discussed. A few studies have been carried out linking exposure to PAHs to changes in DNA methylation and the current state of the knowledge is outlined.

4.1.1 Policy and Compliance

The 2004 European Parliament and Council directive requires all countries with ambient B[a]P concentrations above the minimum threshold of 0.4 ng/m³ to monitor B[a]P levels²⁹. The concentration of B[a]P that corresponds to an excess lifetime cancer risk of 1/10,000 is 1.2 ng/m³ according to the World Health Organisation (WHO)⁶ which is only 0.2 ng/m³ above the target value set by the European Parliament (1 ng/m³). The B[a]P concentration required to increase the excess lifetime cancer risk by one order of magnitude (1/100,000) is 0.12 ng/m³⁶. As recently as 2012, 20% of the European population were exposed to B[a]P levels in air above those stipulated by the EU directive, with central-eastern Europe having the highest concentrations³³⁰. It is important to note that this study was limited to 60% of the total European population, however, even so, exposures at this level would result in 370 new lung cancer cases annually³³⁰. In the winter of 2013 two regions in the Czech Republic had B[a]P concentrations above the EU threshold, with the levels at one of the regions being 5 times higher³³¹. A study conducted where B[a]P concentrations were monitored in the Basque Country in Spain found that B[a]P levels were consistently lower than 1 ng/m³ with the exception of a few days¹⁶. Various monitoring sites within ten European countries were all found to have recorded B[a]P levels below the EU directive threshold between 2008 and 2011²⁵. Taken together, this evidence calls for better monitoring and regulation of B[a]P levels in ambient air.

4.1.2 Measuring exposure to PAHs and Air Pollutants

4.1.2.1 Exposure Assessment

There are several methods available to determine exposure to PAHs, however, they are not all equally reliable. At the bottom of the hierarchy are area-level estimates from air monitors. These are very easy to obtain, and often are publically available however, they cover wide areas thereby making it difficult to assign exposure to individuals living close together. Dispersion and land-use regression (LUR) models allow for exposure levels to be estimated at the individual level. These model estimates are more reliable than air monitor estimates, however these are susceptible to misclassification since most models tend to only take into account data from a single address. Some studies use real measurements from personal exposure monitoring equipment, or estimates obtained from geographical models as measurements of PAH exposure. Personal exposure monitoring equipment allows for accurate and real-time monitoring of exposure to PAHs or other air pollutants, however, measured exposures do not take into account any inter-individual differences in metabolism. This means that these measurements may cause misclassification of an individual's exposure because they do not reflect actual biological burden of exposure. Finally, personal exposure monitoring and model estimates of exposure only take into account a snapshot of exposure over a short span of time.

4.1.2.2 Urinary Metabolites and Biomarkers of PAH Exposure

In many studies, urinary metabolites of PAHs such as 1-hydroxypyrene and other hydroxylated PAHs are used as proxies to measure exposure to PAHs. Other studies use DNA or albumin adduct levels from blood samples. The use of a variety of biomarkers to estimate exposure to air pollutants has been reviewed³³². The use of urinary metabolites or DNA and albumin adducts are useful measurements because they give an indication of total exposure to PAHs or pollutants, irrespective of route of exposure. These methods may not give reliable results due to underlying genetic variants that may increase or decrease metabolism of PAHs. Thus urinary metabolites and adducts are more indicative of biological impacts of exposure, rather than the exposure itself.

Biological measurements of exposure tend to be the most frequently used in epidemiological studies. At least when considering prenatal exposures, PAH-DNA adducts are considered to be the most reliable biomarkers ³³³. Other consistent air pollution biomarkers in prenatal studies are increased levels of oxidative stress markers, and a decrease in global methylation which is described in more detail below ³³³. Generally however, urinary biomarkers have been used because they tend to correlate well with exposure and are obtained non-invasively ³³⁴. Despite this, one study has shown that urinary metabolite levels do not correlate well with DNA adduct levels ³³⁵. In this study carried out on subjects from three European countries, the subjects with the highest levels of urinary metabolites were not the same subjects that had the highest number of DNA adducts ³³⁵. The authors also found that all subjects had comparable genotypes and that PAH levels in ambient air were similar across the cohort leading them to conclude that these differences were probably due to other lifestyle factors ³³⁵. Urinary metabolites only represent recent exposures and do not account for potential bioaccumulation as these metabolites are formed and excreted quickly by the body. Another consideration when considering the use of urinary metabolites as biomarkers of PAH exposure is that a single metabolite is unlikely to provide the whole picture, particularly when each parent compound may form multiple metabolites. A new urinary metabolite for PAH exposure has recently been proposed: 7,8,9,10-tetrahydroxy-7,8,9,10-tetrahydrobenzo(a)pyrene which is produced from the hydrolysis of BPDE ³³⁶. This biomarker was recommended by the authors due to the low inter-individual variation observed and an elimination half-life of 31.5 hours ³³⁶. Tetrahydroxylated PAH metabolites have also been recommended by a study carried out in rats which found that the concentration of these metabolites was dose-dependent, and that they can also be measured from hair samples ³³⁷.

4.1.2.3 Comparing Methods

A study comparing the use of urinary metabolites and geographic information system (GIS) based methods found that urinary concentrations of 1-hydroxypyrene-*O*-glucuronide (a metabolite of pyrene) did not correlate with self-reported or GIS-determined sources of PAH exposure like wood

burning or traffic levels³³⁸. Urinary metabolites and occupational indoor PAH exposure were found to be significantly correlated with each other in a cohort of firemen, with the authors attributing lower correlations to higher PAH exposure outside the workplace³³⁹. Taken together, further work is required in order to fully determine which method of PAH exposure assessment is better, and in the meantime, the limitations and advantages described above should be considered when choosing which method to use.

4.1.3 PAHs in Air and DNA Methylation

There is a large degree of heterogeneity between the epidemiological studies that have been carried out to date assessing the effects of PAH exposure on methylation, making the direct comparison of results difficult^{165,340}. Most studies used urinary PAH metabolites as indicators of PAH exposure^{146,313,349–352,341–348}, others used DNA or albumin adducts^{49,345,346,353–355}, and a few used occupation to differentiate between high and low exposure groups^{342,343,351,356}. Only a small number of studies have used measurements from personal air monitoring equipment or air quality monitoring stations^{345,348,353,357}. Some occupations such as coke production, aluminium production and brick-making are known to expose workers to high concentrations of PAHs and therefore these exposures are not representative of the general population. The results of published studies assessing the effects of PAH exposure on DNA are summarised in Table 4.1.

A few studies have investigated the effects of prenatal PAH exposure on DNA methylation^{345,353,357,358}. A study including 164 pregnant women from a cohort where PAH exposure was measured using personal air monitors reported some contradictory findings: higher maternal PAH exposure was associated with lower global methylation levels in cord blood, but in new-borns with detectable levels of DNA adducts, the global methylation was higher than in those where no adducts were detected³⁴⁵. It would be expected that babies born to mothers with higher PAH exposures would have more DNA adducts and that the effect on DNA methylation would be the same, however, this was not the case.

Two studies have investigated the effects of PAH exposure on DNA methylation in children who have been reported to be more susceptible to the effects of PAHs than adults ^{359,360}.

Table 4.1 Table summarising published results of associations between PAH exposure and DNA methylation.

Gene	Population	Exposure	Exposure Measurement	Direction	Reference
<i>ACSL3</i>	Pregnant women	Normal	Personal exposure monitors, urinary metabolites; DNA adducts in cord blood	+	Perera, 2009 ³⁵³
<i>AHRR</i>	Chimney sweeps and creosote workers	Occupational	Urinary metabolites	-	Alhamdow, 2018 ³⁶¹
<i>Alu</i>	Coke oven workers	Occupational	Urinary metabolites and blood DNA adducts	+	Pavanello, 2009 ³⁴⁶
	Pregnant women	Normal	Urinary metabolites	-	Yang, 2017 ³⁶²
<i>APEX</i>	Children	Ambient PAH exposure	HPLC	+	Alvarado-Cruz, 2017 ³⁶⁰
<i>BRCA1</i>	Breast cancer patients	Use of synthetic logs for heating	NA	-	White, 2016 ³⁶³
<i>CDH1</i>	Breast cancer patients	Use of synthetic logs for heating	NA	+	White, 2016 ³⁶³
<i>DUSP22</i>	Firefighters	Occupational	NA	-	Ouyang, 2012 ³⁵⁶
<i>F2RL3</i>	Pregnant women	Normal	Personal exposure monitors, urinary metabolites; DNA adducts in cord blood	-	Alhamdow, 2018 ³⁶¹
<i>FOXP3</i>	Children with asthma/ allergic rhinitis	Ambient PAH exposure	Spatio-temporal model	+	Hew, 2015 ³⁵⁹
<i>GSTP</i>	Hepatocellular carcinoma patients	Normal	Albumin DNA adducts	+	Tian, 2016 ³⁵⁴
<i>HIC1</i>	Coke oven workers	Occupational	Urinary metabolites and blood DNA adducts	-	Pavanello, 2009 ³⁴⁶
<i>HIN1</i>	Breast cancer patients	Use of synthetic logs for heating	NA	+	White, 2016 ³⁶³
<i>IFN-γ</i>	Pregnant women	Normal	Personal exposure monitors	+	Tang, 2012 ³⁵⁷
<i>IL-12</i>	Brick makers	Occupational	Urinary metabolites	-	Alegria-Torres, 2013 ³⁴¹
<i>IL-6</i>	Coke oven workers	Occupational	Urinary metabolites and blood DNA adducts	+	Pavanello, 2009 ³⁴⁶
<i>IRS2</i>	Non-smoking Korean women	Normal	PAHs in adipose tissue	+	Kim, 2016 ³⁶⁴

<i>LINE-1</i>	Coke oven workers	Occupational	Urinary metabolites		Duan, 2013 ³⁴³
	Pregnant women	Coal factory emissions	DNA adducts in cord blood		Lee, 2017 ³⁵⁸
	Breast cancer patients	Use of synthetic logs for heating	NA	-	White, 2016 ³⁶³
	Coke oven workers	Occupational	Urinary metabolites		Yang, 2018 ³⁴⁷
	Pregnant women	Normal	Urinary metabolites		Yang, 2017 ³⁶²
	Children	Ambient PAH exposure	HPLC		Alvarado-Cruz, 2017 ³⁶⁰
<i>MGMT</i>	Coke oven workers	Occupational	Urinary metabolites and blood DNA adducts	+	Pavanello, 2009 ³⁴⁶
	Coke oven workers	Occupational	Urinary metabolites		Duan, 2013 ³⁴³
<i>OGG1</i>	Diesel engine exhaust particle exposed workers	Occupational	Urinary metabolites	-	Zhang, 2015b ³⁴⁹
	Children	Ambient PAH exposure	HPLC	+	Alvarado-Cruz, 2017 ³⁶⁰
<i>p14^{ARK}</i>	Coke oven workers	Occupational	Urinary metabolites	+	Zhang, 2015a ³⁴⁸
<i>p15^{INK4b}</i>	Coke oven workers	Occupational	Urinary metabolites	+	Zhang, 2015a ³⁴⁸
<i>p16</i>	Diesel engine exhaust particle exposed workers	Occupational	Urinary metabolites	-	Zhang, 2015b ³⁴⁹
<i>p16^{INK4α}</i>	Coke oven workers	Occupational	Urinary metabolites		Yang, 2011
	Coke oven workers	Occupational	Urinary metabolites	+	Zhang, 2015a ³⁴⁸
<i>p53</i>	Brickmakers	Occupational	Urinary metabolites	-	Alegria-Torres, 2013 ³⁴¹
<i>p53</i>	Coke oven workers	Occupational	Urinary metabolites and blood DNA adducts	-	Pavanello, 2009 ³⁴⁶
<i>PARP1</i>	Children	Ambient PAH exposure	HPLC	+	Alvarado-Cruz, 2017 ³⁶⁰
<i>RAD21</i>	Pregnant women	Normal	Personal exposure monitors, urinary metabolites; DNA adducts in cord blood	+	Perera, 2009 ³⁵³
<i>RARβ</i>	Breast cancer patients	Use of synthetic logs for heating	NA	+	White, 2016 ³⁶³
<i>RASSF1A</i>	Diesel engine exhaust particle exposed workers	Occupational	Urinary metabolites	-	Zhang, 2015b ³⁴⁹
	Coke oven workers	Occupational	Urinary metabolites	+	He, 2015 ³⁵¹

<i>SCD5</i>	Pregnant women	Normal	Personal exposure monitors, urinary metabolites; DNA adducts in cord blood	+	Perera, 2009 ³⁵³
<i>SFMBT2</i>	Pregnant women	Normal	Personal exposure monitors, urinary metabolites; DNA adducts in cord blood	+	Perera, 2009 ³⁵³
<i>TRIM36</i>	Coke oven workers	Occupational	Urinary metabolites	+	He, 2017 ³¹³
<i>WVOX</i>	Pregnant women	Normal	Personal exposure monitors, urinary metabolites; DNA adducts in cord blood	+	Perera, 2009 ³⁵³

Only two studies have been published relating PAH exposure in “normally exposed” adults to DNA methylation. Hypermethylation of *IRS2* was associated with the PAH levels measured in the visceral adipose tissue of 53 Korean women ³⁶⁵. In a cohort of 539 Chinese subjects, urinary PAH metabolites were found to be associated with methylation age and methylation aging rate ³⁴⁴. A 1 unit increase in 1-hydroxypyrene led to 0.53 year increase in methylation age and a 1.17% increase in aging rate, and 1 unit increase in 9-hydroxyphenanthrene led to a 0.54 year increase in methylation age and a 1.15% increase in aging rate ³⁴⁴. 1-hydroxypyrene and 9-hydroxyphenanthrene were associated with 3 and 6 CpGs from the methylation age predictor model respectively ³⁴⁴.

PAH exposure and differential DNA methylation have also been linked in cancer in hepatocellular carcinoma patients ³⁵⁴ and breast cancer ^{355,363}. PAH-DNA adducts were measured in the blood of breast cancer patients and controls, and the methylation status of selected genes was measured in tumour tissue ³⁵⁵. Patients were more likely to have hormone-receptor positive tumours if they had detectable levels of DNA adducts, and if the *RARβ* and *APC* genes were methylated ³⁵⁵.

No studies have been published trying to characterise the methylome-wide effects of PAH exposure in humans, and only a few have been carried out in animal models as mentioned in the previous chapter. Most recently, a study linked PAH exposure measured through urinary metabolites with accelerated DNA methylation aging in a Chinese cohort ³⁵². One of the major limitations of many of the studies referred to above is the small cohort sizes, however, this was not the case for all. Additionally, almost all of the investigations looked onto a unique set of genes, making comparison very difficult, and where the same loci were interrogated, in some instances opposing effects were reported, such as those for LINE-1 and global methylation. This heterogeneity calls for further genome-wide methylation studies to be carried out which would allow for better inter-study comparisons and hopefully less variable results.

4.1.4 Effects of Other Air Pollutants on DNA Methylation

Associations between various known air pollutants and DNA methylation have been made, but PM₁₀ and PM_{2.5} are the most common. Using pollutant levels from air monitors has been the most used method, followed by the use of estimates calculated using models built using real measurements. Few studies have used occupation or residential status as proxies for exposure to air pollutants. The results of these studies are summarised in Table 4.2. Taken together, these studies show that the body of epidemiological evidence linking exposure to air pollutants to differential methylation is growing, with several differentially methylated genes identified. However, many of these studies only interrogated the methylation status of a handful of genes, usually by pyrosequencing. Due to this, the majority of the genes mentioned above have only been reported by a single study, or two studies carried out by the same authors in the same cohorts. Despite this, one consistent finding does emerge: several air pollutants are associated with a decrease in global methylation. Only one study looked at the genome-wide effects of air pollutants on methylation¹⁹¹ and more studies of this kind are required in order to properly identify differential methylation patterns and validate reported findings.

4.1.5 Aims

One of the principal hypotheses of this PhD project was to determine whether route of exposure to PAHs is important in determining the downstream effects of exposure. The primary aim of this chapter was to carry out an epigenome-wide study of the association between DNA methylation and exposure to PAHs via air inhalation specifically. In order to achieve this, land-use regression models were used to estimate air exposure to eight of the most carcinogenic PAHs known as PAH8: B[a]A, B[b]Fl, B[k]Fl, B[ghi]P, B[a]P, DB[a,h]A, Chr, and I[cd]P. Following this, an EWAS was carried out to identify differentially methylated probes that were then used to build a methylation index of air PAH8 exposure.

Table 4.2 Table summarised published results of associations between air pollutants and DNA methylation

Gene	Exposure	Exposure Measurement	Direction	Reference
<i>Alu</i>	Black carbon	Air monitor	-	Madrigano, 2011 ³⁶⁶
	PM10 in steel workers	Occupation	-	Tarantini, 2009 ³⁶⁷
	PM10 in steel workers	Occupation and measurements	+	Byun, 2013 ³⁶⁸
<i>APC</i>	Steel workers	Occupation	+	Hou, 2011 ³⁶⁹
<i>CCND2</i>	Traffic emissions at first birth	Model estimates	-	Callahan, 2018 ³⁷⁰
<i>F3</i>	Particulates; Black carbon	Air monitor	-	Bind, 2014 ³⁷¹
<i>FOXP3</i>	Ambient air pollution	Air monitor	+	Kohli, 2012 ³⁷²
	Ambient air pollution	Air monitor	+	Nadeau, 2010 ²³
<i>GCR</i>	NO ₂ ; PM10; PM2.5; Ozone	Model estimates	-	De Prins, 2013 ³⁷³
<i>HIC1</i>	Industrial estate workers	Occupational	+	Peluso, 2012 ³⁷⁴
<i>ICAM1</i>	Particulates; Ozone	Air monitor	-	Bind, 2014 ³⁷¹
<i>IFN-γ</i>	Ozone	Air monitor	+	Bind, 2014 ³⁷¹
	Ambient air pollution	Air monitor	+	Kohli, 2012 ³⁷²
<i>IL-6</i>	Industrial estate workers	Occupation	-	Peluso, 2012 ³⁷⁴
<i>iNOS</i>	PM2.5	Air monitor	-	Salam, 2012 ³⁷⁵
	PM10 in steel workers	Occupation	-	Tarantini, 2009 ³⁶⁷
<i>LINE-1</i>	Black carbon, PM2.5	Model estimates	-	Baccarelli, 2009 ³⁷⁶
	PM10 in steel workers and benzene in gas-station workers	Occupation and measurements	-	Byun, 2013 ³⁶⁸
	Sulphates	Air monitor	-	Madrigano, 2011 ³⁶⁶
	Industrial estate workers	Occupation	-	Peluso, 2012 ³⁷⁴
	PM10 in steel workers	Occupation	-	Tarantini, 2009 ³⁶⁷
<i>NBC2</i>	PM10	Air monitor	-	Guo, 2014 ³⁷⁷

<i>NOS2A</i>	PM2.5	Model estimates	+	Breton, 2010 ³⁷⁸
<i>p16</i>	Steel workers	Occupation	+	Hou, 2011 ³⁶⁹
<i>p53</i>	Steel workers	Occupation	-	Hou, 2011 ³⁶⁹
	Industrial estate workers	Occupational	-	Peluso, 2012 ³⁷⁴
<i>RASSF1A</i>	Steel workers	Occupational	-	Hou, 2011 ³⁶⁹
<i>SATα</i>	PM2.5; PM10	Air monitor	-	Guo, 2014 ³⁷⁷
<i>SCGB3A1</i>	Total suspended particulates at first birth	Model estimates	-	Callahan, 2018 ³⁷⁰
<i>SYK</i>	Total suspended particulates at first birth; Traffic emissions at menarche	Model estimates	+	Callahan, 2018 ³⁷⁰
<i>TLR2</i>	Air pollution	Air monitor	-	Lepeule, 2014 ³⁷⁹

4.2 Results

4.2.1 Comparison of Cohort Characteristics

The characteristics of the cohorts used in this analysis are summarised in Table 4.3 and Figure 4.1 . No differences ($p > 0.05$) in any of the characteristics were observed between the Training and Testing datasets, both of which originated from the EPIC-Italy cohort. This was not the case however, for the EPIC-NL dataset. The distributions for age and air PAH8 exposure were significantly different ($p < 0.05$) between the EPIC-NL and Training and Testing datasets, with the EPIC-NL subjects being older and having a lower, narrower range of air PAH8 exposure. The EPIC-NL dataset also did not include any male subjects, current smokers, or subjects diagnosed with cancer, all of which were present in fairly equal proportions in the other two datasets. The Training dataset had the largest number of subjects ($N = 493$), followed by the Testing dataset ($N = 208$), and the EPIC-NL dataset ($N = 132$).

Comparison of the distribution of the mean global methylation, and mean methylation at various genomic regions across the three datasets showed significant differences. These distributions are shown in Figure 4.2. Statistically significant ($p < 0.05$) differences between the EPIC-Italy derived datasets and the EPIC-NL dataset were observed for mean global, shores, CpG islands, promoters, and gene body methylation.

4.2.2 Effects of Air PAH8 Exposure on Global Methylation

For each dataset, the air PAH8 exposure was divided into quartiles and regressed against the average methylation of all CpG probes in the three datasets independently. The air PAH8 exposure quartiles for each dataset are summarised in Table 4.3. The results of the regression are summarised in Table 4.4. No significant differences in global methylation between the air PAH8 exposure quartiles was observed across any of the datasets when compared to the global methylation levels of the subjects in the lowest quartile (Table 4.4). In the Training dataset, the trend was statistically significant for a decrease in global methylation with increasing air PAH8 exposure (β coefficient = -0.003; $p = 0.006$), but this was not reflected in the quartile analysis (Table 4.4).

Table 4.3 Table of cohort characteristics for the training, testing and EPIC-NL EPIC subjects.

		EPIC-Italy Training	EPIC-Italy Testing	EPIC-NL
Number of Subjects		493	208	132
Centres	Varese	106	45	NA
	Turin	387	163	NA
	Utrecht	NA	NA	115
	Bilthoven	NA	NA	17
Age	Range	35.04-72.05	36.25-71.94	31.34-69.01
	Mean	53.26	53.95	58.71
	Median	54.06	54.54	59.10
Gender	Male	192	76	NA
	Female	301	132	132
Smoking Status	Never	215	109	78
	Ex	136	49	54
	Current	142	50	NA
Cancer Case Status	Controls	296	127	132
	Cases	207	81	NA
PAH8 Exposure (ng/m ³) – Air Pollution	Range	1.36-4.54	1.29-4.37	0.56-2.25
	Mean	2.05	2.03	1.34
	Median	1.99	1.99	1.34
Air PAH8 Exposure Quartiles (ng/m ³)	Q1	1.36 – 1.73	1.29 – 1.73	0.56 – 1.30
	Q2	1.73 – 2.00	1.73 – 1.99	1.30 – 1.34
	Q3	2.00 – 2.16	1.99 – 2.15	1.34 – 1.38
	Q4	2.16 – 4.54	2.15 – 4.37	1.38 – 2.25

Figure 4.1 Summary of cohort characteristics. A: Distribution and air PAH8 exposure in the training, testing and EPIC-NL cohorts. B: Distribution of age in the three datasets. C: Gender proportions within each dataset. D: Smoking status proportions within each dataset. E: Proportion of subjects with cancer case or control status in each dataset.

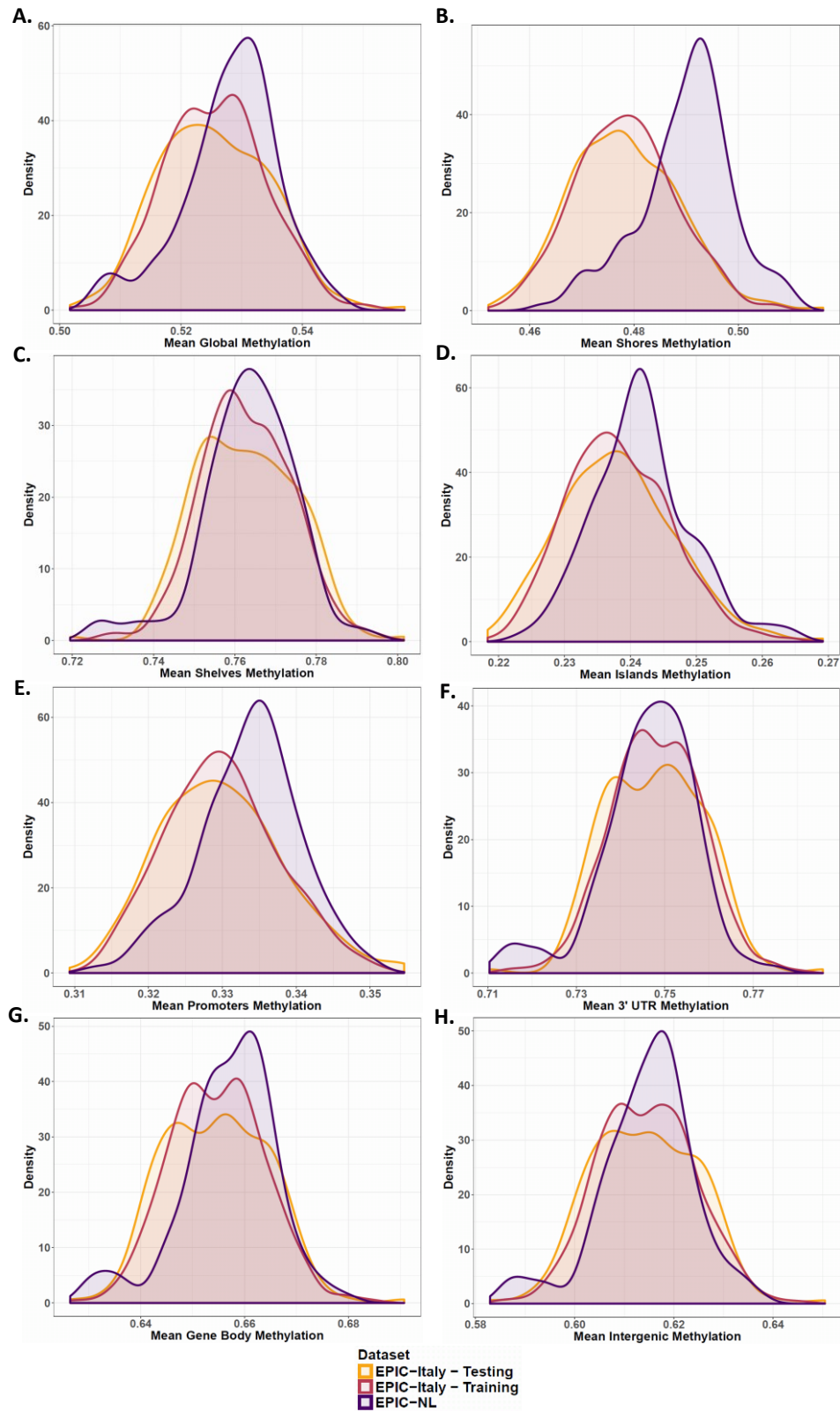


Figure 4.2 Mean methylation distributions of global methylation (A), shores (B), shelves (C), CpG islands (D), promoters (E), 3' UTRs (F), gene bodies (G), and intergenic regions (H) in the Training (yellow), Testing (pink), and EPIC-NL (purple) datasets

Table 4.4 Table of beta regression results looking for differences in global methylation between quartiles of air PAH8 exposure. The lowest quartile (Q1) was used as the reference quartile.

PAH Quartile	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-Netherlands - EPIC-NL (N=132)		
	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value
Q2	-0.006	-0.003; 0.002	0.57	-0.003	-0.008; 0.001	0.16	-0.004	-0.01; 0.004	0.32
Q3	0.002	-0.0005; 0.004	0.13	-0.0006	-0.005; 0.004	0.80	-0.002	-0.01; 0.007	0.65
Q4	-0.001	-0.003; 0.001	0.31	-0.001	-0.006; 0.003	0.54	-0.002	-0.01; 0.006	0.58
Trend	-0.003	-0.004; -0.0008	0.006	0.0007	-0.003; 0.004	0.72	-0.008	-0.02; 0.008	0.34

4.2.3 Effects of Air PAH8 Exposure on Methylation at Genomic Regions

The same quartiles described above were regressed against the average methylation of probes located at various genomic features and regions of interest (Table 4.5 and Table 4.6). For CpG islands, shores, and shelves, no significant differences were observed in their average methylation across quartiles of air PAH8 exposure (Table 4.5). The observed directions of the non-significant differences were also heterogeneous across the three datasets, and in some cases across the quartiles within the same dataset. The trend between methylation at shore regions and air PAH8 exposure was found to be statistically significant (β coefficient = -0.003; $p = 0.003$) in the Training dataset (Table 4.5).

Similar observations were made when comparing the average methylation at promoters, gene bodies, 3' UTRs, and intergenic regions with air PAH8 exposure quartiles (Table 4.6). No statistically significant differences between any of the exposure quartiles compared to the lowest quartile of exposure across all datasets were observed. In the Training Dataset however, the trends between methylation at all four of these regions and air PAH8 exposure were all found to be significant (Promoters: β coefficient = -0.003, $p = 0.04$; 3'UTR: β coefficient = -0.004, $p = 0.05$; Gene body: β coefficient = -0.003, $p = 0.01$; Intergenic regions: β coefficient = -0.003, $p = 0.03$) (Table 4.6).

Table 4.5 Table of beta regression results looking differences in methylation levels at CpG islands, shores and shelves between quartiles of air PAH8 intake. The lowest quartile (Q1) was used as the reference quartile. Entries in bold indicate statistical significance ($p < 0.05$).

Genomic Region	PAH Quartile	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-Netherlands - EPIC-NL (N=132)		
		β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value
Shores	Q2	-0.001	-0.0004; 0.001	0.32	-0.004	-0.009; 0.001	0.15	-0.006	-0.01; 0.002	0.16
	Q3	0.002	-0.005; 0.005	0.12	0.0001	-0.005; 0.005	0.96	-0.002	-0.01; 0.007	0.60
	Q4	-0.002	-0.005; 0.0005	0.11	-0.002	-0.007; 0.004	0.56	-0.0008	-0.009; 0.007	0.85
	Trend	-0.003	-0.005; -0.001	0.003	0.0003	-0.004; 0.004	0.90	-0.006	-0.02; 0.01	0.46
Shelves	Q2	-0.0004	-0.004; 0.003	0.83	-0.002	-0.01; 0.009	0.75	-0.01	-0.03; 0.004	0.14
	Q3	0.001	-0.002; 0.005	0.50	-0.006	-0.02; 0.004	0.27	-0.006	-0.02; 0.001	0.51
	Q4	-0.002	-0.006; 0.002	0.34	-0.002	-0.01; 0.008	0.71	-0.01	-0.03; 0.005	0.19
	Trend	-0.003	-0.007; 0.001	0.14	0.0008	-0.007; 0.009	0.84	-0.018	-0.05; 0.01	0.27
CpG Islands	Q2	-0.002	-0.007; 0.003	0.43	-0.008	-0.02; -0.003	0.16	0.002	-0.01; 0.01	0.71
	Q3	0.003	-0.001; 0.008	0.13	0.005	-0.006; 0.017	0.35	0.003	-0.01; 0.02	0.68
	Q4	-0.003	-0.008; 0.002	0.19	0.0002	-0.01; 0.01	0.97	0.008	-0.003; 0.02	0.18
	Trend	-0.004	-0.009; 0.0001	0.055	0.003	-0.006; 0.01	0.49	0.002	-0.02; 0.03	0.89

Table 4.6. Table of beta regression results looking differences in methylation levels at promoter, 3' UTR, gene body, and intergenic regions between quartiles of air PAH8 intake. The lowest quartile (Q1) was used as the reference quartile. Entries in bold indicate statistical significance (p < 0.05).

<u>Genomic Region</u>	<u>PAH Quartile</u>	<u>EPIC-Italy – Training (N=493)</u>			<u>EPIC-Italy – Testing (N=208)</u>			<u>EPIC-Netherlands - EPIC-NL (N=132)</u>		
		<u>β coefficient</u>	<u>Confidence Interval</u>	<u>P value</u>	<u>β coefficient</u>	<u>Confidence Interval</u>	<u>P value</u>	<u>β coefficient</u>	<u>Confidence Interval</u>	<u>P value</u>
Promoters	Q2	-0.001	-0.004; 0.002	0.49	-0.006	0.01; 0.001	0.11	-0.0005	-0.009; 0.008	0.90
	Q3	0.002	-0.004; 0.005	0.10	0.001	-0.005; 0.008	0.69	-0.0007	-0.01; 0.009	0.89
	Q4	-0.001	-0.004; 0.002	0.42	-0.002	-0.009; 0.005	0.62	0.001	-0.007; 0.009	0.75
	Trend	-0.003	-0.006; -0.0002	0.04	0.002	-0.004; 0.007	0.52	-0.005	-0.02; 0.01	0.58
3' UTR	Q2	-0.001	-0.004; 0.002	0.52	-0.003	-0.011; 0.006	0.54	-0.009	-0.02; 0.005	0.20
	Q3	0.0008	-0.002; 0.004	0.63	-0.004	-0.01; 0.004	0.33	-0.005	-0.02; 0.01	0.53
	Q4	-0.003	-0.006; 0.0004	0.09	-0.002	-0.01; 0.007	0.62	-0.008	-0.02; 0.006	0.25
	Trend	-0.004	-0.007; -0.0000002	0.05	-0.0004	-0.007; 0.007	0.92	-0.02	-0.04; 0.01	0.30
Gene Body	Q2	-0.0009	-0.003; 0.001	0.42	-0.002	-0.007; 0.003	0.47	-0.006	-0.02; 0.004	0.26
	Q3	0.001	-0.001; 0.003	0.28	-0.001	-0.007; 0.004	0.63	-0.002	-0.01; 0.009	0.75
	Q4	-0.002	-0.004; 0.0002	0.07	-0.0005	-0.006; 0.005	0.85	-0.004	-0.01; 0.006	0.42
	Trend	-0.003	-0.005; -0.0006	0.01	0.0008	-0.003; 0.005	0.72	-0.010	-0.03; 0.01	0.34
Intergenic	Q2	0.00004	-0.003; 0.003	0.98	-0.004	-0.01; 0.002	0.19	-0.1	-0.02; 0.0008	0.07
	Q3	0.002	-0.0009; 0.005	0.17	-0.003	-0.009; 0.003	0.28	-0.006	-0.02; 0.006	0.32
	Q4	0.0001	-0.003; 0.003	0.94	-0.004	-0.01; 0.002	0.21	-0.004	-0.01; 0.006	0.40
	Trend	-0.003	-0.006; -0.0003	0.03	-0.001	-0.006; 0.004	0.68	-0.010	-0.03; 0.01	0.35

4.2.4 EWAS Results

An EWAS was carried out using the Training dataset. Results were obtained for 362,404 probes since for some loci the model did not converge suggesting that the model results for those probes were not reliable. Additionally, probes known to cross-hybridise, probes on SNPs, and probes on the sex chromosomes were removed. At the FDR level (FDR $q < 0.05$; $p = 2.8 \times 10^{-5}$), 204 probes were found to be significantly associated with air PAH8 exposure, of which 26 were significant at the strict Bonferroni threshold ($p < 1.38 \times 10^{-7}$) (Figure 4.3A). Hypomethylation was observed at 130 of the FDR-significant probes, and hypermethylation at 71 (Figure 4.3B), while 14 Bonferroni probes were found to be hypomethylated compared to 12 hypermethylated CpG sites. Some inflation of the P values was observed, and the inflation factor λ was calculated to be 1.14 (Figure 4.3C).

Since the beta coefficients of the beta regression models were not interpretable as a percentage in methylation per unit increase in air PAH8 exposure, a mixed effects model was applied to the FDR significant probes to obtain interpretable coefficients. Approximately 34% of probes showed a minimum of 1% change in methylation per ng/m^3 increase in air PAH8 exposure. The biggest differences observed were 4.2% hypomethylation (cg05703053; subject with lowest air PAH8 exposure = 57.9% methylated; subject with highest air PAH8 exposure = 26.8% methylated) and 3.32% hypermethylation (cg27286337; subject with lowest air PAH8 exposure = 75.7% methylated; subject with highest air PAH8 exposure = 95.8% methylated) per unit change in air PAH8 exposure.

Models for the 26 Bonferroni-significant probes were run on the Testing and EPIC-NL datasets in order to identify any replication. Due to the reduced number of probes passing quality control in the EPIC-NL dataset, 4 of the 26 probes were not available in this dataset. The model results for these probes in all datasets are summarised in Table 4.7. None of the probes were statistically significant in all datasets, however, 3 were significant ($p < 0.05$) in the Testing dataset (cg18031747, cg14494451, and cg12826791) (Table 4.7). 8 probes showed a consistent direction of change, and the 4 probes missing in the EPIC-NL dataset were consistent across the Training and Testing datasets (Table 4.7).

The results of the 204 FDR significant probes in all three datasets can be found in Appendix 2 (Table 9.7).

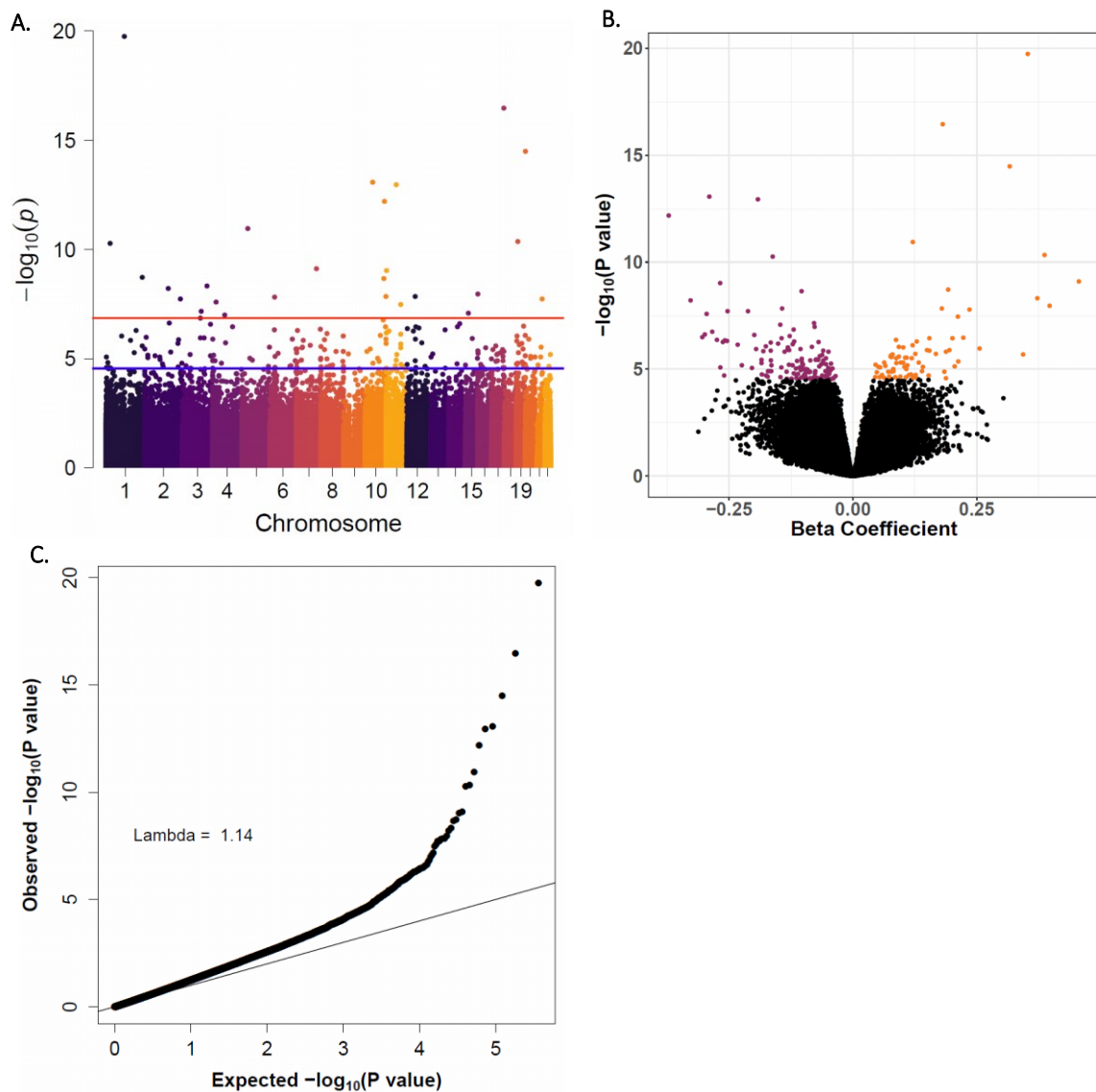


Figure 4.3 A: Manhattan plot showing the $-\log_{10}$ transformed p values of the 362,404 CpG probes tested arranged by chromosome. The red line indicates the threshold for Bonferroni correction for multiple testing ($p < 1.38 \times 10^{-7}$), and the blue line indicates the FDR threshold ($FDR q < 0.05; p = 2.8 \times 10^{-5}$). **B:** Volcano plot showing the $-\log_{10}$ transformed p values of the 362,404 CpG probes against the β -coefficient for dietary PAH8 intake. Coloured points indicate significance after FDR correction, with red points indicating a decrease in methylation and orange points indicating an increase in methylation. **C:** QQ plot showing the observed $-\log_{10}$ transformed p values against the expected $-\log_{10}$ transformed p values from the EWAS.

Table 4.7 Model results for the Bonferroni significant ($p < 1.38 \times 10^{-7}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship between air PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status.

Probe ID	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-NL - EPIC-NL (N=132)		
	B Coefficient	95% Confidence Interval	P Value	B Coefficient	95% Confidence Interval	P Value	B Coefficient	95% Confidence Interval	P Value
cg00466488	0.353	0.278; 0.428	1.82E-20	0.005	-0.117; 0.128	0.932	0.333	-0.162; 0.829	0.187
cg18576374	0.181	0.139; 0.224	3.44E-17	0.012	-0.072; 0.096	0.779	0.049	-0.316; 0.413	0.794
cg14083397	0.316	0.238; 0.395	3.30E-15	-0.011	-0.076; 0.054	0.738	0.019	-0.362; 0.400	0.921
cg14677909	-0.290	-0.366; - 0.214	8.55E-14	0.152	-0.081; 0.385	0.200	-0.352	-0.846; 0.142	0.163
cg01981334	-0.192	-0.242; - 0.141	1.13E-13	0.010	-0.038; 0.058	0.681	0.091	-0.161; 0.342	0.480
cg15275103	-0.371	-0.473; -0.27	6.38E-13	0.042	-0.142; 0.225	0.657	0.062	-0.396; 0.520	0.791
cg26496372	0.121	0.086; 0.156	1.15E-11	0.009	-0.058; 0.076	0.791	-0.047	-0.253; 0.159	0.654
cg18031747	0.387	0.272; 0.502	4.47E-11	0.150	0.034; 0.266	0.011	NA	NA	NA
cg12653146	-0.162	-0.21; -0.113	5.33E-11	0.025	-0.028; 0.078	0.362	0.164	-0.188; 0.515	0.361
cg04678743	0.456	0.311; 0.602	7.92E-10	0.008	-0.262; 0.279	0.952	NA	NA	NA
cg27261733	-0.268	-0.354; - 0.182	9.45E-10	-0.040	-0.133; 0.052	0.390	-0.188	-0.696; 0.321	0.469
cg03317082	0.192	0.130; 0.255	1.92E-09	0.021	-0.072; 0.114	0.661	0.139	-0.331; 0.610	0.562
cg23881299	-0.104	-0.137; -0.07	2.23E-09	0.027	-0.024; 0.077	0.301	0.015	-0.250; 0.280	0.912
cg18592273	0.372	0.248; 0.497	4.63E-09	0.057	-0.170; 0.283	0.623	NA	NA	NA

cg14494451	-0.327	-0.438; - 0.217	6.06E-09	0.089	0.002; 0.176	0.045	0.254	-0.334; 0.842	0.397
cg07482202	0.397	0.261; 0.533	1.09E-08	0.140	-0.070; 0.350	0.192	0.225	-0.657; 1.107	0.616
cg18308755	-0.143	-0.192; - 0.093	1.42E-08	0.038	-0.038; 0.115	0.326	-0.286	-0.706; 0.135	0.183
cg02574894	0.179	0.117; 0.242	1.45E-08	0.006	-0.086; 0.097	0.905	0.267	-0.189; 0.722	0.251
cg16495982	0.235	0.154; 0.317	1.58E-08	0.016	-0.128; 0.160	0.832	NA	NA	NA
cg12826791	-0.253	-0.341; - 0.165	1.89E-08	0.095	0.006; 0.184	0.037	0.146	-0.245; 0.538	0.464
cg22374586	-0.212	-0.286; - 0.138	1.92E-08	-0.103	-0.219; 0.014	0.083	0.341	-0.299; 0.980	0.297
cg13291296	-0.295	-0.399; - 0.191	2.64E-08	0.067	-0.193; 0.328	0.613	0.125	-0.817; 1.067	0.795
cg22049858	0.212	0.137; 0.287	3.42E-08	-0.049	-0.180; 0.081	0.461	-0.490	-1.052; 0.073	0.088
cg10178498	-0.078	-0.107; -0.05	6.88E-08	0.012	-0.041; 0.065	0.654	0.169	-0.061; 0.399	0.151
cg14209037	-0.146	-0.200; - 0.093	8.12E-08	0.012	-0.079; 0.104	0.794	0.126	-0.331; 0.584	0.589
cg03931518	-0.077	-0.106; - 0.049	1.04E-07	-0.011	-0.050; 0.027	0.566	-0.112	-0.383; 0.159	0.419

The methylation status of the probes with the biggest methylation changes (cg27286337 and cg05703053) were assessed in the subjects with the lowest and highest air PAH8 exposures in the Testing and EPIC-NL datasets. In the Testing dataset, probe cg27286337 in the subject with the lowest exposure was 79.8% methylated, and the most highly exposed subject was 92.4% methylated. These observations are very similar to those observed in the Training dataset outlined in the previous paragraph, and the model statistics for this probe were similar in the Training and Testing datasets (Appendix 2 Table 9.7). Probe cg27286337 was not available in the EPIC-NL dataset. Probe cg05703053 was 50.1% methylated in the subject with the lowest air PAH8 exposure and 60.1% methylated in the subject with the highest air PAH8 exposure in the Testing dataset. In the EPIC-NL dataset, there was no difference in the methylation of probe cg05703053 between the most highly and lowly exposed subjects (22.0% and 22.5% respectively).

Table 4.8 summarises the characteristics of the 26 Bonferroni-significant probes. The majority of the loci were located either within gene body or promoter-associated regions. Half of the probes were also located within a CpG island, with a further five in shore or shelf regions flanking islands. Table 9.8 in Appendix 2 shows the characteristics of the 204 FDR-significant probes.

Table 4.8 Table of characteristics of probes found to be significantly associated with air PAH8 exposure at the Bonferroni level ($p < 1.38 \times 10^{-7}$) in the training dataset.

Probe ID	Chromosome	Position	UCSC RefGene Name	Gene Location	Relation to CpG Island	Methylation Change Direction
cg12653146	1	25919290				-
cg00466488	1	118148927	<i>FAM46C</i>	5'UTR	Island	+
cg03317082	1	234748618			South Shore	+
cg14494451	2	153575552	<i>ARL6IP6</i>	Body	Island	-
cg22374586	2	232220566				-
cg10178498	3	124103021	<i>KALRN</i>	Body		-
cg18592273	3	161089930	<i>C3orf57</i>	TSS200	Island	+
cg13291296	4	22390126	<i>GPR125</i>	Body		-
cg03931518	4	79106308	<i>FRAS1</i>	Body	South Shelf	-
cg26496372	5	37379396	<i>WDR70</i>	TSS200	Island	+
cg16495982	6	30641015	<i>DHX16</i>	TSS200	South Shore	+
cg04678743	7	130353515	<i>TSGA13</i>	3'UTR	Island	+
cg14677909	10	48807341	<i>PTPN20B</i>	Body		-
cg23881299	10	121116990	<i>GRK5</i>	Body		-
cg15275103	10	124893024			Island	-
cg18308755	10	134065956	<i>STK32C</i>	Body		-
cg27261733	11	1891872	<i>LSP1</i>	5'UTR	North Shore	-
cg01981334	11	64877237	<i>C11orf2</i>	Body	North Shore	-
cg22049858	11	94884121			Island	+
cg02574894	12	53693825	<i>C12orf10</i>	Body	Island	+
cg14209037	15	41228521	<i>DLL4</i>	Body	Island	-
cg07482202	16	745687	<i>FBXL16</i>	Body	Island	+
cg18576374	17	78549371	<i>RPTOR</i>	Body		+
cg18031747	19	9929709	<i>FBXL12</i>	1stExon	Island	+
cg14083397	20	388473	<i>RBCK1</i>	TSS1500	Island	+
cg12826791	21	45926719	<i>C21orf29</i>	Body	Island	-

When looking at the distribution of the 204 FDR-significant probes across various genomic regions, it was found to be similar to the underlying distribution of sites tested (N = 362,404) (Table 4.9; Figure 4.4A). The only deviations from the expected distribution were a statistically significant enrichment of methylation changes occurring at exons (OR = 2.14; p = 0.0002), and significantly less methylation changes at promoters than expected (OR = 0.69; p = 0.03) (Table 4.9; Figure 4.4A). When comparing the ratio of hypomethylation events to hypermethylation events stratified by genomic location, exons (OR = 2.82; p = 0.028) and introns (OR = 2.36; p = 0.026) were observed to have significantly more hypomethylation events than expected, with more hypermethylation events occurring within LINE regions (OR = 0.089; p = 0.0098) (Table 4.10; Figure 4.4B).

In some instances, multiple FDR-significant CpG sites located close together in the genome were found to be differentially methylated. In order to determine whether other CpG sites in the vicinity of these probes were also differentially methylated but did not pass the significance threshold, the model coefficients of all CpG sites within a 2 kb region were checked. Two interesting regions emerged within the promoter regions of ADAM32 (Figure 4.5) and RP11-712B9.2 (Figure 4.6). Two CpG sites in the promoter of ADAM32 passed FDR correction (cg22848598 and cg26394257) and both showed similar levels of hypomethylation (1.77% and 1.11% methylation per ng/m³ increase in PAH8 exposure respectively). Four additional CpG probes around these were also found to be similarly hypomethylated, one of which was borderline FDR-significant and the rest all had p values less than 0.005 (Figure 4.5). Four FDR-significant probes located in the promoter region of RP11-712B9.2 were all found to be hypermethylated by between 0.61% to 0.96% methylation per ng/m³ increase in PAH8 exposure (Figure 4.6). One other adjacent probe was also found to be hypermethylated to similar levels and had a p-value of less than 0.005.

Table 4.9. Table of Fisher’s Exact Test results comparing the number of differentially methylated probes (N = 274) and all tested probes (N = 362,394) in the training dataset EWAS at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred.

Genomic Region	Odds Ratio	Confidence Interval	P Value
3' UTR	0.69	0.18 – 1.81	0.67
5' UTR	1.36	0.49 – 3.02	0.46
Exon	2.14	1.42 – 3.14	0.0002
Intergenic	0.85	0.53 – 1.32	0.53
Intron	1.17	0.83 – 1.63	0.35
Non-coding	1.47	0.47 – 3.49	0.40
Promoter	0.69	0.48 – 0.97	0.03
TTS	1.06	0.34 – 2.51	0.81
CpG Island	0.91	0.53 – 1.47	0.81
LINE	1.09	0.43 – 2.30	0.69
SINE	0.64	0.17 – 1.66	0.54
LTR	0.82	0.22 – 2.13	1
Other	0.61	0.074 – 2.24	0.78

Table 4.10. Table of Fisher’s Exact Test results comparing the number of hypermethylation changes (N = 99) to hypomethylation changes (N = 175) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected.

Genomic Region	Odds Ratio	Confidence Interval	P Value
3' UTR	Inf	0.38 – Inf	0.30
5' UTR	0.28	0.024 – 1.98	0.19
Exon	2.82	1.06 – 8.83	0.028
Intergenic	0.48	0.18 – 1.26	0.11
Intron	2.36	1.08 – 5.51	0.026
Non-coding	0.14	0.0027 – 1.42	0.059
Promoter	0.72	0.35 – 1.52	0.38
TTS	0.37	0.031 – 3.32	0.36
CpG Island	1.53	0.49 – 5.72	0.61
LINE	0.089	0.0019 – 0.75	0.0098
SINE	1.72	0.14 – 91.71	1
LTR	0.56	0.040 – 7.94	0.62
Other	0.57	0.0072 – 45.02	1

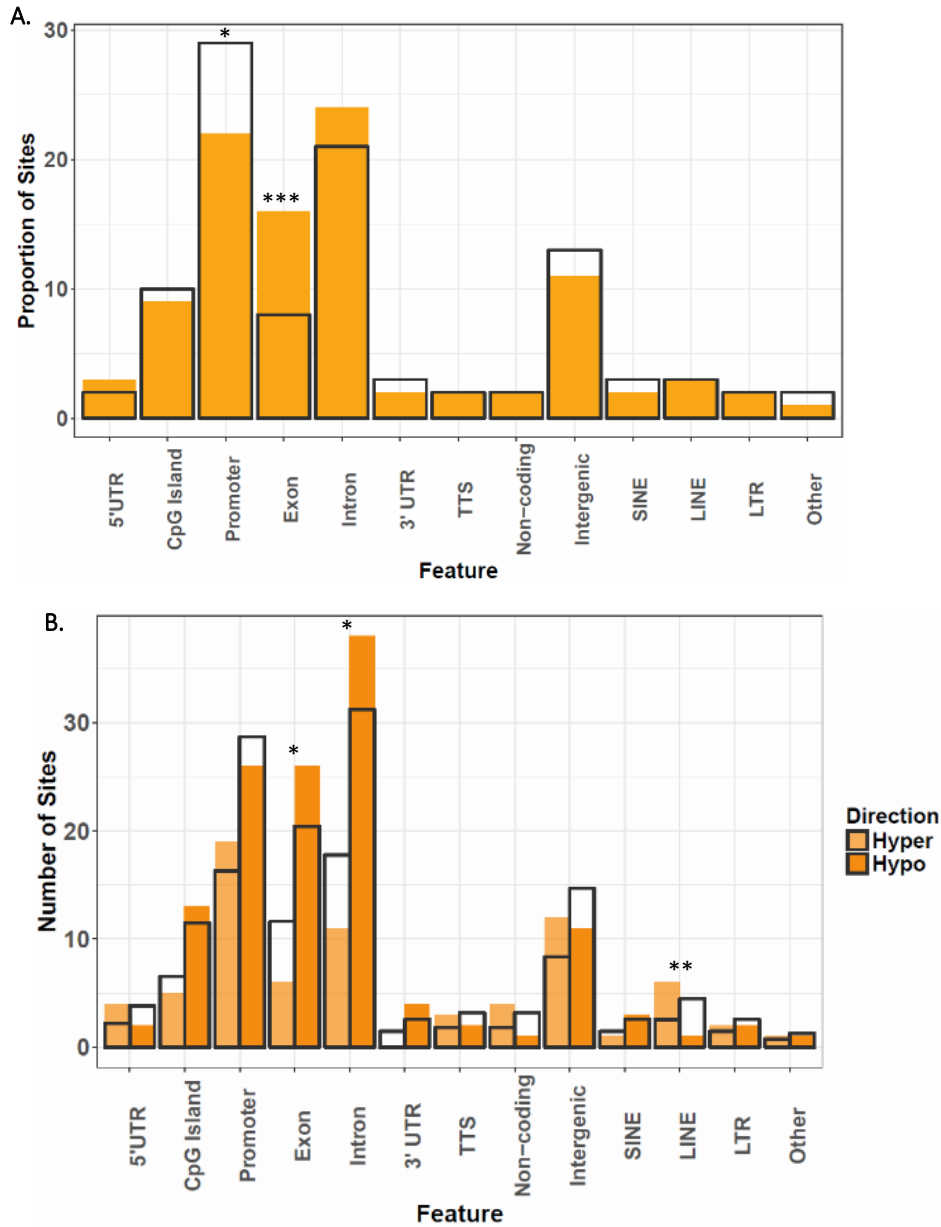


Figure 4.4 A: Comparison of the genomic distribution of differentially methylated probes (N = 204) and all tested probes (N = 362,404) in the training dataset EWAS. The filled yellow bars show the proportion of significant probes, the grey outline bars show the proportion of all probes tested, i.e. the expected distribution. **B:** Comparison of the genomic distribution of hypermethylated (N = 74) and hypomethylated (N = 130) probes. The filled yellow bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. For both plots, * indicates $p < 0.05$, ** indicates $p < 0.01$ and *** indicates $p < 0.001$ following Fisher's Exact test.

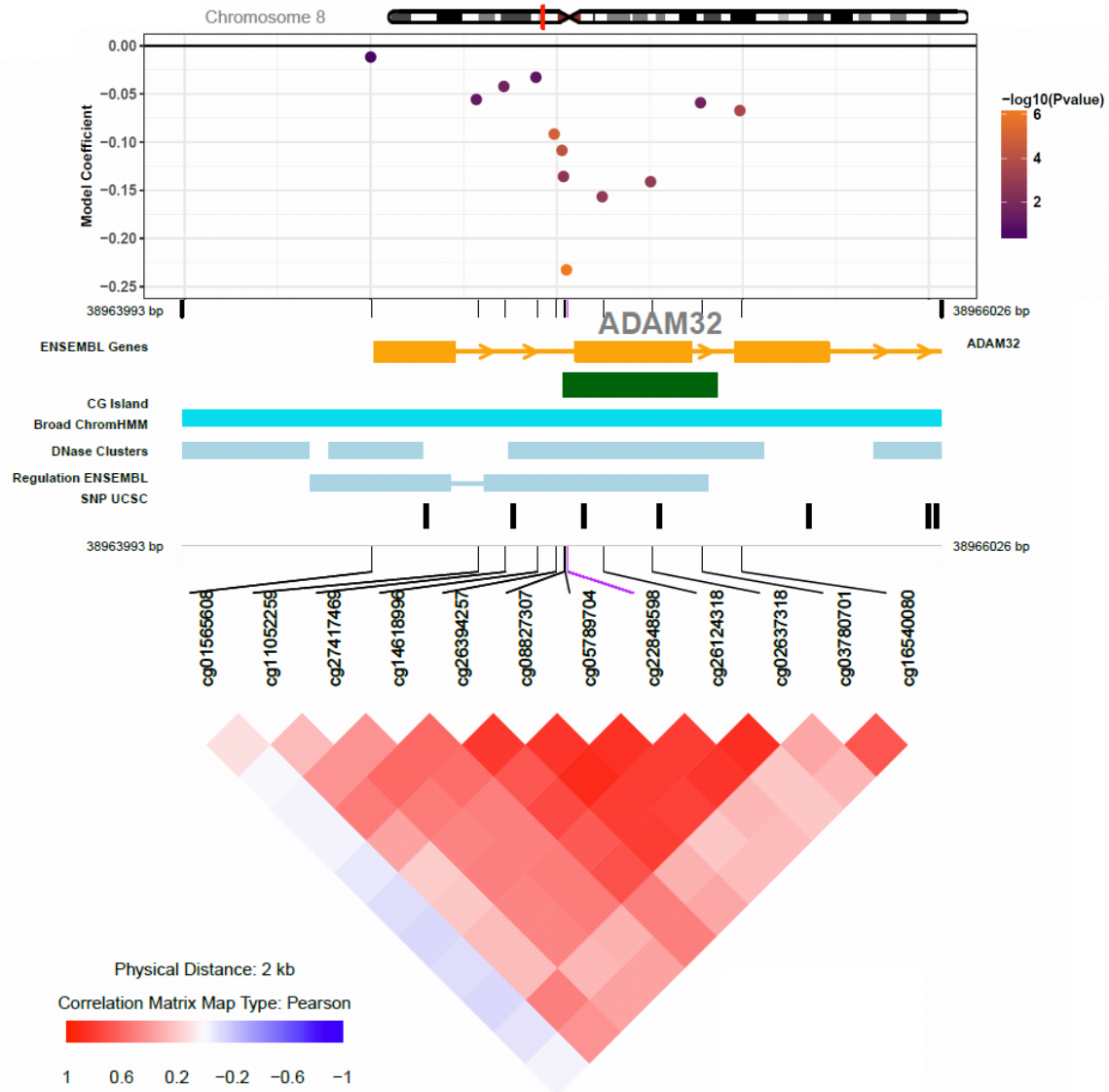


Figure 4.5. Modified coMET plot of a 2kb region on chromosome 8 where multiple CpG sites were found to be hypomethylated. The top panel of the figure is a regional association plot showing the beta coefficients from the beta regression models of these probes from the EWAS by genomic position. The colour of the points corresponds to the \log_{10} P value, where orange indicates FDR-significant probes (FDR $q < 0.05$; $p = 2.8 \times 10^{-5}$). The central panel shows the genomic landscape of the region with respect to genes, CpG islands, chromatin state, clusters of DNase, SNPs and regulatory features. The bottom panel shows a correlation matrix of the methylation values for each probe where red indicates a strong positive correlation, blue a strong negative correlation, and white a lack of correlation between the probes.

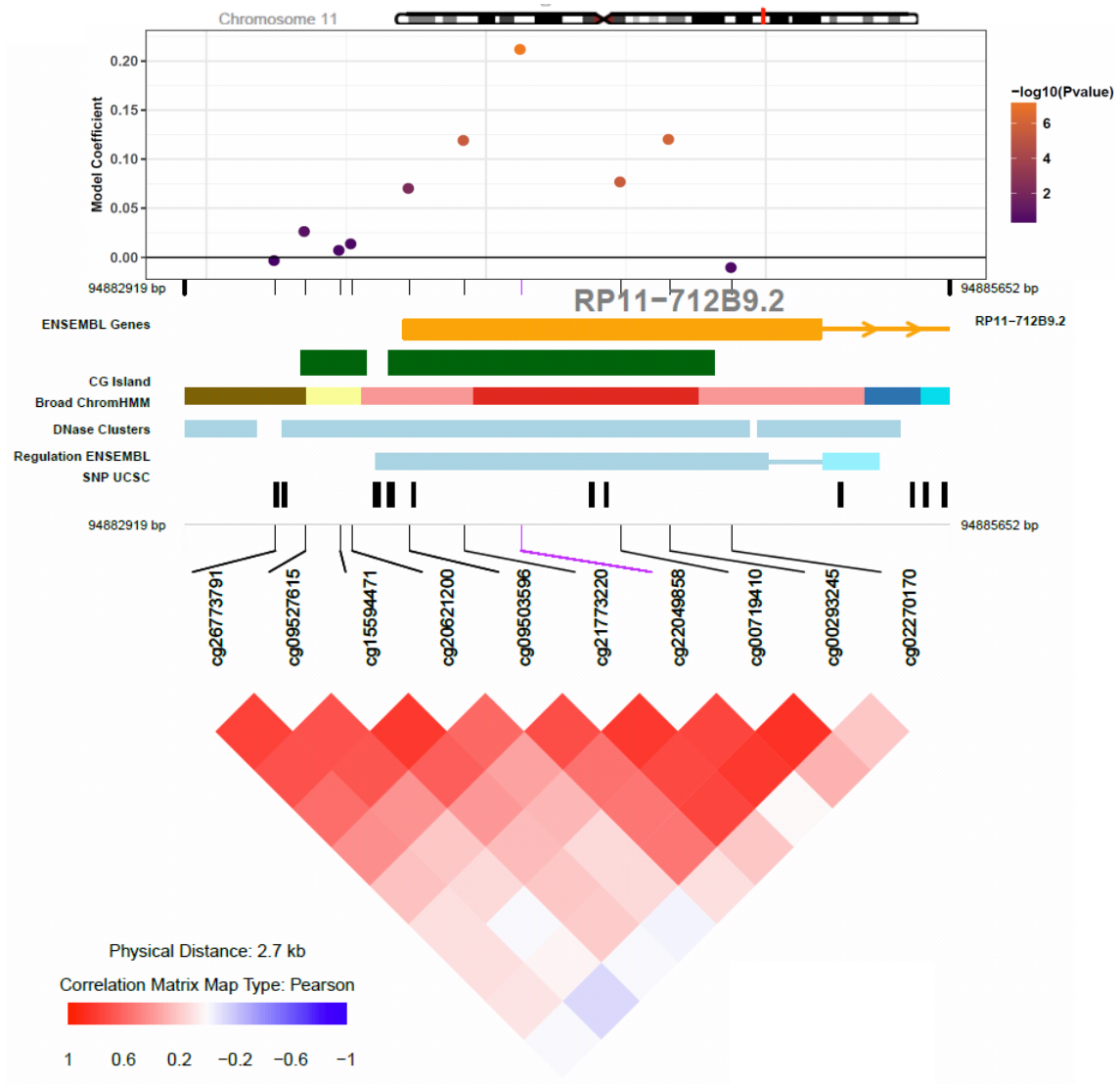


Figure 4.6. Modified coMET plot of a 2kb region on chromosome 11 where multiple CpG sites were found to be hypomethylated. The top panel of the figure is a regional association plot showing the beta coefficients of these probes from the EWAS by genomic position. The colour of the points corresponds to the \log_{10} P value, where orange indicates FDR-significant probes (FDR $q < 0.05$; $p = 2.8 \times 10^{-5}$). The central panel shows the genomic landscape of the region with respect to genes, CpG islands, chromatin state, clusters of DNase, SNPs and regulatory features. The bottom panel shows a correlation matrix of the methylation values for each probe where red indicates a strong positive correlation, blue a strong negative correlation, and white a lack of correlation between the probes.

4.2.5 Building a Methylation Index of Air PAH8 Exposure

The 204 differentially methylated probes identified from the EWAS were used to build a methylation index of air PAH8 exposure. A model was built which included all 204 CpG sites, in addition to age, sex, cancer status, and smoking status. This model was trained on the Training dataset, and its performance was tested in the Testing dataset only due to the differences in methylation distribution between the EPIC-Italy derived datasets and the EPIC-NL dataset. The optimal model was a ridge regression model ($\alpha = 0$) which meant that all probes and covariates were included, and had a penalty factor (λ) of 1.26. The RMSE of this model was 0.44 ng/m³ PAH8 exposure, but it only explained 27.9% of the variance ($R^2 = 0.279$) in the Training dataset. As expected, since the model was built on the Training dataset, it performed well on the same data (Figure 4.7A) with the predicted exposures highly, and significantly positively correlated with the real exposure estimates (Spearman's Rho = 0.78, $p < 2.2 \times 10^{-16}$). Model performance decreased dramatically when applied to the Testing dataset (Figure 4.7B). The predicted and real PAH8 exposures were only weakly correlated, but this was still statistically significant (Spearman's Rho = 0.19, $p = 0.007$).

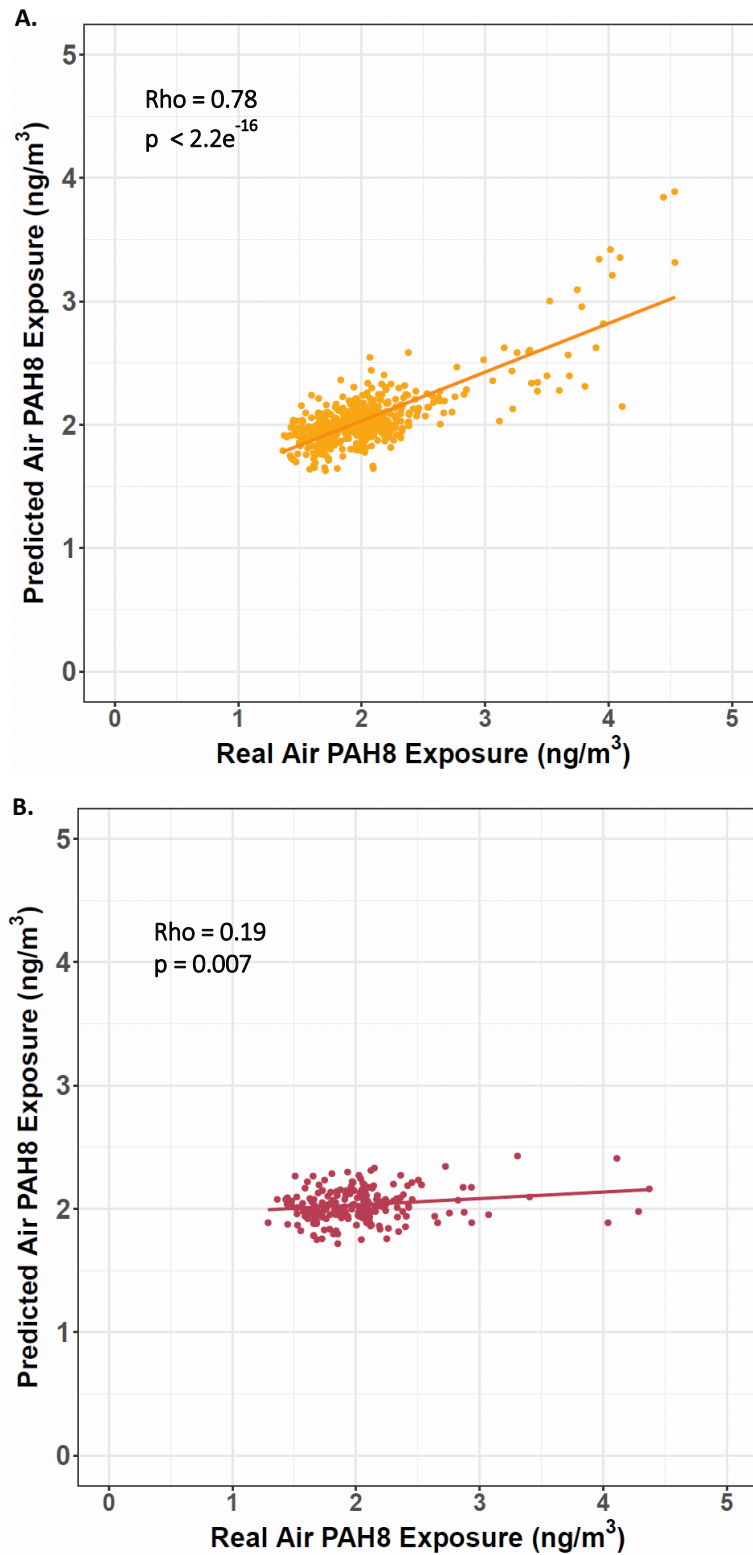


Figure 4.7 Plots showing the correlation between the air PAH8 exposure predicted by the elastic net model against the real air PAH8 exposure for each subject. Figures A and B show the results from the full model for the training and testing sets respectively.

4.3 Discussion

This is the first epidemiological study investigating the methylome-wide effects of air exposure to the eight most carcinogenic PAHs. Previous studies have investigated the effects of PAH exposure on the methylation of various genetic loci as summarised in the introduction of this chapter. In the results presented here, differential methylation of 204 CpG probes was found to be associated with air PAH8 exposure estimated using land-use regression models. Of these changes, significantly more than expected occurred in exon regions, and less than expected occurred at promoter regions. Multiple CpG probes in the promoter region of the ADAM32 and RP11-712B92 genes were found to be differentially methylated. Finally, a methylation index of air PAH8 exposure was developed and tested in an independent dataset, with a weak, but statistically significant performance in the Testing dataset.

4.3.1 Air PAH8 Exposure

The distribution of PAH exposures in the Training and Testing datasets was reasonably spread out, particularly when compared to that of the EPIC-NL subjects which all had an almost identical exposure, but the range of the Italian subjects was still narrow (1.29-4.54 ng/m³) with the majority of subjects having an exposure around 2 ng/m³. The use of land use regression models to estimate exposures, while helpful when other measures are not available, has several drawbacks. The biggest drawback is that these models estimate exposure based on a single address, usually the home address, of each subject, however, the vast majority of people only spend a part of their day in their home. Additionally, depending on the environment of their homes compared to that of their workplaces (rural or urban), the models may under- or over-represent their exposure, resulting in misclassification errors. In the case of this current study, this misclassification has been exacerbated by the fact that the blood on which DNA methylation analysis was carried out was drawn in the early 1990's, whereas the measurements used in the model to estimate exposure were taken between 2008 and 2011. Given that PAH levels in air have been steadily declining, this would suggest that

the PAH8 exposures calculated using LURs from over a decade later would probably underestimate the real exposures.

4.3.2 Cohort Differences

There were fundamental differences between the EPIC-Italy sub-cohorts (Training and Testing) and the EPIC-NL dataset. The EPIC-NL dataset was made up of only women, who were never- or ex-smokers and had never been diagnosed with cancer. Additionally these women were comparatively older. These differences in anthropometric characteristics were compounded by statistically significant differences in mean global methylation and the mean methylation of the major genomic features. For these reasons, the EPIC-NL dataset was not included in the assessment of the methylation index as the model performance would have been significantly impacted. Additionally, these differences go some way to explain why none of the probes identified in the Training dataset replicated in the EPIC-NL dataset. The lack of replication between the Training and Testing datasets however, is not explained by such differences as none of the cohort or methylation characteristics were significantly different between the two. Some loci were statistically significant in the Testing dataset and there were CpG probes that were found to be hypo- or hypermethylated in all three cohorts but these were not significant. One possible explanation is the limited statistical power of the Testing dataset due to the low number of subjects. It is also important to point out that the statistical power of the Training dataset was also limited, but the analyses were carried out to maximise statistical power by having as many subjects as possible in the Training dataset, while also having a similar, independent dataset (Testing dataset) in which the findings could be tested. This could have been improved by maintaining the Training and Testing dataset as a single dataset, however, by doing so reasonable evaluation of the performance of the methylation index would have been impossible.

4.3.3 Statistical and Other Considerations

4.3.3.1 Statistical Considerations for EWAS

The question of statistical power has already been discussed above and is an important consideration in the interpretation of the results presented in this chapter. The inflation of the test statistics of the EWAS was measured and a moderate level of inflation ($\lambda = 1.14$) was observed which suggests the possibility of some false positive results. The results are not shown in this thesis, however, during the course of carrying out the analyses described here it was observed that inflation increased with additional covariates. This is at odds with the usual explanation of inflation which is that higher inflation indicates that there is some confounding or that important covariates were not included. Inflation has not been reported in the few EWAS studies that have used beta regression, which makes it difficult to compare and explain these observations and to know whether the method used to measure inflation is suitable for use on the test statistics from beta regression. Despite these considerations, it is not possible to discount unmeasured and/or unknown confounders which were not accounted for.

4.3.3.2 Methylation Index Performance

The performance of the methylation index developed using the results from the EWAS was reasonable in the Training dataset in which the probes were identified, but this decreased dramatically in the Testing and Training datasets. The methylation index explained only a small proportion of the variance in the Training dataset (27.9%) which suggests that one or several covariates that could better explain the variation were not included in the model. These covariates could have been CpG probes that were not identified in the EWAS or anthropometric variables.

In general, the methylation index tended to underestimate the exposure of subjects with higher exposures, and overestimate the exposure of subjects with lower real exposures. The purpose of building the methylation index was as a reverse test of the EWAS results – if the probes identified in the EWAS were associated with PAH exposure, then their methylation status should be able to

reasonably predict the exposure. This was not the case, which indicates that perhaps the EWAS model did not account for all necessary covariates, or identification of the most important methylation changes based on correction for multiple testing meant that other influential probes were left out. Future work could include assessing models with other covariates, however, this would require larger cohorts due to the number of adjustments already made in the model presented given the sample size.

4.3.3.3 Selection of CpG Probes of Interest

The differentially methylated loci identified in the EWAS were chosen based on their significance level, however, effect size (magnitude of methylation difference) has been used in other studies to identify probes of interest. The biggest methylation changes observed in the EWAS were $\approx 4\%$ methylation per unit increase in air PAH8 exposure, which would represent a methylation difference of $\approx 20\%$ between subjects with lowest and highest exposures in the Training dataset. This is not a small difference, however, the majority of the observed methylation differences were considerably smaller, corresponding to an overall difference of $\leq 4\%$ between the subjects at either end of the exposure range. It is important to consider whether these small differences have any biological consequences with respect to gene expression or chromatin state. While it seems unlikely, this would need to be assessed further before a conclusion can be reached.

4.3.3.4 Measuring DNA Methylation in Blood

There are further points to consider that may go some way to explain the results observed. Methylation measurements were made using DNA from blood samples and it has been well-established that DNA methylation changes are tissue-specific. The relatively small methylation changes reported here may be due to the fact that since blood is not a known target organ of PAH exposure, therefore these compounds only have small effects in blood. It is known, however, that PAH-DNA adducts do form in blood, indicating that that blood cells are capable of metabolising these

compounds. Additionally, macrophages have previously been reported to be susceptible to PAH exposure, specifically B[a]P, in a similar manner to other cells¹⁰⁰.

4.3.3.5 PAH Exposure and DNA Methylation

It is also possible, that at such low concentrations, the effects of exposure are limited or easily repaired. This is difficult to establish for several reasons. The *in vitro* and *in vivo* model experiments that have been used to characterise the effects of PAH exposure tend to use high concentrations, several orders of magnitude higher than those to which humans are exposed. Additionally, single PAH compounds are tested in these studies, and little is known about the behaviour of different PAH mixtures, with current hypotheses suggesting that mixture behaviour is dependent on the relative composition of the components in the mixture^{4,30-33}. It also has yet to be established whether the effects of PAH exposure on methylation are cumulative, or simply “switched on or off”. If the effects are indeed cumulative, this may explain the lack of correlation between exposures and biological markers in humans when length of exposure is not also taken into account, but rather a snapshot of exposure is used. Lastly, as discussed in previous and future chapters, the effects of PAH exposure on DNA methylation may be consequences of DNA adduct formation. While adducts form preferentially at guanines adjacent to methylated cytosines, the downstream consequences of this are not completely clear. Additionally, this preference does not necessarily implicate specific genes, meaning that adducts may form in any genes in any person, and that adduct patterns may differ from person to person along with DNA methylation patterns as well. Lastly, the levels of PAH-DNA adducts can vary from individual to individual due to differences in metabolism and the presence or absence of specific polymorphisms as described in section 1.2.4.1. Therefore, being able to correlate DNA methylation changes with genomic maps of DNA adducts is essential in order to understand how these biomarkers are linked.

4.3.4 Comparison of EWAS Findings to Previously Published Results

None of the differentially methylated loci or their associated genes had been previously reported to be associated with PAH or air pollution exposures with the exception of MGMT. Two previous studies assessing occupational exposures to PAHs reported hypomethylation of the MGMT gene promoter^{343,349} however, cg14677612 located in the promoter region of MGMT was found to be hypermethylated in the current study.

Unlike several previous studies^{343,345–347,358,360}, no differences were observed between quartiles of PAH exposure and global methylation, but the trend did indicate a statistically significant decrease in global methylation with increasing air PAH8 exposure. The results from previous epidemiological studies are contradictory, with some citing LINE-1 hypomethylation^{343,347,358}, others LINE-1 hypermethylation^{346,360}, and one study reporting both global hypo- and hypermethylation³⁴⁵. Given the heterogeneity of previous reports with respect to PAH exposure sources, measurement of PAH exposure, and the developmental stage at which exposure was considered, it is difficult to interpret them and reach an overall conclusion. The lack of statistically significant differences in global methylation between PAH quartiles observed in the analysis presented here may indeed be due to global methylation levels not changing with PAH exposure, but there may be other reasons for this. The first is that the air PAH8 exposure ranges in the three datasets are very narrow and quite low, meaning that the differences between subjects in the lowest and highest quartiles of exposure are very small. A further reason is that the global methylation was calculated using the average methylation of a small subset of CpG probes present on the Infinium HumanMethylation450 BeadChip array. This suggests that there may indeed be differences in global methylation, however, these are not captured adequately by the probes included on the array. Recently, a study of various air pollutants was able to observed differences in global methylation based on the average of all probes on the Infinium HumanMethylation450 BeadChip array¹⁹¹. This study also reported differences in the average methylation at CpG shores and shelves, and gene bodies¹⁹¹, but no changes were observed in the analysis presented above at any genomic features.

The Comparative Toxicogenomics Database (CTD) contains 12,723 unique gene interactions associated with at least one of the eight PAHs in PAH8. Of the significantly differentially methylated CpG sites identified in the EWAS, 92 of the 163 genes associated with the 204 FDR-significant probes have previously been reported in the CTD to have altered gene expression levels associated with PAH exposure¹⁸⁸. It is important to note that predominantly, the studies reported in the CTD were *in vitro* human cell-line experiments or *in vivo* animal model experiments. The CTD has reports of PAHs interacting with over half the genome since the human genome contains between 19,000 – 20,000 genes (approximately 63%), and the gene overlaps found here are about 56% (92 of 163 genes). While this suggests that overlaps identified are possibly due to chance, it also supports the theory that PAHs affect DNA methylation in a random pattern dependent on DNA adduct formation and chromatin structure.

The recently published study by Tryndyak *et al.* (2018)³²⁵ carried out RRBS on a B[a]P-exposed human liver cell line (HepaRG cell line). The authors noted over 6500 differentially methylated regions in these cells compared to controls. A comparison of the EWAS results presented above, and the findings by Tryndyak *et al.* (2018)³²⁵ found four genes reported in all studies, however the location or the differentially methylated sites and the direction of change were not consistent between the two studies. The results are shown in Appendix 2 Table 9.9.

One of the FDR-significant CpG probes, cg02583546 located in the promoter region of *C14orf4* was found to be hypermethylated here and in a previous study published by Besingi *et al.* (2013)³⁸⁰ who carried out an EWAS of smoking, a major source of PAH exposure. At the CpG level, no other overlaps were identified, however differential methylation of 54 of the genes associated with the 204 FDR-significant probes have been previously reported to be associated with tobacco smoke in previous studies. These results are summarised in Appendix 2 Table 9.10.

4.3.5 Conclusions

In summary, this chapter reports the findings of the first EWAS of air PAH8 exposure and shows that this exposure does have an effect on DNA methylation in WBCs. Overall, the results presented in this chapter have not been previously reported in any methylation studies assessing exposure to PAHs or air pollutants, with the exception of the trend towards global hypomethylation with increasing air PAH8 exposure. The differentially methylated probes identified were used to build a methylation index of air PAH8 exposure and this performed well in the Training dataset however this was not replicated in the Testing dataset.

5 Chapter 5 - DNA Methylation and Dietary PAH8 Exposure

5.1 Introduction

Dietary exposure to PAHs in humans has been extensively reviewed by academics and by various organisations such as the Food and Agriculture Organisation (FAO), the Scientific Committee on Food (SCF) and the European Food Safety Authority (EFSA)^{39,270,389,381-388}. The food chain may become contaminated by PAHs through two main pathways. The first is the presence of PAHs in air which then pollute the soil and water thereby contaminating foods from these sources^{39,383,390,391}. The second route is through the processing of food such as drying or smoking, as well as the cooking of food using sources where combustion is not always complete^{39,383,392}. Less significant methods of transfer are through the addition of contaminated smoke flavourings and packaging materials³⁸³. Additionally, PAHs have been reported to bio-accumulate in food webs¹⁷.

5.1.1 Recommendations and Policies

Various committees have defined different groups of priority PAHs in food shown in Table 5.1. The Scientific Committee on Food (SCF) was the first to determine that the following 15 PAHs in foods pose the biggest dietary risk to humans based on the *in vivo* evidence of their genotoxicity and mutagenicity: 5-methylchrysene (5mC), benz[a]anthracene (B[a]A), benzo[b]fluoranthene (B[b]Fl), benzo[j]fluoranthene (B[j]Fl), benzo[k]fluoranthene (B[k]Fl), benzo[ghi]perylene (B[ghi]P), benzo[a]pyrene (B[a]P), chrysene (Chr), cyclopenta[cd]pyrene (C[cd]P), dibenz[a,h]anthracene (DB[a,h]A), dibenzo[a,e]pyrene (DB[a,e]P), dibenzo[a,h]pyrene (DB[a,h]P), dibenzo[a,i]pyrene (DB[a,i]P), dibenzo[a,l]pyrene (DB[a,l]P) and indeno[1,2,3-cd]pyrene (I[cd]P)³⁸⁶. Following the recommendations of the SCF, the European Commission (EC) issued regulation 208/2005 setting maximum permissible levels of B[a]P in foods with a high oil content that undergo smoking or drying procedures³⁹³. The Joint FAO/WHO Expert Committee on Food Additives (JECFA) reached similar conclusions as the SCF but recommended that benzo[c]fluorene (B[c]F) is also monitored, and that

the carcinogenicity of both B[ghi]P and C[cd]P is unclear and requires further investigation ³⁹⁴. Both the SCF and JECFA concluded that B[a]P is a suitable marker for dietary exposure to PAHs ^{382,394}.

Table 5.1 Table showing groups of PAHs as described by various European committees. Shading indicates that a PAH is included in that group.

PAHs	Groups of PAHs				
	JECFA ³⁹⁴ 15+1	SCF ³⁸² 15	CONTAM ²⁸³ PAH8	CONTAM ²⁸³ PAH4	CONTAM ²⁸³ PAH2
<u>Benz[a]anthracene</u>					
<u>Benzo[b]fluoranthene</u>					
<u>Benzo[j]fluoranthene</u>					
<u>Benzo[k]fluoranthene</u>					
<u>Benzo[c]fluorene</u>					
<u>Benzo[ghi]perylene</u>					
<u>Benzo[a]pyrene</u>					
<u>Chrysene</u>					
<u>Cyclopenta[cd]pyrene</u>					
<u>Dibenz[a,h]anthracene</u>					
<u>Dibenzo[a,e]pyrene</u>					
<u>Dibenzo[a,h]pyrene</u>					
<u>Dibenzo[a,i]pyrene</u>					
<u>Dibenzo[a,l]pyrene</u>					
<u>Indeno[1,2,3-cd]pyrene</u>					
<u>5-methylchrysene</u>					

The EC recommended in regulation 2005/108 ³⁹⁵ that the levels of PAHs in food be investigated further which resulted in over 10,000 results for PAH level from 18 member states. Of these, B[a]P was found in approximately 50% of the samples, while 30% did not contain B[a]P but detectable levels of genotoxic and carcinogenic PAHs were measured ²⁸³. Based on these samples, the EFSA Panel on Contaminants in the Food Chain (CONTAM) calculated the median dietary exposure in mean consumers and high consumers within Europe. The results are shown in Table 5.2. A near perfect

correlation was found between PAH4 and PAH8, and since they are a measure of carcinogenic PAHs, this led the CONTAM panel to conclude that they may be used as suitable indicators of PAH content in food, with PAH8 not adding much value compared to PAH4. The panel also recommended against the use of B[a]P as a marker of exposure and against using the toxic equivalency factor (TEF) approach because these do not represent the modes of action of PAH mixtures which results in an inability to accurately anticipate their carcinogenic potency. Following this, the EC issued regulation 835/2011 amending previous regulations and recommending maximum PAH4 levels and that B[a]P concentrations should still be monitored to allow comparability to previous assessments ³⁹⁶.

Table 5.2 Median dietary intake in European consumers²⁸³

<u>Measurement</u>	<u>Average Consumers</u>	<u>High Consumers</u>
B[a]P only	235 ng/day	389 ng/day
PAH2	641 ng/day	1077 ng/day
PAH4	1168 ng/day	2068 ng/day
PAH8	1729 ng/day	3078 ng/day

5.1.2 Incidence of PAHs in Food

Most dietary PAH exposure is from charcoal-broiled and smoked meats, leafy vegetables, grains, fats and oils ²¹. The JECFA reported that the foods that contribute the highest PAH exposure are cereals, vegetable fats and oils, and vegetables ³⁹⁷. The Scientific Cooperation on Food (SCOOP) taskforce reported in 2004 that high levels of PAHs were found in dried fruits, olive pomace oil, smoked fish, grape seed oil, smoked meat products, fresh molluscs, spices, sauces and condiments ³⁹⁸. The CONTAM panel also found that cereals and cereal products had the highest concentrations of carcinogenic PAHs the most PAH exposure, followed by seafood and seafood products ²⁸³. The contributions of smoked meats and fish have been shown to vary depending on the extent to which these foods play a role in the diet ³⁹⁷, with the level of contribution from other meat products depending on the amount of consumption and cooking method ³⁹. The FSA found no PAHs of the PAH8 group in canned vegetables, eggs and milk in their most recent Total Diet Study (TDS) ²⁴⁵ and

while all levels were below those specified by EC regulation 835/2011, those in sugars and preserves, and other vegetables were higher than the 2001 TDS ²⁷².

The levels of PAHs in certain food groups such as vegetables and grains are dependent on the concentrations present in the surrounding air and soil ²¹. In such foods, PAHs are not metabolised since they take no part in translocation given that they are lipophilic and therefore these are passed on to the consumer ¹⁷. Vegetation grown in industrial or urban areas may have a PAH concentration up to 5 orders of magnitude higher than those grown in rural areas ¹⁷. Consequently, diet type must be taken into consideration when considering an individual's PAH exposure. It has been reported by Menzie *et al.* (1992) that a heavy meat diet contributes to a higher dietary PAH exposure compared to a vegetarian diet due to higher PAH content in meats, despite vegetables also containing a high concentration of PAHs. However, a balanced diet was reported by the authors as having the lowest dietary PAH content due to the relatively lower consumption of meat and vegetables compared to heavy meat and vegetarian diets²¹.

5.1.3 Human Exposure to PAHs in Food

In Chapter 1, a number of studies linking dietary B[a]P exposure to colorectal adenoma were discussed. The findings of these studies showed that increased dietary B[a]P exposure was associated with a moderately increased risk of colorectal adenoma ^{115,116,399}. However, a further two studies found no association ^{117,118}.

A study characterising the pollutant mixtures that pregnant French women are exposed to identified a mixture containing PAHs as well as trace elements and furans as one of the major mixtures and recommended the future monitoring of these compounds ⁴⁰⁰. Another study of pregnant women showed that higher consumption of PAH-rich foods was associated with lower birth weight ⁴⁰¹. PAHs have been shown to cross the placenta in mice and rats ^{48,402}, and PAHs have been detected in human breast milk ⁴⁰³ indicating that exposure to PAHs occurs throughout the entire life course.

A study on the presence of PAHs in food carried out in the Netherlands in the 80's found that B[b]Fl, B[k]Fl and Fl were the most frequently occurring compounds ¹⁹⁸. This same study also reported 5 µg/day as a low estimate for dietary PAH exposure and 17 µg/day as a high estimate. Lioy *et al.* 1988 (as cited in Menzie *et al.* 1992²¹) reported that in the late 1980's the range of B[a]P levels as measured from 58 meals prepared in Phillipsburg, New Jersey was 0.004-1.2 µg/kg of wet food. A Chinese study of the dietary PAH intake of 100 subjects reported that incremental lifetime cancer risk of dietary PAHs was 6.65×10^{-5} ⁴⁰⁴ and dietary intake of B[a]P has been shown to be associated with colorectal adenoma in a case-control study of 146 cases and 228 controls ¹¹⁵. Another study reported that every 10 ng of B[a]P consumed per day corresponded to a 6% increased risk of large colorectal adenoma ¹¹⁶. However other studies have reported a null association between dietary B[a]P intake and colorectal cancer ^{117,118} and reported that smoked meat consumption did not pose a significant carcinogenic risk ⁴⁰⁵. Recently, ingestion of charbroiled foods has been shown to be associated with PAH-DNA adduct levels in peripheral lymphocytes in humans ⁴⁰⁶.

The CONTAM Panel reported that the UK had the lowest population dietary intake of PAH8 (1415 ng/day) while Norway had the highest (2136 ng/day) which may be explained by the increased consumption of smoked foods in Scandinavian diets however, the population intakes of the other Scandinavian countries were all around or below the European median ²⁸³. Median dietary exposures in Europe are summarised in Table 5.2 above. A recent review carried out by Ma and Harrad ⁴⁰⁷ cited studies carried out in the UK, Italy and the Netherlands in the 80's and 90's and the population dietary intakes were significantly higher, approximately twice as much as those cited above from more recent years. Also, PAH intakes in the UK have been shown to have decreased between 1979 and 2000 ²⁷².

This may possibly be explained by two factors. The first is that given the EC regulations published in recent years, more care has been taken to ensure that foods are exposed to minimal sources of PAHs. Some examples of this include growing food items such as vegetables away from industrial areas and busy roads, both of which are significant sources of PAHs ^{245,391,397} and the use of liquid smokes ²⁴² instead of traditional smoking methods. The second factor is that in recent years, the amount of PAHs

present in air have decreased ¹⁴, which is likely to also have affected the concentrations of PAHs in soil and water.

5.1.4 Measuring Human Exposure to Dietary PAHs

Bansal *et al.* (2017) reviewed the measurement techniques currently available for measuring PAHs in food samples ⁴⁰⁸. PAHs first need to be extracted and cleaned-up from food samples before detection can take place⁴⁰⁸. Currently, gas chromatography coupled with mass spectroscopy (GC-MS) is the most sensitive method for detection and is one of the most common techniques used alongside liquid chromatography coupled with mass spectroscopy (LC-MS), and high performance liquid chromatography (HPLC) with a fluorescence detector coupled with MS ⁴⁰⁸. For increased sensitivity with small sample amounts chromatographic techniques with advanced detectors are the leading choices ⁴⁰⁸. New techniques are currently in development which include bio-sensing that make use of immunoassays and enzymatic assays, in addition to electrochemical (amperometric/ voltametric) techniques which require less sample preparation ⁴⁰⁸.

As for measuring intake, multiple methods exist including the duplicate plate method, total diet studies (TDS), and food frequency questionnaires (FFQs). The WHO, FAO, and EFSA released joint guidelines on the best approach for TDSs in order to ensure comparability across multiple studies ⁴⁰⁹. TDSs support public health policies by allowing the exposure of the population to both harmful and beneficial food components to be measured. For FFQs to be used in the estimation of dietary PAH intake, they must be used in conjunction with duplicate plate methods or databases containing the concentrations of PAHs in various foods ^{404,410}. The duplicate plate method allows for more accurate measurements of PAH exposure since they are carried out on identically prepared foods, however this method is not always feasible, particularly in large or longitudinal studies. Creating databases of PAH concentrations in foods allows them to be applied to any number of subjects as long as they have filled in a FFQ. However this method also has several limitations, and is particularly prone to bias and

misclassification errors. FFQs and duplicate plate methods are both susceptible to misclassification error if subjects' responses or plates are not truly representative of their normal intakes and habits.

5.1.5 Minimising Human Exposure to PAHs from Food

Recommendations have been made by the JECFA with regards to minimising exposure to PAHs from dietary sources³⁹⁴. As much as possible, direct contact between the food and open flames should be avoided, particularly, allowing fat to drip into the flames should be avoided. Cooking methods, such as broiling, where the heat source is on top of the food should be used whenever possible, and when it is not, low and medium heat should be used, keeping the food as far away from the cooking/heat source as possible. With regards to food processing, direct smoking methods should be replaced with indirect methods. Finally, produce like fruits and vegetables should be washed and peeled prior to consumption. These and some other methods have been recently reviewed by Bansal and Kim³⁹⁰, but of particular interest is their recommendation to ingest antioxidants which, due to their acidic nature, engulf carcinogenic compounds such as PAHs and the free radicals their metabolism produces. Increased consumption of micronutrients, particularly retinol, α -tocopherol and γ -tocopherol have been shown to be associated with lower DNA adduct levels⁴¹¹. The same has been reported for increased intake of fresh fruit and vegetables, olive oil, antioxidants and vitamins, with the observations being attenuated in smokers^{412,413}.

5.1.6 Aims

The overarching aim of this chapter was to find any associations between DNA methylation and dietary PAH8 exposure which were expected to be different from those associated with air PAH8 exposure due to the different exposure routes. The diet is the major route of PAH exposure in humans and no study to date has been carried out to investigate the effects of total dietary PAH exposure and DNA methylation. This was done by first calculating dietary PAH8 exposure using FFQ data and estimates of PAH exposure collated from published literature. Once the dietary PAH8

exposures were calculated, an EWAS was carried out, with differentially methylated probes used to build a methylation index of dietary PAH8 exposure.

5.2 Results

5.2.1 Presence of PAHs in Foods Dataset

The food items dataset contained information for 1352 food items from 77 studies, while the food types dataset had 251 entries from 10 studies. The number of entries in each dataset by food class is summarised in Table 5.3. The food items dataset had more records than the food types dataset for all food classes with the exception of cereals and cereal products, condiments and sauces, eggs and egg products, and potatoes and other tubers. There were no data in the food items dataset for the soups and bouillons food class. Meat and meat products, fish and shellfish, and fats were the most popular food items for which PAHs were measured. While there were also a large number of non-alcoholic beverages in the food items dataset, they were obtained from only a handful of studies which is more indicative of the size of the studies than the popularity of the food type for analysing PAH content. The high number of items in the miscellaneous group is also misleading because the vast majority of items contributing to this group are supplements from four studies. The entries in the food types dataset were more evenly distributed, probably because the aim of the studies from which the data were obtained was to look at the whole diet. However, the meat and meat products group still had the most entries.

Between the two datasets, there were data available from at least one study for each of the 54 PAHs listed in

Table 5.4. The PAH with the most data were B[a]P for which there were only 62 entries in the food items dataset which had missing data. Smoked paprika, olive oil, coconut oils, lapsang souchang tea, gravy from grilled pork chops, tea infusions, black tea, smoked sausage, smoked herring and beef burgers all had total PAH concentrations of more than 1000 µg/kg. Smoked meat from Latvia, tea infusions, lapsang souchong tea, and smoked paprika had more than 1000 µg/kg of PAH8. In the food types, meat and meat products had the highest sum of all PAHs (39.51 µg/kg), and non-alcoholic beverages, more specifically coffee and tea, had the highest PAH8 (7.81 µg/kg). All the reported values are from the un-imputed datasets, meaning that the true concentration may be higher.

Table 5.3 Table of the number of food items and number of type entries in the datasets.

<u>Food Class Name</u>	<u>N Food Items</u>	<u>N Food Types</u>
Alcoholic beverages	26	4
Cakes & Biscuits	28	16
Cereals & Cereal Products	10	16
Condiments & Sauces	2	6
Dairy & Dairy Products	98	24
Eggs & Egg Products	2	8
Fats	153	19
Fish & Shellfish	281	21
Fruits, Nuts & Seeds	14	20
Legumes	8	6
Meat & Meat Products	358	34
Miscellaneous	137	16
Non-alcoholic beverages	175	13
Potatoes & Other Tubers	5	10
Soups & Bouillons	NA	3
Sugar & Confectionary	6	3
Vegetables	49	26

Table 5.4 List of polycyclic aromatic hydrocarbons included in the datasets

1-methylchrysene	Anthanthrene	Cyclopenta[<i>cd</i>]pyrene
1-methylnaphthalene	Anthracene	Dibenz[<i>a,c</i>]anthracene
1-methylphenanthrene	Benz[<i>a</i>]anthracene	Dibenz[<i>a,h</i>]anthracene
1-methylpyrene	Benzo[<i>a</i>]fluoranthene	Dienzo[<i>a,e</i>]pyrene
1,2-dimethylphenanthrene	Benzo[<i>b</i>]fluoranthene	Dibenzo[<i>a,h</i>]pyrene
11H-benzo[<i>b</i>]fluorene	Benzo[<i>ghi</i>]fluoranthene	Dibenzo[<i>a,i</i>]pyrene
2-methylantracene	Benzo[<i>j</i>]fluoranthene	Dibenzo[<i>a,l</i>]pyrene
2-methylnaphthalene	Benzo[<i>k</i>]fluoranthene	Dibenzothiophene
2-methylphenanthrene	Benzo[<i>a</i>]fluorene	Fluoranthene
2,4-dimethylphenanthrene	Benzo[<i>b</i>]fluorene	Fluorene
4,5-methylphenanthrene	Benzo[<i>c</i>]fluorene	Indeno[<i>1,2,3-cd</i>]perylene
5-methylchrysene	Benzo[<i>b</i>]naphthol[<i>2,1-d</i>]thiophene	Indeno[<i>1,2,3-cd</i>]pyrene
9-methylantracene	Benzo[<i>ghi</i>]perylene	Naphthacene
9-methylphenanthrene	Benzo[<i>c</i>]phenanthrene	Naphthalene
9,10-dimethylantracene	Benzo[<i>a</i>]pyrene	Perylene
Acenaphthelene	Benzo[<i>e</i>]pyrene	Phenanthrene
Acenaphthene	Chrysene	Pyrene
Acenaphthylene	Coronene	Triphenylene

The food items studies came from a total of 23 European countries, with Spain, Italy and Germany providing the most data with respect to number of items from a single country (Figure 5.1A). This was also the case for number of studies from a single country (Figure 5.1C). The 10 studies that made up the food types dataset came from 5 European countries and from a study which analysed foods from all over the EU. The EU study had the highest number of entries from a single location followed by the UK (Figure 5.1B). The countries contributed a single study each, with Spain and the UK being the exceptions having three studies each (Figure 5.1D).

Some of the food classes in the two datasets were made up with data from a single country. In the food items dataset, studies measuring PAH concentrations in Spanish foods accounted for 100% of

the data in the alcoholic beverages, cereals and cereal products, eggs and egg products, legumes, and potatoes and other tubers food classes (Figure 5.1E). Italian foods were the sole source of data in the condiments and sauces class, and Polish foods accounted for all the data in the sugar and confectionary class. Within the food types dataset, most countries contributed to most food classes (Figure 5.1F). The only exception was within the alcoholic beverages class where the data came only from the EU-wide study.

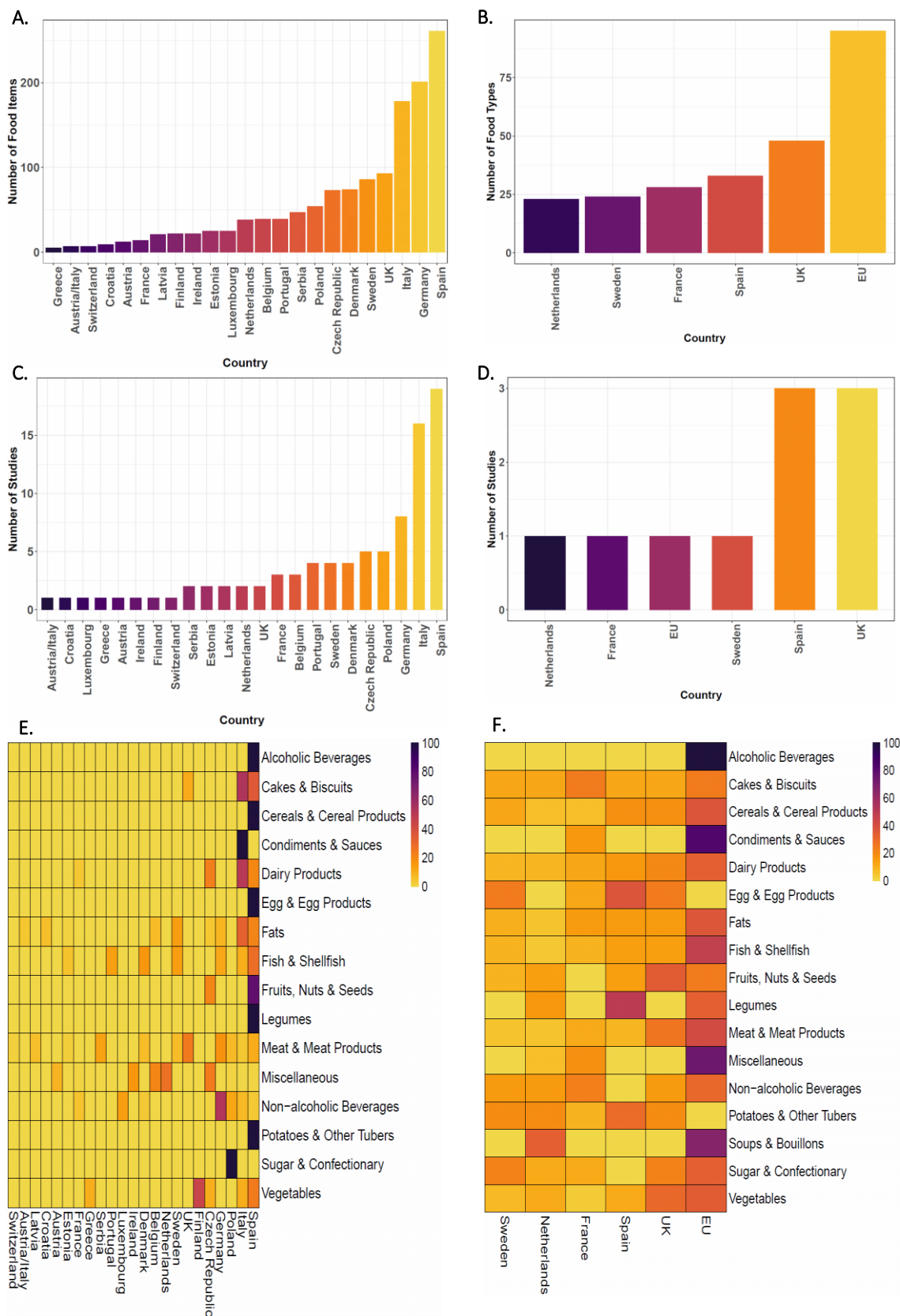


Figure 5.1 Data by country. A&B: Number of food items (A) and food types (B) by country. C&D: Number of studies within the food items (C) and food types (D) datasets by country. E&F: Heatmaps representing the proportion of data in each food class coming from each country for the food items (E) and food types (F) datasets.

5.2.2 Estimation of Dietary PAH8 Exposure

The total levels of the eight most carcinogenic PAHs (PAH8) were calculated following imputation of missing individual values. When comparing the total PAH8 levels across the two datasets within each food class, large differences were observed in the distributions between the food items and food types (Figure 5.2A). Wilcoxon tests showed significant differences in PAH8 distributions between the two datasets for: non-alcoholic beverages ($p = 2.8 \times 10^{-8}$), meat and meat products ($p = 1.2 \times 10^{-5}$), fats ($p = 4.6 \times 10^{-5}$), miscellaneous ($p = 1.1 \times 10^{-4}$), cakes and biscuits ($p = 1.6 \times 10^{-4}$), and alcoholic beverages ($p = 8.7 \times 10^{-3}$) food classes. In all of these cases, the food items' PAH8 levels were higher than those for the food types.

When comparing the medians for each food class from the two datasets, the same results were observed (Figure 5.2B). For many food classes, the median PAH8 concentration for the food items was higher than that for the food types. As mentioned earlier, one possible explanation for this is the heterogeneity in the methods used to measure PAH content in food by the studies in the food items dataset. Another explanation may be that studies looking to optimise measurement methodologies may use foods known to have high PAH content, therefore skewing the results. Given these observations, only the medians for the food groups were used to estimate dietary PAH8 intake in the cohort subjects.

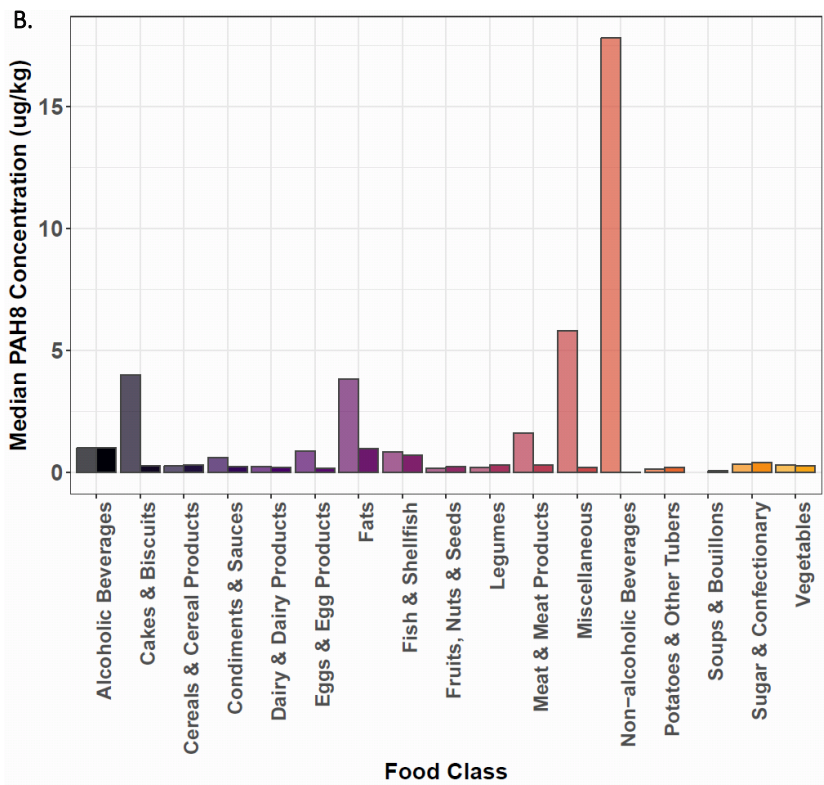
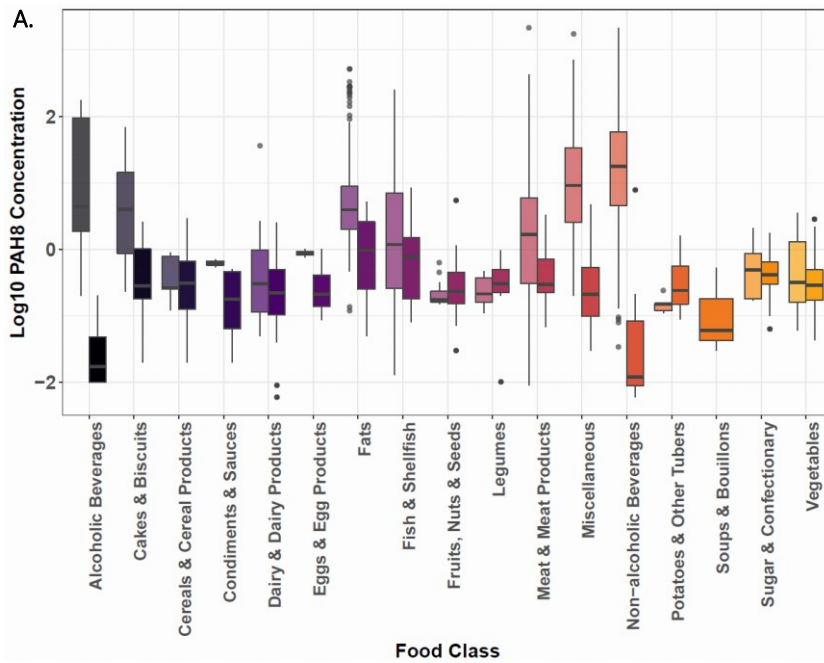


Figure 5.2 A: Boxplot showing PAH8 distributions from the food items (lighter colour on the left) and food types (darker colour on the right) datasets grouped by food class. B: Bar chart showing the median PAH8 concentration from the food items (lighter colour on the left) and food types (darker colour on the right) datasets grouped by food class.

The median PAH8 concentrations for each food class were multiplied by the intake of each subject for that class. The results were then summed in order to have a single value of total PAH8 dietary exposure per subject. The subjects for which dietary PAH8 exposure was estimated were the same as those used in the air pollution PAH8 analysis in the previous chapter. They were also divided into the same subsets for Training, Testing and EPIC-NL datasets. Table 5.5 shows the dietary PAH8 exposures of the three datasets, and the quartile distribution of dietary PAH8 exposure. The Training dataset had the highest mean dietary intake (520.7 ng/day), but the median was similar to that of the test dataset. The test dataset also had the widest range. The EPIC-NL dataset had the narrowest exposure range, as well as a lower mean and median. However, none of these differences were statistically significant and are shown in Figure 5.3. Additionally, as shown in Figure 5.4, there was no correlation between air and dietary PAH8 exposures in any of the datasets.

Table 5.5 Table of cohort characteristics for the Training, Testing and EPIC-NL EPIC subjects.

		EPIC-Italy Training (N=493)	EPIC-Italy Testing (N=208)	EPIC-NL (N=132)
DietaryPAH8 Intake (ng/day)	Range	181.0-1307.5	138.4-1324.9	168.6-790.9
	Mean	520.7	509.4	491.2
	Median	507.4	508.8	479.6
Dietary PAH8 Intake Quartiles (ng/day)	Q1	181.0-412.1	138.4-410.6	168.6-406.8
	Q2	412.1-507.4	410.6-508.8	406.8-479.6
	Q3	507.4-616.5	508.8-576.5	479.6-567.0
	Q4	616.5-1307.5	576.5-1324.9	567.0-790.6

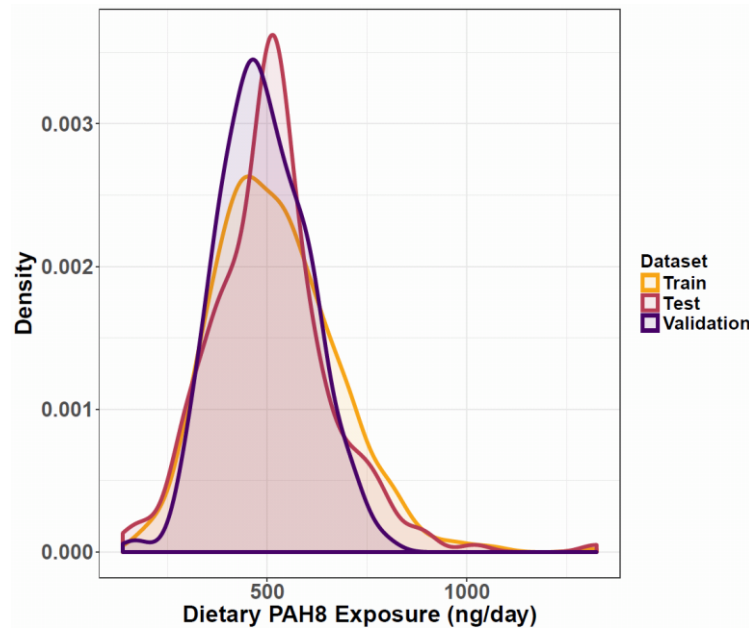


Figure 5.3 Distribution of dietary PAH8 intake in ng/day for each of the three datasets: Training (yellow), Testing (red) and EPIC-NL (purple).

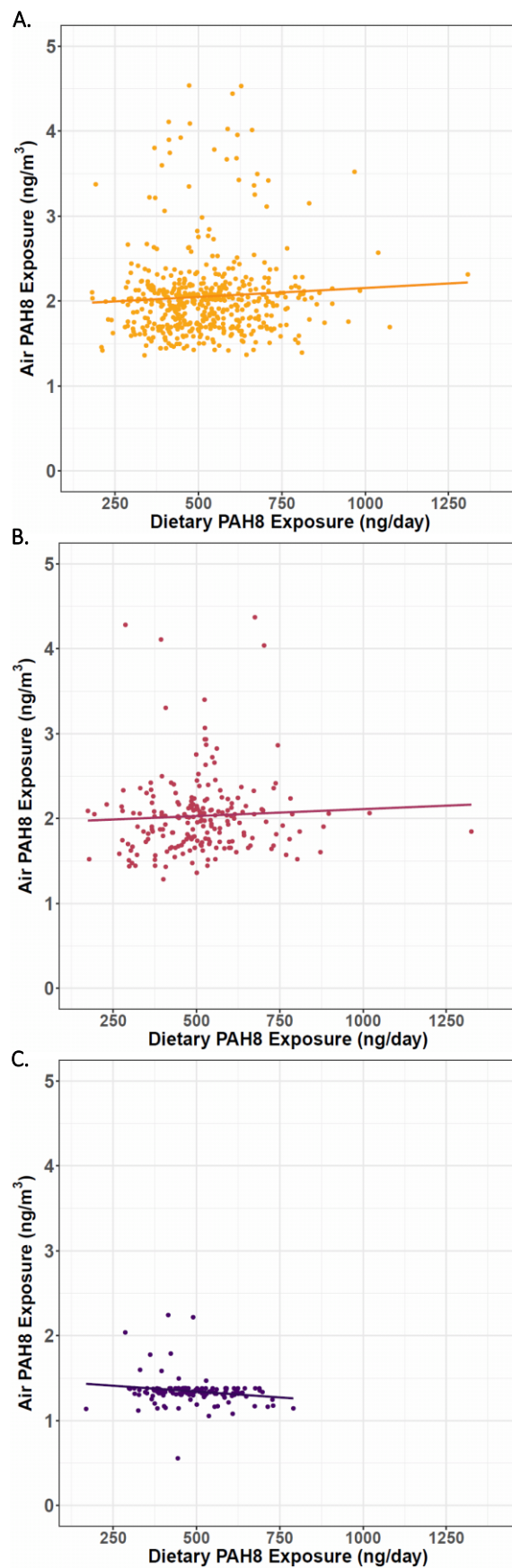


Figure 5.4 Correlation between Air (ng/m³) and Dietary (ng/day) PAH8 exposures in the Training (A), Testing (B), and EPIC-NL (C) datasets.

5.2.3 Effects of Dietary PAH8 Exposure on Global Methylation

When comparing the average methylation of all CpG sites between subjects in the lowest dietary PAH8 intake quartile (Q1) and those in the higher quartiles (Q2, Q3, and Q4)(quartiles summarised in Table 5.5), no statistically significant differences were observed in any of the datasets (Table 5.6). This was also the case for the trend in the three datasets. In the Training dataset, although not significant, the higher exposure quartiles indicated hypomethylation compared to Q1, while in the Testing and EPIC-NL datasets the beta coefficients indicated hypermethylation. In all cases, no dose-dependent methylation effects were observed.

5.2.4 Effects of Dietary PAH8 Exposure on Methylation at Genomic Regions

Table 5.7 summarises the results from the dietary PAH8 quartile analysis in CpG islands, shores and shelves. In the Training and EPIC-NL datasets none of the dietary PAH8 exposure quartiles were significantly associated with the methylation at CpG islands, shores, or shelves. In the Testing dataset, subjects in Q2 had significantly hypermethylated CpG islands and shores, and significantly hypomethylated shelves compared to subjects in Q1. Subjects in Q3 also had significantly hypermethylated CpG islands compared to Q1 subjects, however, the difference in methylation was greater between Q1 and Q2 subjects. As with the global methylation quartile analysis, the results for the Training dataset indicate that the subjects in the higher quartile had lower methylation at CpG islands, shores, and shelves while the results from the EPIC-NL dataset indicated the opposite. In the Testing dataset, the CpG islands and shores were found to be hypermethylated overall, and the shelves were hypomethylated. None of the results from the trend analyses were statistically significant.

Table 5.8 summarises the results from the dietary PH8 quartile analysis in gene promoter, 3' UTR, gene body and intergenic regions of the genome. Similar to results reported above, none of the dietary PAH8 exposure quartiles were associated with methylation differences at any of these genomic regions in the Training and EPIC-NL datasets. In the Testing dataset, the subjects in Q2 had

significantly hypermethylated promoters compared to subjects in Q1 and while the subjects in the higher quartiles also had hypermethylated promoters, this was not statistically significant. In the EPIC-NL dataset the results for all genomic regions indicated hypermethylation at the higher quartiles compared to subject in Q1, and the differences were consistent in the gene body. The results in the Training dataset were mixed in the 3' UTR, but consistently hypomethylated in the remaining genomic regions. None of the genomic regions showed any statistically significant trend.

Table 5.6 Table of beta regression results looking for differences in global methylation between quartiles of dietary PAH8 intake. The lowest quartile (Q1) was used as the reference quartile.

PAH Quartile	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-Netherlands - EPIC-NL (N=132)		
	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value
Q2	-0.002	-0.004; 0.0005	0.14	0.003	-0.001; 0.007	0.18	0.003	-0.005; 0.01	0.47
Q3	-0.0009	-0.003; 0.001	0.43	0.0002	-0.005; 0.005	0.94	0.002	-0.006; 0.01	0.64
Q4	-0.001	-0.004; 0.001	0.34	0.001	-0.003; 0.006	0.54	0.002	-0.006; 0.01	0.68
Trend	-8.86×10^{-7}	-7.4×10^{-6} ; 5.6×10^{-6}	0.79	4.54×10^{-6}	-6.6×10^{-6} ; 1.6×10^{-5}	0.43	2.46×10^{-6}	-2.3×10^{-5} ; 2.8×10^{-5}	0.85

Table 5.7. Table of beta regression results looking differences in methylation levels at CpG islands, shores and shelves between quartiles of dietary PAH8 intake. The lowest quartile (Q1) was used as the reference quartile. Entries in bold indicates statistically significant results (p < 0.05)

Genomic Region	PAH Quartile	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-Netherlands - EPIC-NL (N=132)		
		β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value
Shores	Q2	-0.001	-0.004; 0.002	0.45	0.006	0.0004; 0.01	0.04	0.006	-0.003; 0.01	0.20
	Q3	-0.0009	-0.004; 0.002	0.52	0.002	-0.003; 0.007	0.45	0.005	-0.003; 0.01	0.25
	Q4	-0.002	-0.004; 0.001	0.26	0.003	-0.003; 0.008	0.34	0.005	-0.003; 0.01	0.23
	Trend	-2.19×10^{-6}	$-.97 \times 10^{-6}$; 5.3×10^{-6}	0.57	4.44×10^{-6}	8.4×10^{-6} ; 1.73×10^{-5}	0.50	1.19×10^{-5}	-1.4×10^{-5} ; 3.8×10^{-5}	0.37
Shelves	Q2	-0.003	-0.006; 0.001	0.15	-0.01	-0.02; -0.0003	0.04	0.006	-0.01; 0.02	0.51
	Q3	-0.002	-0.004; 0.004	0.90	-0.007	-0.02; 0.003	0.16	0.007	-0.009; 0.02	0.37
	Q4	0.006	-0.003; 0.005	0.77	-0.002	-0.01; 0.008	0.73	0.005	-0.01; 0.02	0.58
	Trend	8.28×10^{-5}	-1.4×10^{-5} ; 1.5×10^{-5}	0.91	1.01×10^{-6}	-2.3×10^{-5} ; 2.5×10^{-5}	0.94	2.06×10^{-5}	-3.0×10^{-5} ; 7.1×10^{-5}	0.42
CpG Islands	Q2	-0.002	-0.006; 0.003	0.48	0.02	0.01; 0.03	4.1×10^{-5}	0.004	-0.008; 0.02	0.50
	Q3	-0.002	-0.007; 0.002	0.30	0.01	0.0006; 0.02	0.04	0.0007	-0.01; 0.01	0.91
	Q4	-0.005	-0.009; 0.0004	0.07	0.009	-0.002; 0.02	0.12	0.0008	-0.01; 0.01	0.90
	Trend	-5.5×10^{-6}	-2.1×10^{-5} ; 1.1×10^{-5}	0.50	1.92×10^{-5}	-8.7×10^{-6} ; 4.7×10^{-5}	0.18	-1.34×10^{-5}	-5.2×10^{-5} ; 2.5×10^{-5}	0.50

Table 5.8. Table of beta regression results looking differences in methylation levels at promoter, 3'UTR, gene body, and intergenic regions between quartiles of dietary PAH8 intake. The lowest quartile (Q1) was used as the reference quartile. Entries in bold indicate statistically significant results (p < 0.05)

Genomic Region	PAH Quartile	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-Netherlands - EPIC-NL (N=132)		
		β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value
Promoters	Q2	-0.001	-0.004; 0.002	0.37	0.001	0.003; 0.02	0.005	0.002	-0.006; 0.01	0.59
	Q3	-0.0007	-0.004; 0.002	0.66	0.004	-0.003; 0.01	0.30	0.00009	-0.008; 0.008	0.98
	Q4	-0.002	-0.005; 0.001	0.27	0.003	-0.004; 0.01	0.45	0.0006	-0.009; 0.008	0.89
	Trend	-1.15×10^{-6}	$-1.1 \times 10^{-5}; 8.5 \times 10^{-6}$	0.82	6.06×10^{-6}	$-1.1 \times 10^{-5}; 2.3 \times 10^{-5}$	0.48	-8.50×10^{-6}	$-3.4 \times 10^{-5}; 1.8 \times 10^{-5}$	0.52
3' UTR	Q2	-0.002	-0.005; 0.002	0.31	-0.006	-0.01; 0.003	0.19	0.005	-0.001; 0.02	0.48
	Q3	0.0002	-0.003; 0.004	0.89	-0.005	-0.01; 0.003	0.23	0.008	-0.006; 0.02	0.26
	Q4	0.0005	-0.003; 0.004	0.76	-0.0004	-0.009; 0.009	0.93	0.005	-0.009; 0.02	0.49
	Trend	3.08×10^{-6}	$-1.0 \times 10^{-5}; 1.6 \times 10^{-5}$	0.65	-1.46×10^{-6}	$-2.3 \times 10^{-5}; 2.0 \times 10^{-5}$	0.89	1.94×10^{-5}	$-2.5 \times 10^{-5}; 6.4 \times 10^{-5}$	0.39
Gene Body	Q2	-0.002	-0.004; 0.0003	0.08	-0.0008	-0.006; 0.004	0.77	0.004	-0.006; 0.01	0.40
	Q3	-0.0006	-0.003; 0.002	0.58	-0.002	-0.007; 0.004	0.52	0.004	-0.006; 0.01	0.45
	Q4	-0.0008	-0.003; 0.002	0.51	0.001	-0.004; 0.006	0.72	0.004	-0.06; 0.01	0.44
	Trend	-6.07×10^{-7}	$-8.2 \times 10^{-6}; 7.0 \times 10^{-6}$	0.88	3.81×10^{-6}	$-9.1 \times 10^{-6}; 1.7 \times 10^{-5}$	0.56	1.12×10^{-5}	$-1.9 \times 10^{-5}; 4.2 \times 10^{-5}$	0.47
Intergenic	Q2	-0.002	-0.005; 0.001	0.19	0.002	-0.004; 0.008	0.51	0.002	-0.006; 0.002	0.37
	Q3	-0.002	-0.005; 0.001	0.18	-0.0005	-0.007; 0.006	0.88	0.005	-0.006; 0.002	0.35
	Q4	-0.001	-0.005; 0.002	0.43	0.002	-0.004; 0.009	0.45	0.005	-0.006; 0.002	0.38
	Trend	-1.36×10^{-6}	$-1.1 \times 10^{-5}; 7.8 \times 10^{-6}$	0.77	8.44×10^{-6}	$-6.3 \times 10^{-6}; 2.31 \times 10^{-5}$	0.26	1.32×10^{-5}	$-2.1 \times 10^{-5}; 4.7 \times 10^{-5}$	0.45

5.2.5 EWAS Results

The EWAS was carried out on the 493 EPIC-Italy subjects in the Training dataset. After the removal of probes whose model did not converge, as well as probes known to cross-hybridise, probes on SNPs, and probes on the sex chromosomes, results were available for 362,310 CpG sites. Of these, 171 were found to be significantly associated with dietary PAH8 exposure (FDR $q < 0.05$; $p < 2.37 \times 10^{-5}$), with 85 sites hypomethylated and 86 sites hypermethylated (Figure 5.5A and Figure 5.5B). Of these, 7 passed the more stringent Bonferroni correction ($p < 1.38 \times 10^{-7}$), 5 of which were hypermethylated and the remaining 2 hypomethylated (Figure 5.5A). The inflation factor (λ) was found to be 1.19 (Figure 5.5C). All of significant probes had a methylation difference of $< 1\%$ per unit increase in dietary PAH8 intake. This biggest differences observed were 0.0074% hypomethylation (cg21811450; subject with lowest dietary PAH8 exposure = 61.9% methylated; subject with highest dietary PAH8 exposure = 37.9% methylated) and 0.0079% hypermethylation (cg10832076; subject with lowest dietary PAH8 exposure = 41.4% methylated; subject with highest dietary PAH8 exposure = 46.3% methylated) per unit change.

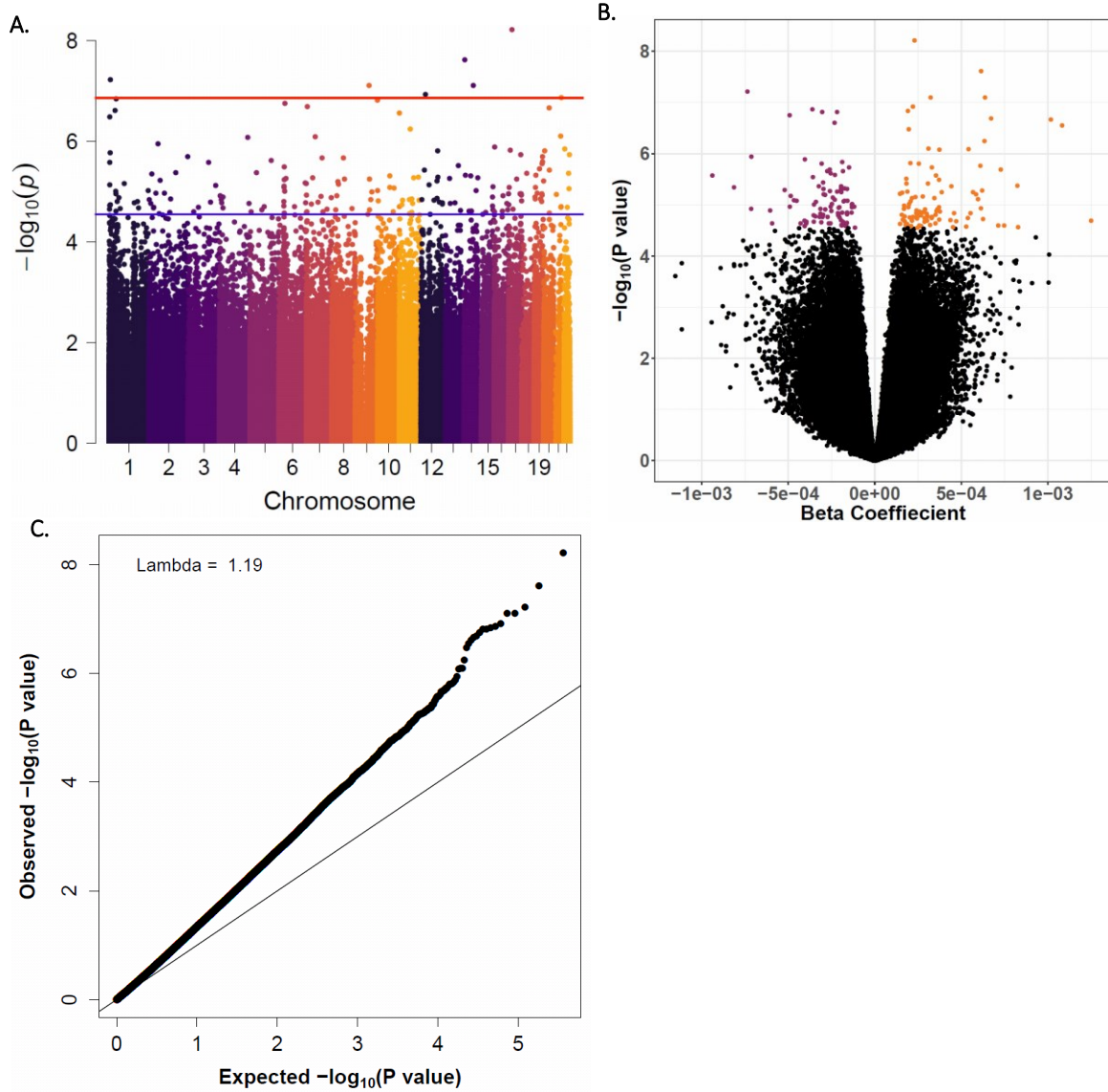


Figure 5.5 A: Manhattan plot showing the $-\log_{10}$ transformed p values of the 362,310 CpG probes tested arranged by chromosome. The red line indicates the threshold for Bonferroni correction for multiple testing ($p < 1.38 \times 10^{-7}$), and the blue line indicates the FDR threshold ($FDR \ q < 0.05$; $p < 2.37 \times 10^{-5}$). B: Volcano plot showing the $-\log_{10}$ transformed p values of the 362,310 CpG probes against the β -coefficient for dietary PAH8 intake. Coloured points indicate significance after FDR correction, with red points indicating a decrease in methylation and orange points indicating an increase in methylation. C: QQ plot showing the observed $-\log_{10}$ transformed p values against the expected $-\log_{10}$ transformed p values from the EWAS.

The results of the 7 Bonferroni significant CpG probes in all three datasets are summarised in Table 5.9, with the results for the 171 FDR-significant probes shown in Appendix 3 Table 9.11. None of the probes were statistically significant ($p < 0.05$) in the EPIC-NL dataset, however 4 of the 7 probes showed a difference in methylation in the same direction as reported in the Training dataset (cg24634746 was hypomethylated; cg05419385, cg10001646, and cg12550399 were hypermethylated). Two probes were statistically significant in the Testing dataset: cg24634746 ($p = 0.013$) and cg10001646 ($p = 0.04$). The latter was hypermethylated in both datasets and non-significantly in the EPIC-NL dataset, however the former was hypomethylated in the Training dataset and hypermethylated in Testing dataset. Other probes showed the same direction of change in the Testing dataset as the Training dataset but were not statistically significant: cg05419385 and cg12940991 were hypermethylated, and cg13588826 was hypomethylated. The methylation status of the probes with the biggest methylation changes (cg21811450 and cg10832076) were assessed in the subjects with the lowest and highest dietary PAH8 exposures in the Testing and EPIC-NL datasets. In the Testing dataset, probe cg21811450 in the subject with the lowest exposure was 37.9% methylated, and the most highly exposed subject was 73.3% methylated. These observations are very similar to those observed in the Training dataset outlined in the previous paragraph, and the model statistics for this probe were similar in the Training and Testing datasets (Appendix 3 Table 9.11). In the EPIC-NL dataset, probe cg21811450 in the subject with the lowest exposure was 65.1% methylated, and the most highly exposed subject was 65.2% methylated. Probe cg10832076 was 41.8% methylated in the subject with the lowest air PAH8 exposure and 40.9% methylated in the subject with the highest air PAH8 exposure in the Testing dataset. This probe in the EPIC-NL dataset was 44.4% methylated and 48.3% methylated in the subjects with the lowest and highest dietary PAH8 exposures respectively.

Table 5.9 Model results for the Bonferroni significant ($p < 1.38 \times 10^{-7}$) EWAS probes in the three datasets: Training, Testing and EPIC-NL. All results are from beta regression models assessing the relationship between dietary PAH8 exposure and the methylation beta values for each probe. The Training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The Testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status.

Probe ID	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-NL - EPIC-NL (N=132)		
	<u>B</u> <u>Coefficien</u> <u>t</u>	<u>95%</u> <u>Confidence</u> <u>Interval</u>	<u>P Value</u>	<u>B</u> <u>Coefficient</u>	<u>95%</u> <u>Confidence</u> <u>Interval</u>	<u>P Value</u>	<u>B</u> <u>Coefficient</u>	<u>95%</u> <u>Confidence</u> <u>Interval</u>	<u>P Value</u>
cg24634746	-0.00074	-0.001; - 0.00047	6.03E-08	0.00030	0.00006; 0.00053	0.013	-0.0002	-0.00087; 0.00048	0.569
cg21207730	0.00064	0.0004; 0.00087	7.89E-08	-0.00010	-0.00047; 0.00026	0.576	-0.00025	-0.00088; 0.00038	0.438
cg05419385	0.00022	0.00014; 0.0003	1.20E-07	0.00005	-0.00009; 0.0002	0.490	0.00018	-0.00021; 0.00056	0.364
cg10001646	0.00061	0.0004; 0.00083	2.45E-08	0.00018	0.00001; 0.00036	0.040	0.00004	-0.00041; 0.00049	0.850
cg12940991	0.00032	0.00021; 0.00044	7.91E-08	0.00015	-0.00004; 0.00034	0.118	-0.00033	-0.00094; 0.00029	0.295
cg12550399	0.00023	0.00015; 0.00031	6.10E-09	-0.00003	-0.00016; 0.00011	0.707	0.00034	-0.00045; 0.00071	0.079
cg13588826	-0.00036	-0.0005; - 0.00023	1.36E-07	-0.00001	-0.0002; 0.00017	0.892	0.00021	-0.00037; 0.0008	0.475

The characteristics of the 7 Bonferroni significant probes are summarised in Table 5.10 and the characteristics of the 171 FDR-significant probes can be found in Appendix 3 Table 9.12. Four of the probes were located in genic regions, specifically the gene body and 3' UTR, while the remaining probes were intergenic. A more detailed analysis of the genomic distribution of the FDR significant probes showed that most changes occurred at the promoter, intron and intergenic regions (Table 5.11, Figure 5.6A). Comparing the genomic distribution of these probes to that of all sites on the array, significantly less changes than expected occurred at promoter regions (OR = 0.67 ; p = 0.038), and significantly more changes occurred at 3' UTR regions (OR = 2.38; p = 0.0093). When comparing the direction of change at all genomic regions to the overall ratio of hypomethylated to hypermethylated probes (ratio = 0.99) (Table 5.12, Figure 5.6B), the changes occurring at CpG islands were significantly more hypomethylated than expected (OR = 5.87; p = 0.0026). Intergenic and 3' UTR regions had borderline significantly more hypermethylation events than expected (Intergenic: OR = 0.39, p = 0.054; 3' UTR: OR = 0.21, p = 0.057). Multiple CpG probes within and around a CpG island located in an open sea region on chromosome 15 were found to be significantly hypomethylated (Figure 5.7). Three of these sites were FDR significant, with a further two being borderline FDR significant, and only one site had p > 0.05.

Table 5.10 Table of characteristics of probes found to be significantly associated with dietary PAH8 exposure at the Bonferroni level ($p < 1.38 \times 10^{-7}$) in the Training dataset.

Probe ID	Chromosome	Position	UCSC RefGene Name	Gene Location	Relation to CpG Island	Methylation Change Direction
cg24634746	1	7538723	<i>CAMTA1</i>	Body		-
cg21207730	9	86821905				+
cg05419385	12	27352945				+
cg10001646	14	24683737	<i>MDP1</i>	Body	South Shore	+
cg12940991	14	77525744				+
cg12550399	17	19482275	<i>SLC47A1</i>	3'UTR	North Shore	+
cg13588826	21	47533197	<i>COL6A2</i>	Body	South Shore	-

Table 5.11. Table of Fisher’s test results comparing the number of differentially methylated probes (N = 171) and all tested probes (N = 362,310) in the Training dataset EWAS at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred.

Genomic Region	Odds Ratio	Confidence Interval	P Value
3' UTR	2.38	1.17-4.38	0.0093
5' UTR	0.80	0.16-2.39	1
Exon	1.19	0.66-1.99	0.48
Intergenic	1.21	0.76-1.84	0.36
Intron	1.10	0.75-1.58	0.64
Non-coding	1.04	0.21-3.10	0.77
Promoter	0.67	0.46-0.98	0.035
TTS	1.53	0.55-3.40	0.30
CpG Island	1.10	0.64-1.80	0.70
LINE	0.55	0.11-1.64	0.38
SINE	0.96	0.31-2.30	1
LTR	0.49	0.058-1.78	0.45
Other	0.73	0.088-2.68	1

Table 5.12. Table of Fisher’s test results comparing the number of hypermethylation changes (N=86) to hypomethylation changes (N=85) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected.

Genomic Region	Odds Ratio	Confidence Interval	P Value
3' UTR	0.21	0.021-1.05	0.057
5' UTR	2.04	0.10-122.15	0.62
Exon	1.33	0.42-4.44	0.61
Intergenic	0.39	0.14-1.03	0.054
Intron	0.56	0.25-1.22	0.14
Non-coding	0.50	0.0084-9.81	1
Promoter	1.90	0.85-4.36	0.097
TTS	2.07	0.29-23.42	0.44
CpG Island	5.87	1.57-32.93	0.0026
LINE	0.50	0.0084-9.81	1
SINE	0.67	0.055-5.99	1
LTR	1.01	0.013-80.31	1
Other	Inf	0.19-Inf	0.25

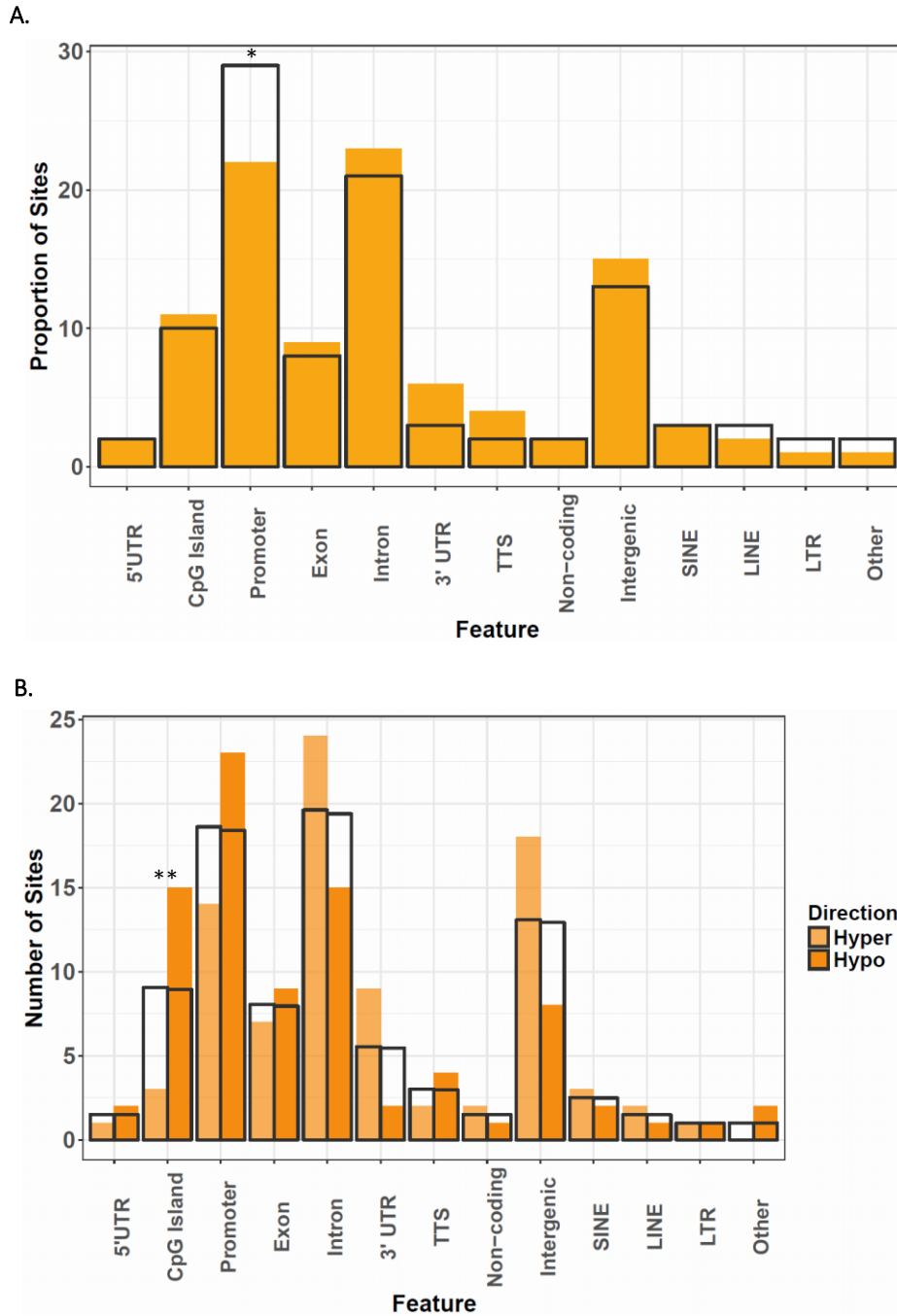


Figure 5.6 A: Comparison of the genomic distribution of differentially methylated probes (N = 171) and all tested probes (N = 362,310) in the Training dataset EWAS. The filled yellow bars show the proportion of significant probes, the grey outline bars show the proportion of all probes tested, i.e. the expected distribution. B: Comparison of the genomic distribution of hypermethylated (N = 86) and hypomethylated (N = 85) probes. As in A, the filled yellow bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. For both plots, * indicates $p < 0.05$ and ** indicates $p < 0.01$ following Fisher's Exact test.

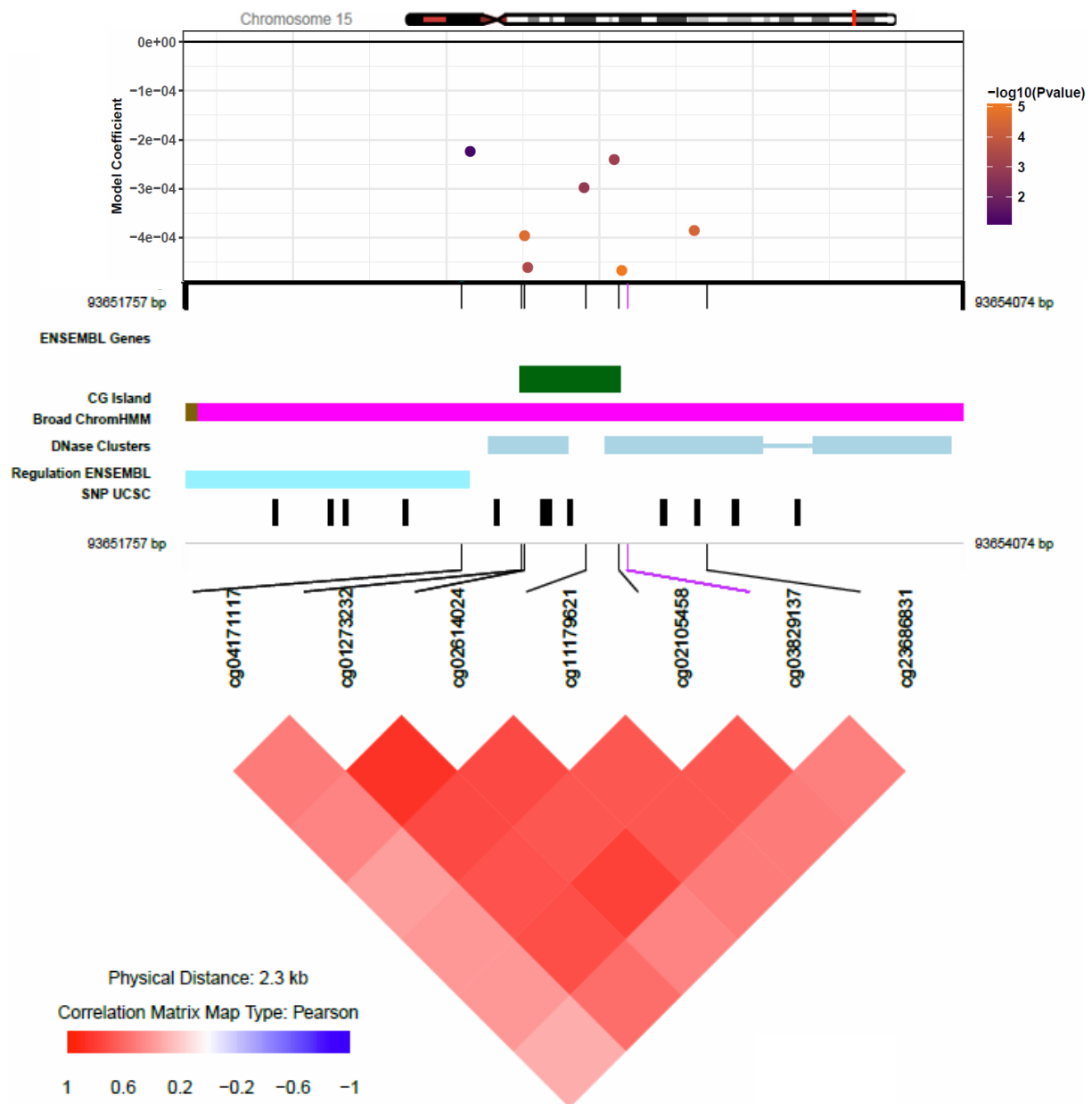


Figure 5.7. Modified coMET plot of a 2kb region on chromosome 15 where multiple CpG sites were found to be hypomethylated. The top panel of the figure is a regional association plot showing the beta coefficients of these probes from the EWAS by genomic position. The colour of the points corresponds to the \log_{10} P value, where orange indicates FDR-significant probes (FDR $q < 0.05$; $p < 2.37 \times 10^{-5}$). The central panel shows the genomic landscape of the region with respect to genes, CpG islands, chromatin state, clusters of DNase, SNPs and regulatory features. The bottom panel shows a correlation matrix of the methylation values for each probe where red indicates a strong positive correlation, blue a strong negative correlation, and white a lack of correlation between the probes.

5.2.6 Building a Methylation Index of Dietary PAH8 Exposure

The full model was trained on the Training dataset and included all 171 FDR significant probes identified in the Training dataset EWAS as well as age, cancer case status, sex and smoking status variables. The full model had an α parameter equal to 0.22 which indicated that elastic net regression produced the best performing model. This meant that for some variables the coefficients were shrunk to 0 and therefore were not included in the final model. This was the case for 74 CpG probes as well as cancer case status and smoking status therefore after Training, the final model included 97 of the 171 CpG probes, age, and sex. For these remaining covariates, the penalty factor applied (λ) was 12.92 which was calculated during the model Training process. The model with the best performance had an RMSE of 120.24 ng/day and the percentage of variance explained by the model (R^2) was 37.9%. Figure 5.8A and Figure 5.8B show how the model performed when reapplied to the Training dataset on which it was trained and model performance on the Testing dataset respectively. The probes included in the methylation index were identified only in the Training dataset and the model was independently tested in the Testing dataset. As expected, the model predicted the dietary PAH8 exposure of the Training dataset subjects reasonably well as determined from the correlation between the predicted and real dietary exposures (Spearman's Rho = 0.78, $p < 2.2 \times 10^{-16}$). The results of the same correlation using the predicted and real exposures of subjects in Testing dataset were less strong but still statistically significant (Spearman's Rho = 0.19, $p = 0.007$).

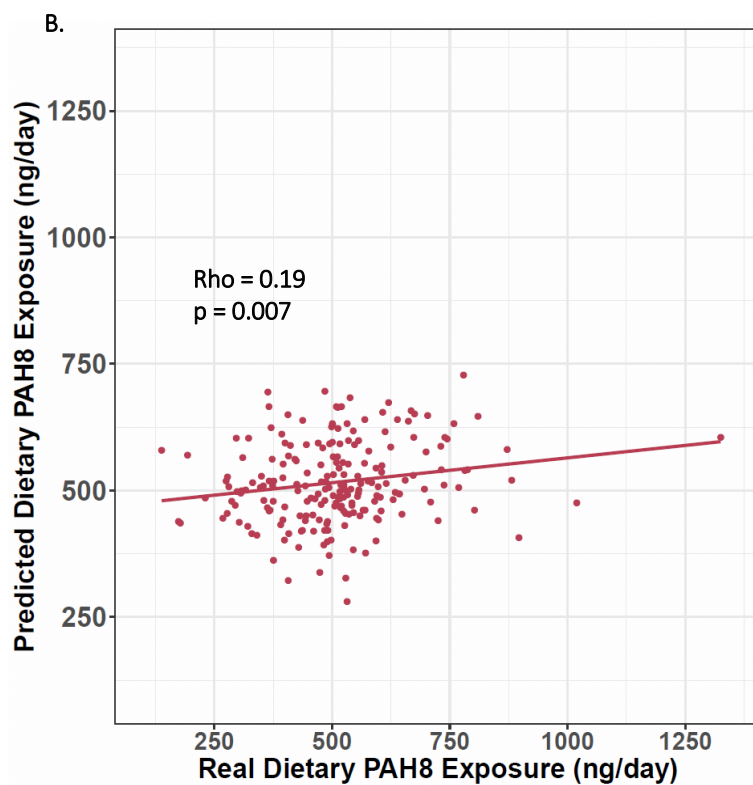
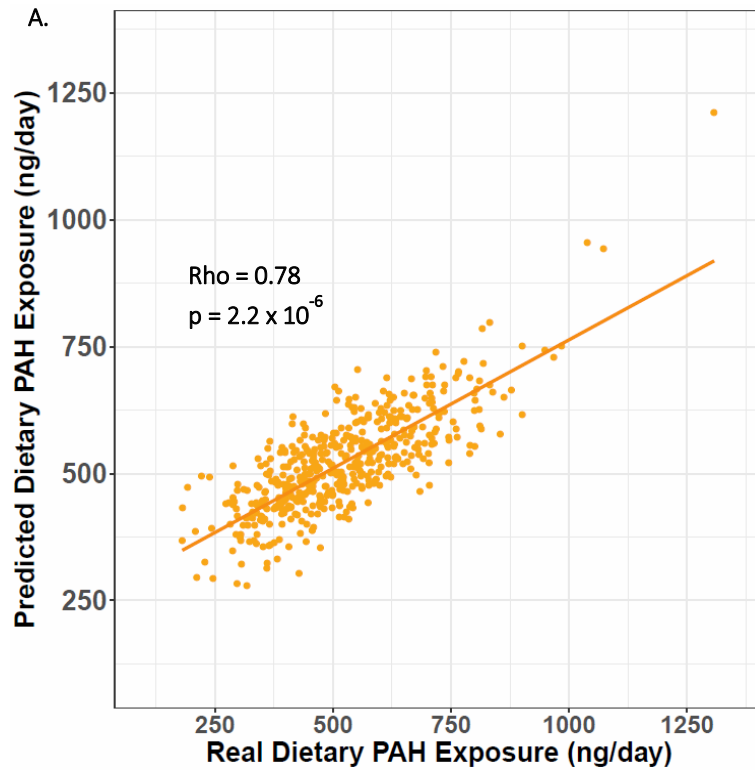


Figure 5.8 Plots showing the correlation between the dietary PAH8 exposure predicted by the elastic net model against the real combined PAH8 exposure for each subject. Figures A and B show the results from the full model for the Training and Testing sets respectively.

5.3 Discussion

This is the first study analysing the association between genome-wide methylation and dietary PAH8 exposure in humans. A previous study analysing human cord blood methylation and prenatal PAH exposure has assessed the relationship between the consumption of PAH-rich foods (specifically smoked, broiled, fried, and barbecued meat) and the number of BPDE-DNA adducts³⁴⁵. No relationship was found by the authors, but the results reported in this thesis indicate that dietary PAH8 exposure estimated from all ingested foods reported by subjects in a FFQ is associated with differences in methylation. The results also showed that of the methylation changes associated with dietary PAH8 exposure, more than expected occurred at the 3'UTR regions, and less than expected occurred at promoter regions. At CpG islands more hypomethylation changes were observed compared to the overall ratio of hypermethylation to hypomethylation changes. Several CpG probes in and around a CpG island on chromosome 15 were found to be significantly or borderline significantly hypomethylated. None of the probes identified in the Training dataset validated in the Testing or EPIC-NL datasets individually, and the methylation index indicated an association rather than prediction in the Testing dataset.

5.3.1 Exposure Estimates of Dietary PAH8 Exposure

Calculating the exposure estimates for these subjects was possible by first integrating data from 87 European studies where PAH measurements of various foods were published. This is currently the largest known dataset of PAH levels in foods and includes over 250 entries in the food types subset of the dataset, and 1300 entries in the food items subset. Use of this dataset with the available FFQs of the subjects made it possible to calculate dietary PAH8 exposure. The lack of correlation between air and dietary PAH8 exposures in the three datasets is not surprising as the two sources of exposure are independent of each other.

5.3.2 Factors Affecting Estimation of Dietary PAH8 Intake

All of the factors discussed below outline possible sources of misclassification and bias introduced to the study due to the nature of FFQs.

5.3.2.1 Seasonal Variability

A previous study has shown high seasonal variability within an individual's diet and the authors suggested that this contributed to the variation within the study population⁴⁰⁴. The FFQ provides only a snapshot of dietary habits and it is also possible that since subjects were recruited to the cohort over a period of years, they may have been recruited and had their blood drawn at different times of year which could have influenced their FFQ responses. Using FFQs as a method for determining diet is prone to misclassification errors and bias for additional reasons. The FFQ cannot account for changes in dietary habits over time, meaning that while a subject may report a reasonably healthy low-PAH diet at the time the FFQ was filled in, this may not always have been the case. Also, the accuracy of estimates made using FFQs as a measure of intake is highly dependent on the honesty of the subject and recall bias. It is known for subjects to underreport what they consider to be "bad habits" such as the quantity and quality of their diet, alcohol intake, and smoking habits and this supports the need to develop biological measures instead of relying on self-reports.

5.3.2.2 Cooking Methods and Food Production

There are other important factors that FFQs do not take into consideration which would influence the estimation of dietary PAH exposure. The first is cooking processes – the dietary PAH burden of a barbecued piece of meat would be significantly different from a roasted piece of meat. A large study looked at the effects of various cooking methods on the amount of PAHs in meats and other foods²⁴⁹. The authors also accounted for distance between food and cooking source and cooking time and found that all these variables impacted the concentration of PAHs in the food item²⁴⁹. These variables are not usually recorded in a FFQ and even if they were, the answers are unlikely to be accurate given that most people would not consciously be aware of all these factors when preparing their food.

Additionally, these variables would vary seasonally, with barbecues being more prevalent in the summer compared to the winter for example, so would also contribute to the seasonal variation considered above.

The methods of production and source of foods are also factors that need to be considered. In a study comparing conventional egg production to free-range and organic methods found that the former contained five times more PAHs than the latter two ⁴¹⁴. These factors need to be included in FFQs in order to obtain more accurate estimates of PAH exposure, however these are not always known. As mentioned above, smoking methods are a major method by which PAHs are introduced to food, with liquid smokes becoming increasingly common. Finally, as discussed above, the location where vegetables are grown and livestock are reared impacts the levels of PAH present in those foods however this would be almost impossible to determine for all foods an individual might eat.

5.3.2.3 Possible Solutions

Many of these limitations in dietary PAH8 exposure estimation could have been overcome by using methods such as the duplicate plate method to measure the PAH levels in the food subjects actually consumed. This would account for differences in cooking methods, production methods, and between brands of food. This method, however, is also subject to a number of limitations. It is possible that subjects would not prepare the same types of meals they would eat in reality for similar reasons as misreporting in the FFQ. Additionally, from a logistical and financial perspective, it would be difficult to include large numbers of subjects, which would have an impact on statistical power. A more accurate, cheaper, and more feasible method to measure PAH8 exposure would be to use biomarkers such as urinary metabolites which have been discussed in the previous chapter. One of the aims of this study was to attempt to determine differences in methylation associated with different sources of PAH exposure, and the use of biomarkers, even if available, would not allow for source attribution.

5.3.3 Comparison of Estimated Dietary PAH8 Intakes to Previous Reports

Comparison of the range of dietary PAH8 intakes estimated in this chapter to previously reported intakes indicate that the exposures calculated were underestimates. An analysis of the Italian diet sampled and published in 1995 found that the average intake of PAHs was 3000 ng/day, which is almost 6 times higher than the average dietary PAH8 intake calculated for the EPIC-Italy subjects (Training and Testing datasets) above. A similar Dutch study which sampled between 1984 and 1986¹⁹⁸ reported the average dietary PAH intake was 5220 ng/day, more than 10 times the average dietary PAH8 intake calculated for the EPIC-NI subjects (EPIC-NL dataset). In 2008, the CONTAM panel²⁸³ reported the average PAH intakes to be 1962 and 1785 ng/day for the Italian and Dutch populations respectively, which are closer to estimates of dietary PAH8 intake calculated above, but still significantly higher. One explanation for this is that the estimates were calculated using the presence of PAHs in food dataset built in this thesis which was built from studies published from 1980 to the present day. The cited studies suggest that the levels of PAHs in foods have been declining over the years, probably due to increased regulation and awareness, and declining air pollution, however the dataset contained studies conducted over a number of decades, but a higher number were from more recent years. This may have contributed to the underestimation, particularly since all subjects were recruited in the 90's. A second explanation is the use of only the food types data from the dataset, which provided overall lower median levels of PAHs for most of the food classes compared to those for the food items. The reason for this decision was that the food items data were influenced by food items with particularly high concentrations of PAHs (e.g. paprika, or tea) which would have skewed the estimates, while the food types data were more consistent across studies. The number of contributing studies also needs to be considered – 77 food items studies vs 10 food types studies – since heterogeneity increases with the number of studies. An additional reason for the discrepancy between the calculated estimates and those published in the literature is the number of PAHs considered. In this thesis, only the eight most carcinogenic PAHs were included, however, the PAHs measured across studies varied in number, ranging from one to more than sixteen, which would

greatly influence the reported values. Imputation of missing values for studies where not all eight PAHs were measured was carried out in the presence of PAHs in food dataset which could have been a contributing factor to the underestimation. The use of food classes in the work presented instead of more granular levels of data may have also influenced the estimated exposures since the food classes are made of a number of food types which spanned a range of PAH concentrations. If, for example, a particular subject prefers cooking with one type of oil which has a very different concentration of PAHs than another, the method employed here would not account for this difference.

5.3.4 Comparison of EWAS Results to Previously Published Findings

The recently published study by Tryndyak *et al.*³²⁵ carried out RRBS on a B[a]P-exposed human liver cell line (HepaRG cell line). The authors noted over 6500 differentially methylated regions in these cells compared to controls. A comparison of the EWAS results presented above, and the findings by Tryndyak *et al.*³²⁵ found two genes reported in both studies, however the location of the differentially methylated sites and the direction of change were not consistent between the two studies. The results are shown in Appendix 3 Table 9.13.

While none of the CpG sites associated with dietary PAH8 exposure were previously reported to be associated with smoking, 38 genes associated with CpG sites overlapped with both the dietary PAH8 exposure and published smoking EWAS results. These overlaps are summarised in Appendix 3 Table 9.14. Cg06420305 located in the WWOX gene was found to be hypomethylated in relation to dietary PAH8 exposure, and it has also been reported to be associated with prenatal PAH exposure albeit in the opposite direction¹⁴⁶. Hypermethylation of other probes in the WWOX gene have previously been reported to be associated with smoking^{415,416}. No other probes or genes previously reported in air pollution, occupational PAH, or prenatal PAH exposure studies overlapped with the results reported above. This may further support the hypothesis that the effects of PAH exposure on DNA methylation that are dependent on the route of exposure, however further work would be required to confirm this.

Of the significantly differentially methylated CpG sites, 70 of the 122 genes associated with the 171 FDR-significant probes have previously been reported in the CTD to have altered gene expression levels associated with PAH exposure¹⁸⁸. At least one of the eight PAHs in PAH8 has been reported in the CTD to be associated with a total of 12,723 unique gene interactions. The 70 overlapping gene interactions are summarised in the appendix in [Table 9.14](#) Table 14 and it is important to note that predominantly, the studies reported in the CTD were *in vitro* human cell-line experiments or *in vivo* animal model experiments. The human genome contains between 19,000 – 20,000 genes, meaning that the CTD has reports of PAHs interacting with over half the genome (approximately 63%), and the gene overlaps found here are about 57% (70 of 122 genes). This suggests that overlaps identified are possibly due to chance. Even so, this does support the theory that PAHs do not induce gene-specific signatures of DNA methylation.

Finally, comparison both of the CpG probes and the genes associated with those probes between the dietary PAH8 results and the air PAH8 results show no overlap. This may suggest that the two different exposure routes do in fact have different effects on DNA methylation, an observation that could not be made using biomarkers such as urinary metabolites. Alternatively, this observation may indicate that both sets of EWAS results are not really associated with PAH8 exposure, but rather are a random set of probes.

5.3.5 Statistical and Other Considerations

It is important to note however, that while these findings are interesting and warrant further investigation, the results reported for the Training dataset were not consistently replicated in either the Testing or EPIC-NL datasets. The cohort split between Training and Testing datasets was kept the same as in previous chapter, i.e. subjects were split based on their air PAH8 exposure as described in the methods chapter and the datasets were then maintained for all subsequent analyses and the two exposures (air and dietary PAH8 exposures) were not correlated in either of the two datasets. Despite this, no statistically significant differences between the dietary PAH8 exposures in the Training and

Testing datasets were found, and this also applied to the EPIC-NL dataset ($p > 0.05$). As described in the previous chapter on air PAH8 exposure, none of the cohort characteristics included in the models applied were significantly different between the Training and Testing datasets but this was not the case for the EPIC-NL dataset which did not include any male subjects, current smokers, or subjects diagnosed with cancer. Moderate inflation of the test statistics was observed as in the previous chapter. Several discussion points related to the EWAS results that were made in the previous chapter about air PAH8 exposure are also applicable here, such as using p-values over effect sizes to choose the CpG probes of interest, inflation of test statistics, the lack of validation across the three datasets, overfitting of the methylation index, and underlying differences in the methylation between the three datasets. These will not be discussed again here but will be summarised in the final chapter of this thesis.

5.3.6 Conclusions

In summary, dietary PAH8 exposure does have some, albeit small effect on DNA methylation in WBCs where 171 CpG probes were found to be differentially methylated. The CpG probes and genes affected are different to those previously reported in air pollution and PAH exposure publications, as well those reported in the previous chapter on air PAH8 exposure. Of the observed methylation changes, significantly less than expected occurred at gene promoter regions, and more hypomethylation events than expected occurred at CpG island regions. The methylation index developed performed well in the Training dataset despite the percentage of variance explained not being very high, but performance decreased greatly in the Testing dataset. The reasons for this warrant further investigation but some possible reasons are discussed in Chapter 7.

6 Chapter 6 – Combined Air and Dietary PAH8 Exposure

6.1 Introduction

In the two previous chapters, EWASs were carried out to analyse the relationship between blood DNA methylation and PAH8 exposure from air inhalation and dietary sources separately. There was no overlap between the 204 and 171 differentially methylated probes identified in the air PAH8 exposure EWAS and the dietary PAH8 exposure EWAS respectively. In this chapter, air and dietary PAH8 exposures are combined and the relationship of combined PAH8 exposure with DNA methylation is assessed.

6.1.1 The Effects of Smoking on DNA Methylation

Smoking is considered to be one of the main sources of PAH exposure and several epidemiological studies have been carried out to determine the effects of smoking on the methylation of various genes. PAH exposure from smoking has not been considered in this thesis, but the main findings from previous studies are discussed below. Additionally, the results from 31 smoking EWASs were collated and have been compared with the results from the EWASs carried out in the current and previous chapters to identify any overlaps. These 37 published smoking EWAS describe methylation changes at over 10,400 unique CpG sites mapping to almost 5000 genes^{49,380,423–432,415,433–442,416,443–449,417–422}. The genes most frequently reported to be differentially methylated are summarised in [Table 6.1](#) including the number of differentially methylated CpG probes that have been reported and the direction of methylation change. One study found that over 97% of the differentially methylated CpG sites were found to be hypomethylated in association with smoking exposure and that 149 out of 751 loci remain differentially methylated more than 35 years after smoking cessation⁴²⁵. The authors hypothesised that the reason for this is a higher smoking-related difference at these sites rather than it taking a longer time for their methylation states to revert back to normal indicating that the magnitude of difference in methylation changes must also be considered. The functions of the genes most commonly found to be differentially methylated due to smoking include the development and

function of the cellular, haematological, immune, cardiovascular, tumorigenic and reproductive systems.

6.1.2 Aims

The previous two chapters looked separately at the effects of air and dietary PAH8 exposure on DNA methylation. The results from the two analyses did not overlap. In this chapter, the hypothesis was that by adding the air and dietary PAH8 exposures to create a single combined exposure, the results from the EWAS would overlap with those from the previous chapters. Additionally, combining two of the principal sources of PAH exposure allows for a more comprehensive measure of exposure. The analytical processes employed here followed the statistical methodology used in the previous two chapters as described in the Methods Chapter.

Table 6.1 Table of genes most frequently reported to be differentially methylated in association with smoking. The total number of reports column includes the total number of times each CpG probe was reported by each study, with some studies reporting multiple probes associated with the same gene, and in many instances the same probe was reported by more than one study.

Gene	Number of Reports	Number of CpGs Reported	Direction	References
<i>AHRR</i>	239	52	-	49,380,429–431,433,434,436– 438,441,444,415,446,448,449,416,417,421–423,425,428
			+	415,416,439,440,446,448,449,422,425,428–430,433,434,438
			NA	447
<i>GFI1</i>	78	11	-	380,415,438,446,448,449,416,422,425,428–431,433
			+	416,440
			NA	447
<i>MYO1G</i>	58	5	-	415,422,425,431,448
			+	380,415,437–439,446,448,449,416,417,422,425,428–430,433
			NA	444
<i>CYP1A1</i>	46	12	-	415
			+	415,416,428–430,433,438,439,446
			NA	444
<i>PRDM16</i>	43	40	-	415–417,421,429
			+	415,416,421
			-	380,415,438,439,445,446,448,449,416,425,426,428–431,433
<i>CNTNAP2</i>	39	8	+	380,415,416,425,433,440,446,448,449
			-	415,416,433,449
			+	416
<i>RUNX3</i>	36	26	+	416
			NA	432
			-	380,415,416,421,422,425,448,449
<i>C14orf43</i>	36	11	NA	444
			-	380,415,438,442,444,448,449,416,422,425,426,428,429,431,433
			+	416,440
<i>LRP5</i>	28	11	-	380,415,416,425,429,431,444,448,449
			+	380,415,421,440
			-	380,415,416,421,425,431,444,448,449
<i>RARA</i>	24	11	+	415,416,425,448
			-	415,416,428,433
			+	433
<i>GALNT2</i>	22	11	-	415,416,421,425,448
			+	444
			-	415,416
<i>ITGAL</i>	20	7	+	380,415,443,448,449
			-	415,416,429
			NA	416
<i>HIVEP3</i>	20	16	-	380,415,442,444,445,448–450,417,418,421–423,425,426,431
			+	440
			NA	447
<i>F2RL3</i>	19	1	-	415–417,421
			+	415,416,419,428–430,433
			-	415,416,449
<i>RPTOR</i>	18	16	+	415,421,422,448
			-	444
			NA	444
<i>VAR5</i>	18	15	-	415,416,425,448,449
			+	415,440
			-	380,415,422,425,431,444,448,449
<i>GNA12</i>	18	7	+	416,433
			-	
			+	
<i>PRSS23</i>	18	5	-	
			+	
			-	

6.2 Results

6.2.1 Combined Air and Dietary PAH8 Exposure

In order to be able to combine air and dietary PAH8 exposures, these were converted to Z-scores and then added for all subjects in the EPIC-Italy Training and Testing, and the EPIC-NL datasets. Summary statistics and quartiles of the Z-scores of combined PAH8 exposure are shown in Table 6.2, with the distributions of the three datasets in Figure 6.1. There were no statistically significant differences ($p > 0.05$) between the Training, Testing and EPIC-NL datasets for the Z-scores of combined PA8 exposure.

6.2.2 Effects of Combined Air and Dietary PAH8 Exposure on Global Methylation

When comparing the average methylation of all CpG sites between subjects in the lowest combined air and dietary PA8 intake quartile (Q1) and those in the higher quartiles (Q2, Q3, and Q4) (quartile cut-offs are summarised in Table 6.2), no associations were observed in any of the datasets (Table 6.3). When assessing the trend, the Training dataset showed a statistically significant negative relationship between global mean methylation and Z-scores of combined air and dietary PAH8 exposure (β coefficient = -0.00074, $p = 0.031$) but this was not replicated in the Testing or EPIC-NL datasets (Table 6.3). In addition to the lack of statistical significance, across the quartiles and trends, the direction of change was inconsistent both within and across datasets.

Table 6.2 Table of cohort characteristics for the training, testing and EPIC-NL EPIC subjects.

		EPIC-Italy Training (N=493)	EPIC-Italy Testing (N=208)	EPIC-NL (N=132)
Z-score of Combined PAH8 Exposure	Range	-3.28-5.85	-3.29	-4.92
	Mean	0	0	0
	Median	-0.20	-0.05	-0.002
Quartiles of Z-score of Combined PAH8 Exposure	Q1	-3.28 - -0.94	-3.29 - -0.90	-4.92 - -0.72
	Q2	-0.94 - -0.19	-0.90 - -0.047	-0.72 - -0.0017
	Q3	-0.19 - 0.73	-0.047 - 0.70	-0.0017 - 0.76
	Q4	0.73 - 5.85	0.70 - 6.16	0.76 - 5.00

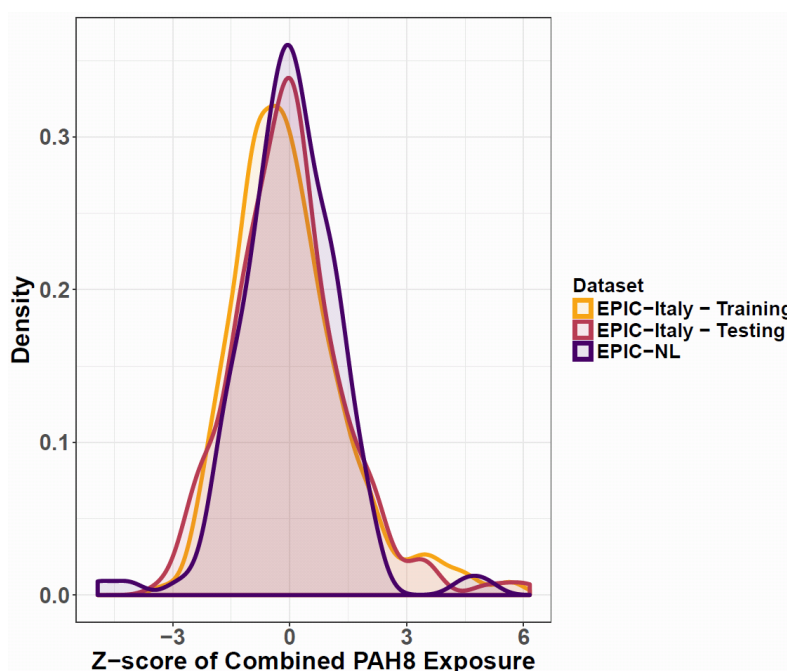


Figure 6.1 Distribution of Z-scores for the combined air and dietary PAH8 exposures for each of the three datasets: EPIC-Italy Training (yellow), EPIC-Italy - Testing (red) and EPIC-NL (purple).

Table 6.3 Table of beta regression results looking for differences in global methylation between quartiles of combined air and dietary PAH8 exposures. The lowest quartile (Q1) was used as the reference quartile.

PAH Quartile	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-NL (N=132)		
	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value
Q2	0.0004	-0.002; 0.003	0.70	-0.003	-0.005; 0.004	0.89	0.004	-0.005; 0.01	0.37
Q3	-0.0002	-0.002; 0.002	0.84	-0.004	-0.008; 0.0006	0.09	0.002	-0.007; 0.01	0.71
Q4	-0.002	-0.004; 0.0008	0.20	0.003	-0.002; 0.008	0.20	-0.0003	-0.008; 0.008	0.95
Trend	-0.00074	-0.0014; -0.000067	0.031	0.00048	-0.00069; 0.0016	0.42	-0.00065	-0.0028; 0.0015	0.56

6.2.3 Effects of Combined Air and Dietary PAH8 Exposure on Methylation at Genomic Regions

No associations were found between subjects in the lowest quartile of combined air and dietary PH8 intake (Q1) and subjects in the higher quartiles (Q2, Q3, and Q4) at shore regions in any of the three datasets (Table 6.4). In the Training dataset, a borderline significant association was found between Q1 and Q4 subjects (β coefficient = -0.003, $p = 0.05$), with Q4 subjects having more hypomethylated shore regions compared to subjects in the lowest quartile of exposure. However, the direction of methylation change, despite being non-significant, at CpG shores in the Testing and EPIC-NL datasets were different. Similar results were observed for CpG shelves: the only significant association was found between Q1 and Q3 subjects in the Testing dataset (β coefficient = -0.01, $p = 0.020$), with Q3 subjects having lower levels of methylation at CpG shelves compared to Q1 subjects (Table 6.4). Analysis of methylation at CpG islands, showed that Q4 subjects in the Training dataset had lower methylation at CpG island regions compared to subjects in the lowest exposure quartile (Q1) (β coefficient = -0.007, $p = 0.002$) (Table 6.4). No other significant associations between exposure and methylation at CpG islands were observed for any of the datasets and the direction of change across quartiles and datasets was inconsistent. Promoter regions were statistically significantly hypomethylated in subjects in the highest exposure quartile (Q4) compared to Q1 subjects in the Training dataset (β coefficient = -0.003, $p = 0.047$) (Table 6.5). Subjects in Q3 of the Testing datasets had lower methylation levels at 3' UTR and intergenic regions compared to subjects in Q1 (β coefficient = -0.01, $p = 0.02$) (Table 6.5). As reported above, these results were not replicated across quartiles or datasets. Analysis of the overall trends for all of these genomic regions (Table 6.4 and Table 6.5) supported the findings of the quartile analyses: no statistically significant associations were observed between the Z-scores of combined air and dietary PAH8 exposure and the mean methylation levels at each of the genomic regions. The only exception to this was mean methylation at CpG shores in the Training dataset which was significantly negatively associated with combined air and dietary PAH8 exposure (β coefficient = -0.00099, $p = 0.011$) (Table 6.4).

Table 6.4 Table of beta regression results looking differences in methylation levels at CpG islands, shores and shelves between quartiles of combined air and dietary PAH8 exposures. The lowest quartile (Q1) was used as the reference quartile.

<u>Genomic Region</u>	<u>PAH Quartile</u>	<u>EPIC-Italy – Training</u>			<u>EPIC-Italy - Testing</u>			<u>EPIC-NL</u>		
		<u>β coefficient</u>	<u>P value</u>	<u>Confidence Interval</u>	<u>β coefficient</u>	<u>P value</u>	<u>Confidence Interval</u>	<u>β coefficient</u>	<u>P value</u>	<u>Confidence Interval</u>
Shores	Q2	-0.0003	0.82	-0.003; 0.002	0.001	0.60	-0.004; 0.007	0.004	0.35	-0.004; 0.01
	Q3	-0.0002	0.86	-0.003; 0.002	-0.002	0.38	-0.008; 0.003	0.004	0.34	-0.004; 0.01
	Q4	-0.003	0.05	-0.005; 0.00004	0.003	0.25	-0.002; 0.009	0.002	0.64	-0.006; 0.01
	Trend	-0.00099	0.011	-0.0018; -0.00022	0.00038	0.58	-0.00096; 0.0017	0.00015	0.889	-0.0021; 0.0024
Shelves	Q2	0.003	0.13	-0.0009; 0.007	-0.009	0.077	-0.02; 0.001	0.01	0.25	-0.007; 0.03
	Q3	-0.001	0.48	-0.005; 0.002	-0.01	0.020	-0.02; -0.002	0.008	0.32	-0.008; 0.02
	Q4	0.001	0.47	-0.003; 0.005	0.0009	0.86	-0.009; 0.01	0.003	0.67	-0.01; 0.02
	Trend	-0.00074	0.33	-0.0022; 0.00075	0.00026	0.84	-0.0023; 0.0028	-0.00046	0.84	-0.0048; 0.0037
CpG Islands	Q2	-0.003	0.14	-0.008; 0.001	0.01	0.061	-0.0005; 0.02	0.0005	0.93	-0.01; 0.01
	Q3	0.0008	0.71	-0.004; 0.005	0.006	0.33	-0.006; 0.02	0.0007	0.92	-0.01; 0.01
	Q4	-0.007	0.002	-0.01; -0.003	0.01	0.075	-0.001; 0.02	-0.004	0.52	-0.02; 0.008
	Trend	-0.0015	0.068	-0.0032; 0.00011	0.002	0.15	-0.00080; 0.0050	-0.00072	0.67	-0.0040; 0.0026

Table 6.5. Table of beta regression results looking differences in methylation levels at promoter, 3' UTR, gene body and intergenic regions between quartiles of combined air and dietary PAH8 exposures. The lowest quartile (Q1) was used as the reference quartile.

Genomic Region	PAH Quartile	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-NL (N=132)		
		β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value	β coefficient	Confidence Interval	P value
Promoters	Q2	-0.0009	-0.004; 0.002	0.54	0.004	-0.003; 0.01	0.28	0.001	-0.008; 0.01	0.80
	Q3	0.0004	-0.002; 0.003	0.79	-0.002	-0.009; 0.005	0.60	-0.0002	-0.009; 0.008	0.96
	Q4	-0.003	-0.006; 0.00004	0.047	0.005	-0.002; 0.01	0.17	-0.004	-0.01; 0.005	0.39
	Trend	-0.00083	-0.0018; 0.00016	0.10	0.00084	-0.00091; 0.0023	0.35	-0.0011	-0.0033; 0.0012	0.35
3' UTR	Q2	0.002	-0.0008; 0.006	0.14	-0.007	-0.02; 0.002	0.13	0.008	-0.007; 0.02	0.28
	Q3	-0.001	-0.004; 0.002	0.54	-0.01	-0.02; -0.001	0.02	0.007	-0.008; 0.02	0.38
	Q4	0.0002	-0.003; 0.004	0.91	0.0003	-0.009; 0.009	0.96	0.004	-0.01; 0.02	0.61
	Trend	-0.00075	-0.0021; 0.00060	0.27	-0.00019	-0.0024; 0.0020	0.87	-0.00024	-0.0040; 0.0036	0.90
Gene Body	Q2	0.001	-0.001; 0.003	0.38	-0.003	-0.008; 0.003	0.34	0.006	-0.004; 0.02	0.24
	Q3	-0.0005	-0.003; 0.002	0.67	-0.005	-0.01; 0.0003	0.07	0.004	-0.006; 0.01	0.47
	Q4	-0.001	-0.003; 0.001	0.42	0.003	-0.003; 0.008	0.34	0.002	-0.007; 0.01	0.62
	Trend	-0.00076	-0.0015; 0.000014	0.054	0.00045	-0.00090; 0.0018	0.51	-0.00022	-0.0028; 0.0024	0.87
Intergenic	Q2	0.001	-0.002; 0.004	0.39	-0.002	-0.008; 0.004	0.53	0.006	-0.005; 0.02	0.31
	Q3	-0.0008	-0.004; 0.002	0.59	-0.006	-0.01; -0.0003	0.04	0.005	-0.006; 0.02	0.41
	Q4	-0.0008	-0.004; 0.003	0.65	0.003	-0.003; 0.009	0.40	0.002	-0.009; 0.01	0.74
	Trend	-0.00084	-0.0018; 0.000096	0.078	0.00038	-0.0012; 0.0019	0.63	-0.0018	-0.0031; 0.0028	0.90

6.2.4 EWAS Results

The EWAS was carried out on the 493 EPIC-Italy subjects in the Training dataset, regressing the Z-scores of combined air and dietary PAH8 exposures against DNA methylation. Results were available for 362,394 probes after removal of probes whose model did not converge, probes known to cross-hybridise, probes located on SNPs, and probes located on the sex chromosomes. Of these probes, 16 passed the threshold of Bonferroni correction for multiple Testing ($p < 1.38 \times 10^{-7}$), and 274 passed FDR correction ($p < 3.76 \times 10^{-5}$; $q < 0.05$) (Figure 6.2A). Approximately two-thirds of the FDR-significant probes were hypomethylated (N = 175 hypomethylated, N = 99 hypermethylated) (Figure 6.2B). Some inflation of the p values was observed, and the inflation factor lambda was calculated to be 1.18 (Figure 6.2C). The biggest changes in methylation of the FDR-significant probes were a loss in methylation of 1.34% (cg05703053; subject with lowest combined air and dietary PAH8 exposure Z-score = 43.0% methylated; subject with highest combined air and dietary PAH8 exposure Z-score = 17.6% methylated), and a gain in methylation of 0.79% (cg06457011; subject with lowest combined air and PAH8 exposure Z-score = 32.1% methylated; subject with highest combined air and dietary PAH8 exposure Z-score = 39.2% methylated) for every unit change in the Z-scores of combined air and dietary PAH8 exposure. These values are interpretable as percentage change in methylation per standard deviation in combined air and diet PAH8 exposure.

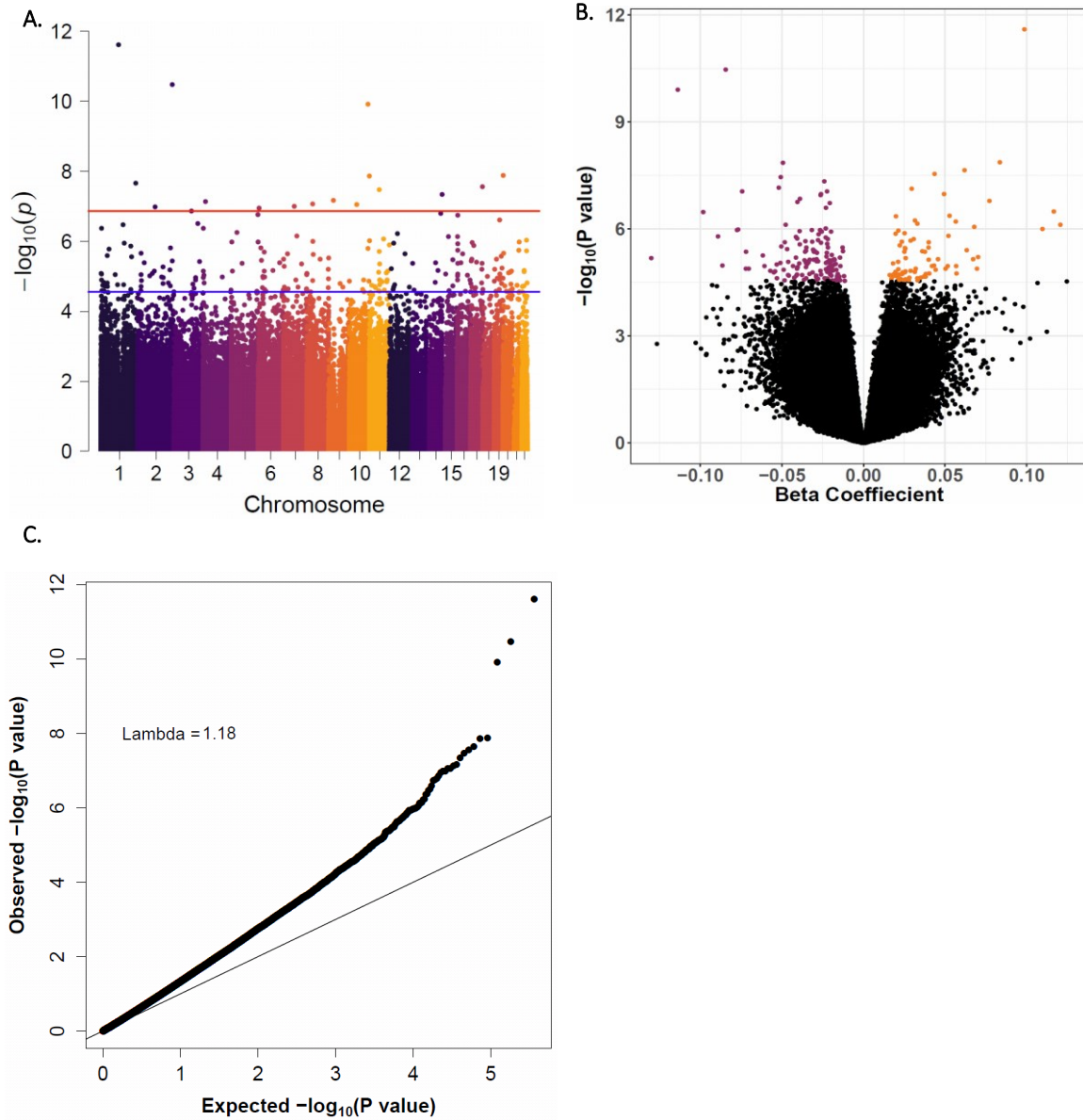


Figure 6.2 A: Manhattan plot showing the $-\log_{10}$ transformed p values of the 362,394 CpG probes tested arranged by chromosome. The red line indicates the threshold for Bonferroni correction for multiple testing ($p < 1.38 \times 10^{-7}$), and the blue line indicates the FDR threshold ($p < 3.76 \times 10^{-5}$; $q < 0.05$). B: Volcano plot showing the $-\log_{10}$ transformed p values of the 362,394 CpG probes against the β -coefficient for dietary PAH8 intake. Coloured points indicate significance after FDR correction, with red points indicating a decrease in methylation and orange points indicating an increase in methylation. C: QQ plot showing the observed $-\log_{10}$ transformed p values against the expected $-\log_{10}$ transformed p values from the EWAS.

The Bonferroni significant probes from the EWAS carried out in the Training dataset were analysed in the Testing and EPIC-NL datasets, and the results are summarised in Table 6.6. The results for the 274 FDR-significant probes in all three datasets can be found in Appendix 4 Table 9.15. None of the 16 probes were significant in all three datasets, and none were significant in both the Training and Testing datasets. Two probes were significant and showed methylation changes in the same direction in both the Training and EPIC-NL datasets: cg00466488 (Training: $\beta = 0.099$, $p = 2.48 \times 10^{-12}$; EPIC-NL: $\beta = 0.069$, $p = 0.049$) and cg14677909 (Training: $\beta = -0.074$, $p = 8.82 \times 10^{-8}$; EPIC-NL: $\beta = -0.076$, $p = 0.029$). Only two probes showed methylation differences in the same direction in all three datasets: cg03317082 (Training: $\beta = 0.062$, $p = 2.24 \times 10^{-8}$; Testing: $\beta = 0.006$, $p = 0.688$; EPIC-NL: $\beta = 0.002$, $p = 0.955$), and cg06009497 (Training: $\beta = -0.022$, $p = 8.80 \times 10^{-8}$; Testing: $\beta = -0.003$, $p = 0.579$; EPIC-NL: $\beta = -0.003$, $p = 0.770$). The methylation status of the probes with the biggest methylation changes (cg05703053 and cg06457011) were assessed in the subjects with the lowest and highest dietary PAH8 exposures in the Testing and EPIC-NL datasets. In the Testing dataset, probe cg05703053 in the subject with the lowest exposure was 38.0% methylated, and the most highly exposed subject was 60.6% methylated which indicates hypermethylation rather than the hypomethylation observed in the Training dataset. Similarly, in the EPIC-NL dataset, probe cg05703053 in the subject with the lowest exposure was 22.5% methylated, and the most highly exposed subject was 40.7% methylated. Probe cg06457011 was 39.6% methylated in the subject with the lowest air PAH8 exposure and 41.7% methylated in the subject with the highest air PAH8 exposure in the Testing dataset. This probe in the EPIC-NL dataset was 34.7% methylated and 5.0% methylated in the subjects with the lowest and highest dietary PAH8 exposures respectively, which again is the opposite trend to that observed in the Training dataset.

Table 6.6 Model results for the Bonferroni significant ($p < 1.38 \times 10^{-7}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship between combined air and dietary PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status.

Probe ID	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-NL (N=132)		
	<u>B</u> <u>Coefficient</u>	<u>95%</u> <u>Confidence</u> <u>Interval</u>	<u>P Value</u>	<u>B</u> <u>Coefficient</u>	<u>95%</u> <u>Confidence</u> <u>Interval</u>	<u>P Value</u>	<u>B</u> <u>Coefficient</u>	<u>95%</u> <u>Confidence</u> <u>Interval</u>	<u>P Value</u>
cg00466488	0.099	0.071; 0.126	2.48E-12	-0.002	-0.042; 0.037	0.915	0.069	0; 0.137	0.049
cg22374586	-0.085	-0.109; -0.06	3.38E-11	-0.009	-0.048; 0.03	0.642	0.083	-0.002; 0.169	0.057
cg15275103	-0.114	-0.148; - 0.079	1.21E-10	0.017	-0.041; 0.076	0.562	0.009	-0.053; 0.071	0.771
cg14083397	0.084	0.055; 0.112	1.32E-08	-0.007	-0.028; 0.015	0.542	0.01	-0.042; 0.063	0.702
cg18308755	-0.049	-0.066; - 0.032	1.39E-08	0.011	-0.014; 0.036	0.388	0.014	-0.044; 0.073	0.63
cg03317082	0.062	0.04; 0.084	2.24E-08	0.006	-0.024; 0.037	0.683	0.002	-0.062; 0.066	0.955
cg18576374	0.044	0.028; 0.059	2.79E-08	-0.011	-0.038; 0.017	0.448	0	-0.05; 0.05	0.993
cg01981334	-0.051	-0.069; - 0.033	3.44E-08	0	-0.015; 0.016	0.978	0.024	-0.009; 0.057	0.157
cg27158340	-0.024	-0.033; - 0.015	4.57E-08	0	-0.016; 0.017	0.953	0.012	-0.024; 0.048	0.499
cg14286514	-0.052	-0.071; - 0.033	6.83E-08	-0.016	-0.045; 0.014	0.298	0.008	-0.045; 0.061	0.775
cg17304168	0.029	0.019; 0.04	7.31E-08	0.006	-0.014; 0.025	0.554	0	-0.053; 0.052	0.993
cg06009497	-0.022	-0.031; - 0.014	8.80E-08	-0.003	-0.016; 0.009	0.579	-0.003	-0.022; 0.016	0.770
cg14677909	-0.074	-0.102; - 0.047	8.82E-08	0.025	-0.047; 0.097	0.494	-0.076	-0.144; - 0.008	0.029

cg09214099	-0.026	-0.036; - 0.016	1.01E-07	0.006	-0.009; 0.022	0.421	0.011	-0.021; 0.042	0.513
cg12448298	0.049	0.031; 0.068	1.05E-07	-0.002	-0.031; 0.027	0.878	-0.004	-0.06; 0.053	0.903
cg03349397	-0.026	-0.036; - 0.017	1.12E-07	0.002	-0.015; 0.018	0.825	0.007	-0.033; 0.047	0.737

Of the 16 Bonferroni significant probes, 11 were located in genic regions, including 3 associated with promoters, and the remaining 5 were located intergenically (Table 6.7). The genes associated with these probes, and other characteristics, are summarised in Table 6.7 and Appendix 4 Table 9.16 for the 274 FDR-significant probes. The genomic distribution of the longer list of 274 FDR-significant probes was analysed and compared to the distribution of all 362,394 probes analysed (Table 6.8; Figure 6.3A). Fisher's tests were carried out to determine whether the results obtained were different from those expected and the results are summarised in Table 6.8 and Figure 6.3A. More methylation changes than expected occurred in exon regions (OR = 2.14, $p = 2.73 \times 10^{-5}$) and less methylation changes than expected occurred at promoter regions (OR = 0.64, $p = 0.0022$). When comparing the direction of change at all genomic regions to the overall ratio of hypomethylated to hypermethylated probes (ratio = 1.77) (Table 6.9; Figure 6.3B), more hypomethylation events occurred than expected based on the ratio at exons (OR = 3.39, $p = 0.003$) and CpG islands (OR = 11.63, $p = 1.01 \times 10^{-5}$). More hypermethylation changes than expected took place in intergenic (OR = 0.43, $p = 0.014$) and intronic (OR = 0.47, $p = 0.017$) regions.

Table 6.7 Table of characteristics of probes found to be significantly associated with combined air and dietary PAH8 exposure at the Bonferroni level ($p < 1.38 \times 10^{-7}$) in the training dataset.

Probe ID	Chromosome	Position	UCSC RefGene Name	Gene Location	Relation to CpG Island	Methylation Change Direction
cg00466488	1	118148927	<i>FAM46C</i>	5'UTR	Island	+
cg03317082	1	234748618			South Shore	+
cg12448298	2	115822039	<i>DPP10</i>	Body		+
cg22374586	2	232220566				-
cg17304168	4	15626104	<i>FBXL5</i>	Body		+
cg03349397	6	6588693	<i>LY86</i>	TSS1500		-
cg09214099	7	72791740			Island	-
cg06009497	8	37695050	<i>GPR124</i>	Body	North Shelf	-
cg14286514	9	32525315	<i>DDX58</i>	Body	North Shore	-
cg14677909	10	48807341	<i>PTPN20B</i>	Body		-
cg15275103	10	124893024			Island	-
cg18308755	10	134065956	<i>STK32C</i>	Body		-
cg01981334	11	64877237	<i>C11orf2</i>	Body	North Shore	-
cg27158340	14	105603389			Island	-
cg18576374	17	78549371	<i>RPTOR</i>	Body		+
cg14083397	20	388473	<i>RBCK1</i>	TSS1500	Island	+

Table 6.8. Table of Fisher’s test results comparing the number of differentially methylated probes (N = 274) and all tested probes (N = 362,394) in the training dataset EWAS at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred.

Genomic Region	Odds Ratio	Confidence Interval	P Value
3' UTR	0.64	0.21 – 1.52	0.46
5' UTR	0.84	0.27 – 1.98	1
Exon	2.14	1.51 – 2.98	2.73 x 10⁻⁵
Intergenic	1.18	0.83 – 1.65	0.32
Intron	0.89	0.65 – 1.21	0.51
Non-coding	1.98	0.90 – 3.83	0.054
Promoter	0.64	0.47 – 0.86	0.0022
TTS	0.47	0.10 – 1.38	0.23
CpG Island	1.42	0.97 – 2.02	0.060
LINE	0.57	0.18 – 1.35	0.29
SINE	0.96	0.41 – 1.93	1
LTR	0.61	0.16 – 1.57	0.43
Other	1.15	0.37 – 2.72	0.63

Table 6.9. Table of Fisher’s test results comparing the number of hypermethylation changes (N=99) to hypomethylation changes (N=175) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected.

Genomic Region	Odds Ratio	Confidence Interval	P Value
3' UTR	0.85	0.095 – 10.29	1
5' UTR	2.29	0.22 – 113.97	0.66
Exon	3.39	1.41 – 9.42	0.0030
Intergenic	0.43	0.21 – 0.88	0.014
Intron	0.47	0.24 – 0.90	0.017
Non-coding	0.70	0.15 – 3.61	0.73
Promoter	0.96	0.51 – 1.86	1
TTS	0	0 – 1.36	0.046
CpG Island	11.63	2.86 – 102.22	1.01 x 10⁻⁵
LINE	0.37	0.031 – 3.30	0.36
SINE	0.94	0.18 – 6.19	1
LTR	1.71	0.13 – 90.60	1
Other	0.37	0.031 – 3.30	0.36

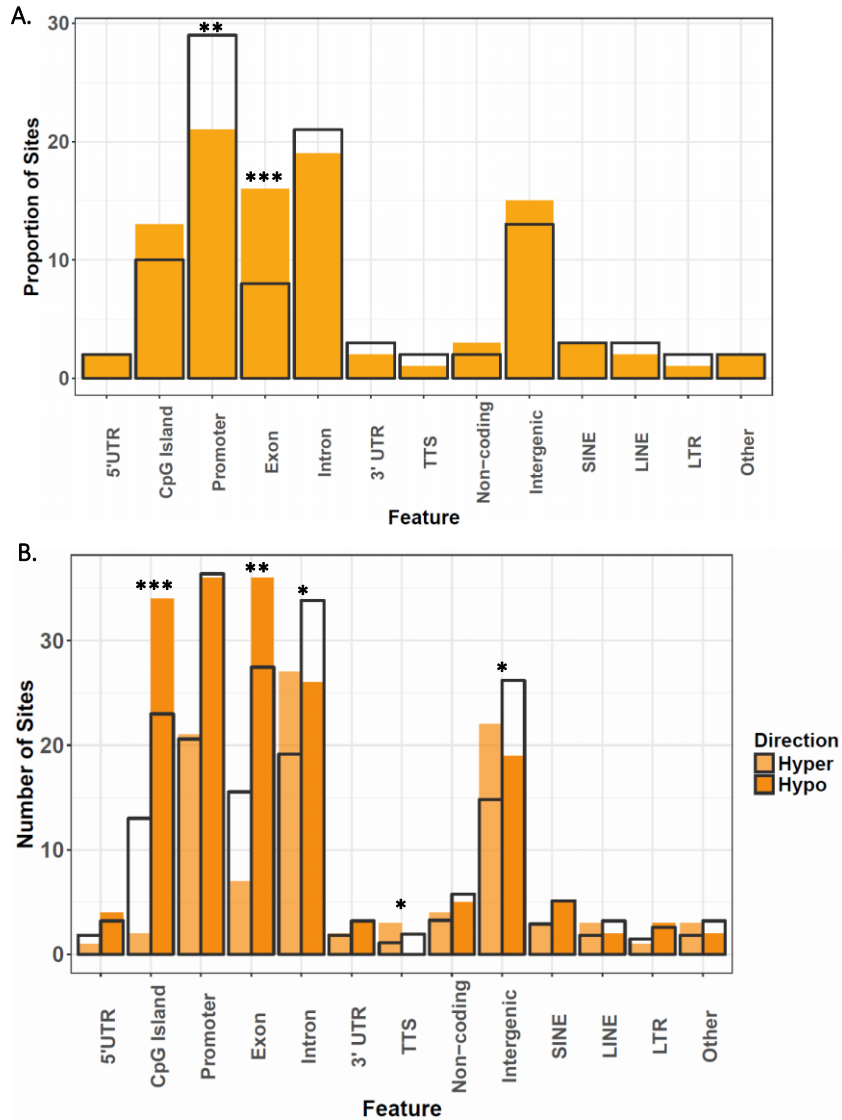


Figure 6.3 A: Comparison of the genomic distribution of differentially methylated probes (N = 274) and all tested probes (N = 362,394) in the training dataset EWAS. The filled yellow bars show the proportion of significant probes, the grey outline bars show the proportion of all probes tested, i.e. the expected distribution. **B:** Comparison of the genomic distribution of hypermethylated (N = 175) and hypomethylated (N = 99) probes. As in **A**, the filled yellow bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. For both plots, * indicates $p < 0.05$, ** indicates $p < 0.01$ and *** indicates $p < 0.001$ following Fisher's Exact test.

6.2.5 Building a Methylation Index of Combined Air and Dietary PAH8 Exposure

A methylation index of combined air and dietary PAH8 exposure was developed using the 274 FDR-significant probes identified from the EWAS. The optimal model parameters were $\alpha = 0$ and $\lambda = 0.95$, meaning that all 274 probes were included in the model in addition to sex, age, cancer case status, and smoking status as indicated and a penalty factor of 0.95 was applied to all model coefficients. These parameters, were found using the train function from the caret package in R using the Training dataset to train the model. The best performing model had a RMSE of 1.08 and explained 44.9% of the variance ($R^2 = 0.449$). Model performance was assessed in both in the Training and Testing datasets and the results are shown in Figures 4A and 4B. Due to the underlying differences in both cohort characteristics and methylation distributions, the methylation index was not tested on the EPIC-NL dataset. As expected, the model performed well on the data on which it was trained (Training dataset; Figure 6.4A) with a strong correlation between the predicted and real Z-scores of combined air and dietary PAH8 exposure (Spearman's Rho = 0.85, $p < 2.2 \times 10^{-16}$). Performance declined significantly when the model was applied to the Testing dataset (Figure 6.4B), with no correlation between predicted and real Z-scores of combined air and dietary PAH8 exposure (Spearman's Rho = 0.11, $p = 0.11$).

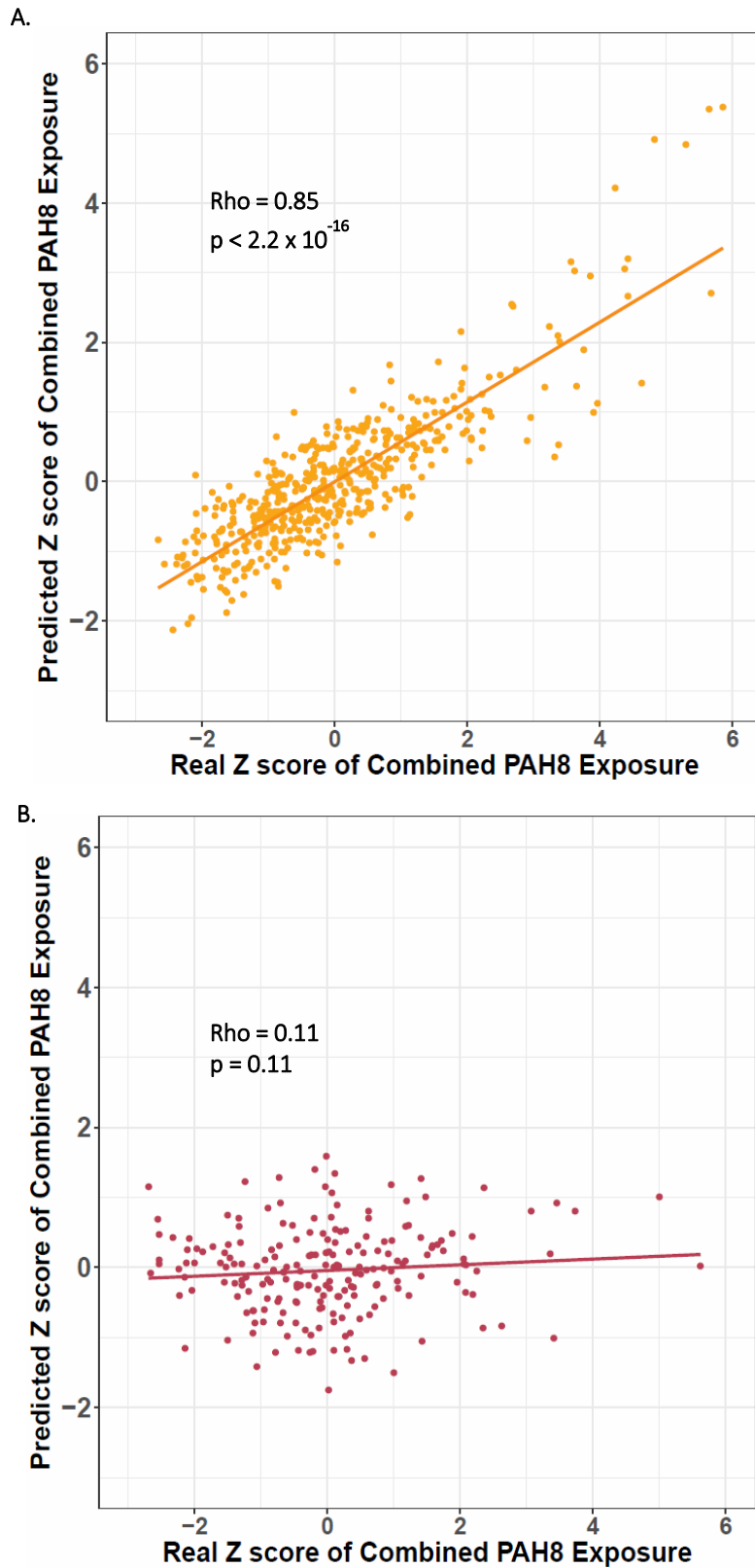


Figure 6.4 A and B: Plots showing the correlation between the combined PAH8 exposure predicted by the elastic net model against the real combined PAH8 exposure for each subject in the training and testing datasets respectively.

6.2.6 Comparison of the EWAS Results of Air, Dietary, and Combined Air and Dietary PAH8 Exposure

The EWAS of air PAH8 exposure found 204 FDR-significant probes, the EWAS of dietary PAH8 exposure found 171 FDR-significant probes, and the EWAS of Z-scores of combined air and dietary PAH8 exposure found 274 FDR-significant probes. There were no overlaps between the air and dietary PAH8 EWAS results, but 35 probes were significant in both the EWAS of dietary PAH8 exposure and combined PAH8 exposure, and 58 probes were significant both in the EWAS of air PAH8 exposure and combined PAH8 exposure (Figure 6.5). Figure 6.6 aims to summarise the results of the Fisher's tests carried out on all three sets of EWAS results looking for more (enrichment) or less (depletion) methylation changes at genomic regions than expected based on the genomic distribution of all tested probes. Significantly less methylation changes than expected occurred at gene promoter regions in all three EWAS, with significantly more changes than expected occurring at exon region in both the air and the combined PAH8 exposure EWASs (Figure 6.6).

6.2.6.1 Air and Combined PAH8 Exposure EWAS Results

Table 6.10 shows the model results for the 58 probes that were FDR-significant in both the air and combined PAH8 exposure EWASs. For all probes, the direction of methylation was the same in both EWASs, however the confidence intervals only overlapped for two probes: cg17304168 and cg12448298 (Table 6.10). The effect sizes (β coefficients) were larger for all probes in the air PAH8 exposure EWAS. Table 6.11 shows the characteristics of all the 58 overlapping probes. The majority of these probes were located in genic regions, mostly the gene body (N = 27), followed by the promoter (N = 15), and the 3' UTR (N = 3) (Table 6.10). The remaining 13 probes were located at intergenic regions (Table 6.11).

6.2.6.2 Dietary and Combined PAH8 Exposure EWAS Results

All 35 of the common CpG probes showed methylation changes in the same direction in both EWASs and the results for these are shown in Table 6.12. The confidence intervals did not overlap for any of

the probes, but interestingly, the effect sizes and confidence intervals from the dietary PAH8 exposure EWAS were approximately 2 orders of magnitude than those from the combined PAH8 exposure EWAS (Table 6.12). The characteristics of these probes are summarised in Table 6.13. The majority of the probes were associated with genic regions, with 13 in a gene body, 9 in a promoter region, and 4 in the 3' UTR of a gene. Only 9 probes were located intergenically.

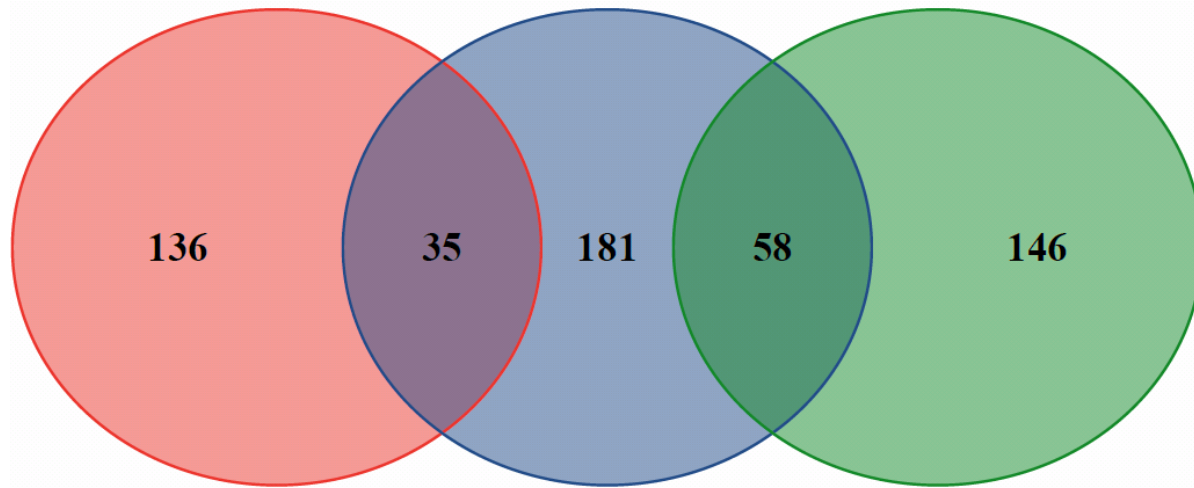


Figure 6.5 Venn diagram showing the overlap between the FDR-significant CpG sites between the air PAH8 exposure model (green), the dietary PAH8 exposure model (red), and the model of combined PAH8 exposure (blue).

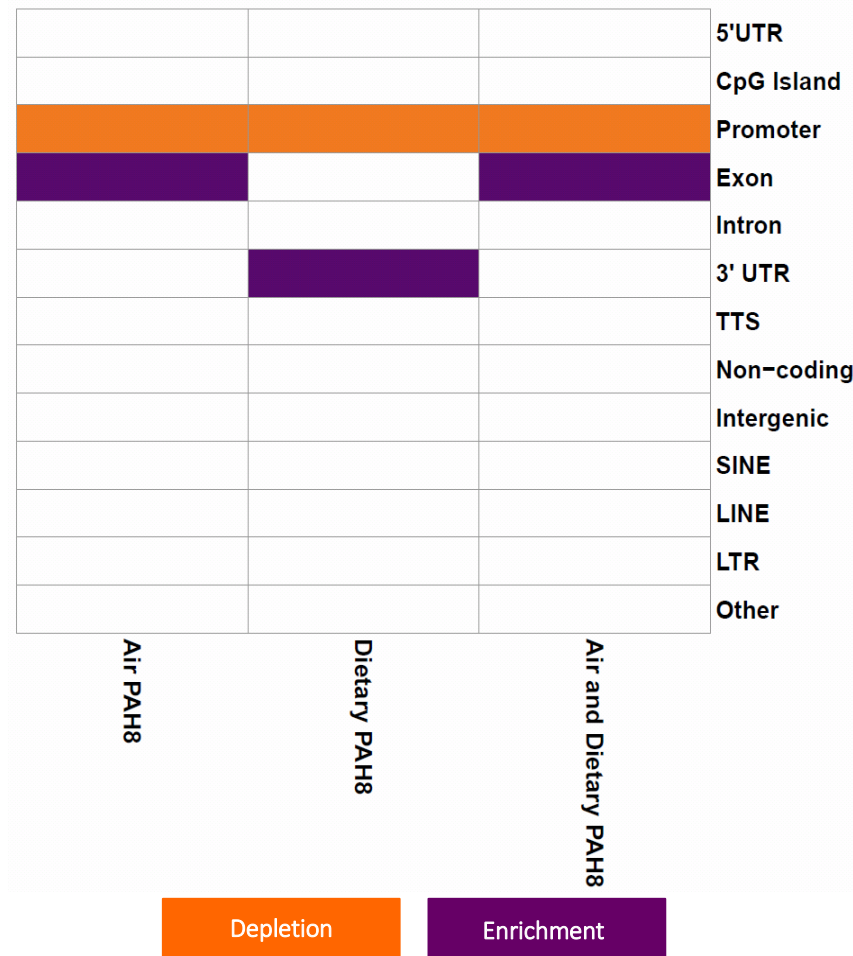


Figure 6.6. Heatmap summarising the Fisher's test results looking for enrichment or depletion of methylation differences at particular genomic features across the three models: Air PAH8 exposure, dietary PAH8 exposure and combined air and dietary PAH8 exposure. White boxes indicate non-significant results ($p > 0.05$), and coloured boxes indicate significant results ($p < 0.05$). Orange indicates less methylation differences than expected (depletion) and purple indicates more methylation differences than expected (enrichment).

Table 6.10 Model results for probes that were FDR significant ($q < 0.05$) in both the combined air and diet PAH8 exposure EWAS model and the air PAH8 exposure only EWAS model. All results are from beta regression models run on the training dataset and were adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre.

Probe ID	Combined Air and Diet PAH8 Exposure			Air PAH8 Exposure		
	<u>B Coefficient</u>	<u>95% Confidence Interval</u>	<u>P Value</u>	<u>B Coefficient</u>	<u>95% Confidence Interval</u>	<u>P Value</u>
cg00466488	0.099	0.071; 0.126	2.48E-12	0.353	0.278; 0.428	1.82E-20
cg22374586	-0.085	-0.109; -0.06	3.38E-11	-0.212	-0.286; -0.138	1.92E-08
cg15275103	-0.114	-0.148; -0.079	1.21E-10	-0.371	-0.473; -0.27	6.38E-13
cg14083397	0.084	0.055; 0.112	1.32E-08	0.316	0.238; 0.395	3.30E-15
cg18308755	-0.049	-0.066; -0.032	1.39E-08	-0.143	-0.192; -0.093	1.42E-08
cg03317082	0.062	0.04; 0.084	2.24E-08	0.192	0.13; 0.255	1.92E-09
cg18576374	0.044	0.028; 0.059	2.79E-08	0.181	0.139; 0.224	3.44E-17
cg01981334	-0.051	-0.069; -0.033	3.44E-08	-0.192	-0.242; -0.141	1.13E-13
cg14286514	-0.052	-0.071; -0.033	6.83E-08	-0.129	-0.185; -0.073	6.06E-06
cg17304168	0.029	0.019; 0.04	7.31E-08	0.071	0.039; 0.103	1.27E-05
cg06009497	-0.022	-0.031; -0.014	8.80E-08	-0.052	-0.077; -0.028	2.49E-05
cg14677909	-0.074	-0.102; -0.047	8.82E-08	-0.29	-0.366; -0.214	8.55E-14
cg12448298	0.049	0.031; 0.068	1.05E-07	0.117	0.064; 0.171	1.76E-05
cg12389423	-0.039	-0.054; -0.024	1.39E-07	-0.106	-0.148; -0.063	1.13E-06
cg12497870	-0.023	-0.032; -0.014	2.52E-07	-0.062	-0.088; -0.036	2.09E-06
cg18592273	0.117	0.072; 0.161	3.18E-07	0.372	0.248; 0.497	4.63E-09
cg06466757	0.053	0.032; 0.073	4.25E-07	0.15	0.089; 0.21	1.28E-06
cg26496372	0.031	0.019; 0.044	5.69E-07	0.121	0.086; 0.156	1.15E-11
cg02574894	0.057	0.034; 0.079	6.04E-07	0.179	0.117; 0.242	1.45E-08
cg06745145	0.033	0.02; 0.046	7.11E-07	0.093	0.054; 0.132	2.51E-06
cg07482202	0.121	0.073; 0.168	7.36E-07	0.397	0.261; 0.533	1.09E-08
cg14209037	-0.048	-0.067; -0.029	7.39E-07	-0.146	-0.2; -0.093	8.12E-08

cg22049858	0.068	0.041; 0.095	8.72E-07	0.212	0.137; 0.287	3.42E-08
cg22848598	-0.077	-0.108; - 0.046	1.02E-06	-0.232	-0.324; - 0.141	6.96E-07
cg12826791	-0.078	-0.109; - 0.046	1.07E-06	-0.253	-0.341; - 0.165	1.89E-08
cg16619935	-0.02	-0.028; - 0.012	1.18E-06	-0.054	-0.078; - 0.031	5.82E-06
cg19485911	0.052	0.031; 0.073	1.56E-06	0.154	0.092; 0.217	1.33E-06
cg04293602	-0.028	-0.039; - 0.016	1.99E-06	-0.076	-0.11; - 0.042	1.21E-05
cg24935556	0.023	0.014; 0.033	2.21E-06	0.066	0.037; 0.094	6.95E-06
cg02216727	-0.035	-0.05; -0.02	3.96E-06	-0.101	-0.144; - 0.057	6.19E-06
cg11315081	-0.072	-0.103; - 0.041	4.32E-06	-0.204	-0.293; - 0.114	8.44E-06
cg05703053	-0.062	-0.088; - 0.035	5.38E-06	-0.184	-0.263; - 0.105	5.04E-06
cg07480373	0.026	0.015; 0.037	6.51E-06	0.08	0.047; 0.113	1.98E-06
cg02583546	0.067	0.038; 0.097	6.90E-06	0.223	0.137; 0.308	3.40E-07
cg01731811	-0.034	-0.05; - 0.019	8.33E-06	-0.1	-0.144; - 0.056	9.55E-06
cg19428444	-0.039	-0.057; - 0.022	9.15E-06	-0.115	-0.166; - 0.064	8.96E-06
cg06856378	-0.087	-0.125; - 0.048	1.06E-05	-0.299	-0.412; - 0.185	2.32E-07
cg15407965	-0.043	-0.063; - 0.024	1.07E-05	-0.144	-0.199; - 0.089	3.48E-07
cg12088773	-0.061	-0.088; - 0.034	1.09E-05	-0.184	-0.263; - 0.106	3.70E-06
cg07356415	-0.021	-0.031; - 0.012	1.14E-05	-0.067	-0.095; - 0.039	2.39E-06
cg09961689	-0.04	-0.059; - 0.022	1.17E-05	-0.131	-0.184; - 0.079	9.35E-07
cg15233880	-0.072	-0.105; - 0.04	1.26E-05	-0.224	-0.321; - 0.127	6.53E-06
cg01003448	0.07	0.038; 0.101	1.28E-05	0.213	0.122; 0.304	4.31E-06
cg14677612	0.044	0.024; 0.064	1.37E-05	0.154	0.095; 0.214	3.54E-07
cg00256932	-0.054	-0.079; - 0.03	1.54E-05	-0.164	-0.235; - 0.093	6.51E-06
cg11060856	-0.022	-0.032; - 0.012	1.58E-05	-0.068	-0.097; - 0.038	7.15E-06
cg04117764	-0.041	-0.059; - 0.022	1.84E-05	-0.119	-0.173; - 0.065	1.54E-05
cg12653146	-0.038	-0.055; - 0.02	1.95E-05	-0.162	-0.21; - 0.113	5.33E-11
cg15289190	0.038	0.021; 0.056	2.17E-05	0.129	0.079; 0.179	4.97E-07

cg12126859	0.033	0.018; 0.049	2.72E-05	0.109	0.063; 0.154	2.79E-06
cg10629004	-0.048	-0.071; - 0.026	2.75E-05	-0.166	-0.233; - 0.099	1.33E-06
cg27190138	0.027	0.015; 0.04	2.81E-05	0.093	0.056; 0.13	8.74E-07
cg24303478	-0.015	-0.022; - 0.008	2.85E-05	-0.047	-0.068; - 0.026	1.06E-05
cg25679475	-0.033	-0.049; - 0.018	2.95E-05	-0.12	-0.164; - 0.075	1.41E-07
cg04678743	0.107	0.056; 0.157	3.18E-05	0.456	0.311; 0.602	7.92E-10
cg27605307	-0.042	-0.062; - 0.022	3.24E-05	-0.123	-0.181; - 0.066	2.72E-05
cg25170034	-0.038	-0.056; - 0.02	3.34E-05	-0.122	-0.175; - 0.069	6.44E-06
cg09576415	-0.016	-0.023; - 0.008	3.53E-05	-0.05	-0.072; - 0.028	8.53E-06

Table 6.11. Table of characteristics of probes found to be significantly associated with combined air and dietary PAH8 exposure and air PAH8 exposure only at the FDR significance levels ($q < 0.05$).

Probe ID	Chromosome	Position	UCSC RefGene Name	Gene Location	Relation to CpG Island	Methylation Change Direction
cg04117764	1	10917451				-
cg12653146	1	25919290				-
cg00466488	1	118148927	<i>FAM46C</i>	5'UTR	Island	+
cg03317082	1	234748618			South Shore	+
cg19428444	2	21023690	<i>C2orf43</i>	TSS1500	South Shore	-
cg24935556	2	21291088				+
cg12448298	2	115822039	<i>DPP10</i>	Body		+
cg06856378	2	160759118	<i>LY75</i>	Body	North Shore	-
cg05703053	2	169769616				-
cg07480373	2	216874286	<i>MREG</i>	Body	North Shelf	+
cg19485911	2	220380542	<i>ACCN4</i>	Body	South Shelf	+
cg22374586	2	232220566				-
cg25679475	3	118705126	<i>IGSF11;IGSF11</i>	Body		-
cg12389423	3	118864836	<i>C3orf30</i>	TSS200		-
cg18592273	3	161089930	<i>C3orf57</i>	TSS200	Island	+
cg06466757	4	1255808				+
cg11060856	4	5895410	<i>CRMP1</i>	TSS1500	South Shore	-
cg17304168	4	15626104	<i>FBXL5</i>	Body		+
cg15407965	4	128707242	<i>HSPA4L</i>	Body	South Shelf	-
cg26496372	5	37379396	<i>WDR70</i>	TSS200	Island	+
cg15289190	6	28831544			North Shore	+
cg06745145	7	90664816	<i>CDK14</i>	Body		+
cg04678743	7	130353515	<i>TSGA13</i>	3'UTR	Island	+
cg01731811	7	157890171	<i>PTPRN2</i>	Body	North Shore	-
cg12126859	8	335281				+
cg06009497	8	37695050	<i>GPR124</i>	Body	North Shelf	-
cg22848598	8	38965026	<i>ADAM32</i>	TSS200	Island	-
cg09961689	8	144590068	<i>ZC3H3</i>	Body		-
cg14286514	9	32525315	<i>DDX58</i>	Body	North Shore	-
cg14677909	10	48807341	<i>PTPN20B</i>	Body		-
cg27190138	10	98479757	<i>PIK3AP1</i>	Body	Island	+

cg15275103	10	124893024			Island	-
cg14677612	10	131263962	<i>MGMT</i>	TSS1500	North Shore	+
cg18308755	10	134065956	<i>STK32C</i>	Body		-
cg01981334	11	64877237	<i>C11orf2</i>	Body	North Shore	-
cg04293602	11	65553660	<i>OVOL1</i>	TSS1500	North Shore	-
cg15233880	11	69454727	<i>CCND1</i>	TSS1500	Island	-
cg22049858	11	94884121			Island	+
cg02574894	12	53693825	<i>C12orf10</i>	Body	Island	+
cg11315081	13	22651243				-
cg02583546	14	77494451	<i>C14orf4</i>	5'UTR	Island	+
cg14209037	15	41228521	<i>DLL4</i>	Body	Island	-
cg01003448	16	745685	<i>FBXL16</i>	Body	Island	+
cg07482202	16	745687	<i>FBXL16</i>	Body	Island	+
cg16619935	16	2037439	<i>GFER</i>	3'UTR	North Shelf	-
cg27605307	16	20357506	<i>UMOD</i>	Body	North Shelf	-
cg24303478	16	89143845			Island	-
cg25170034	17	33288066	<i>ZNF830</i>	TSS1500	North Shore	-
cg02216727	17	38520653	<i>GJD3</i>	1stExon	South Shore	-
cg18576374	17	78549371	<i>RPTOR</i>	Body		+
cg07356415	19	19655352	<i>CILP2</i>	Body	Island	-
cg12497870	19	36210913	<i>MLL4</i>	Body	Island	-
cg12088773	19	44128330	<i>CADM4</i>	Body		-
cg14083397	20	388473	<i>RBCK1</i>	TSS1500	Island	+
cg10629004	20	21696467	<i>PAX1</i>	3'UTR	South Shore	-
cg09576415	20	62059559	<i>KCNQ2</i>	Body	Island	-
cg12826791	21	45926719	<i>C21orf29</i>	Body	Island	-
cg00256932	22	51041732	<i>MAPK8IP2</i>	1stExon	North Shore	-

Table 6.12. Model results for probes that were FDR significant ($q < 0.05$) in both the combined air and diet PAH8 exposure EWAS model and the dietary PAH8 exposure only EWAS model. All results are from beta regression models run on the training dataset and were adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre.

Probe ID	Combined Air and Diet PAH8 Exposure			Dietary PAH8 Exposure		
	<u>B Coefficient</u>	<u>95% Confidence Interval</u>	<u>P Value</u>	<u>B Coefficient</u>	<u>95% Confidence Interval</u>	<u>P Value</u>
cg27158340	-0.024	-0.033; -0.015	4.57E-08	-0.00020	-0.00028; -0.00011	5.34E-06
cg09214099	-0.026	-0.036; -0.016	1.01E-07	-0.00021	-0.0003; -0.00011	1.61E-05
cg06182121	0.02	0.012; 0.028	4.26E-07	0.00020	0.00012; 0.00027	3.32E-07
cg03246584	0.11	0.066; 0.154	9.78E-07	0.00108	0.00067; 0.0015	2.79E-07
cg25930644	0.021	0.013; 0.03	1.10E-06	0.00021	0.00012; 0.00029	1.52E-06
cg12550399	0.02	0.012; 0.028	1.36E-06	0.00023	0.00015; 0.00031	6.10E-09
cg26780022	-0.041	-0.057; -0.024	1.83E-06	-0.00040	-0.00057; -0.00024	1.30E-06
cg08073527	0.031	0.018; 0.043	1.85E-06	0.00031	0.00019; 0.00043	7.94E-07
cg18827332	-0.032	-0.045; -0.018	2.26E-06	-0.00031	-0.00043; -0.00018	1.57E-06
cg15706250	-0.026	-0.037; -0.015	2.35E-06	-0.00024	-0.00035; -0.00014	7.21E-06
cg05262877	-0.022	-0.032; -0.013	2.59E-06	-0.00020	-0.0003; -0.00011	1.11E-05
cg12610917	-0.039	-0.055; -0.023	2.67E-06	-0.00035	-0.00051; -0.00019	1.59E-05
cg12187586	-0.047	-0.067; -0.027	3.63E-06	-0.00045	-0.00065; -0.00025	8.28E-06
cg21548131	0.063	0.036; 0.09	3.89E-06	0.00060	0.00034; 0.00086	7.67E-06
cg23759710	-0.013	-0.018; -0.007	4.10E-06	-0.00012	-0.00017; -0.00007	1.07E-05
cg05881436	-0.031	-0.044; -0.018	4.21E-06	-0.00030	-0.00042; -0.00017	4.87E-06
cg00086493	-0.023	-0.033; -0.013	5.82E-06	-0.00023	-0.00033; -0.00013	2.57E-06
cg16548154	-0.014	-0.021; -0.008	6.30E-06	-0.00015	-0.00021; -0.00009	1.85E-06
cg24413662	-0.033	-0.047; -0.019	7.04E-06	-0.00031	-0.00046; -0.00017	1.47E-05
cg14307471	0.043	0.024; 0.062	1.02E-05	0.00044	0.00025; 0.00063	4.35E-06
cg26913155	0.021	0.012; 0.031	1.31E-05	0.00021	0.00012; 0.00031	7.29E-06
cg14027524	-0.019	-0.027; -0.01	1.40E-05	-0.00022	-0.0003; -0.00014	1.52E-07
cg17583504	-0.031	-0.046; -0.017	1.60E-05	-0.00031	-0.00045; -0.00017	1.20E-05

cg18936620	0.016	0.009; 0.023	1.94E-05	0.00019	0.00012; 0.00026	1.44E-07
cg05419385	0.018	0.01; 0.027	2.05E-05	0.00022	0.00014; 0.0003	1.20E-07
cg00910067	-0.028	-0.041; - 0.015	2.42E-05	-0.00030	-0.00042; - 0.00017	3.59E-06
cg04351156	-0.016	-0.024; - 0.009	2.67E-05	-0.00017	-0.00025; - 0.0001	4.71E-06
cg20585869	-0.06	-0.088; - 0.032	2.81E-05	-0.00059	-0.00087; - 0.00032	2.29E-05
cg19312314	0.125	0.066; 0.183	2.88E-05	0.00125	0.00067; 0.00182	2.03E-05
cg14494090	-0.02	-0.029; - 0.01	3.17E-05	-0.00020	-0.00029; - 0.00011	1.98E-05
cg00686197	-0.017	-0.025; - 0.009	3.35E-05	-0.00018	-0.00026; - 0.0001	8.39E-06
cg15659420	0.029	0.015; 0.043	3.44E-05	0.00030	0.00016; 0.00044	1.48E-05
cg19697911	-0.023	-0.034; - 0.012	3.47E-05	-0.00026	-0.00037; - 0.00015	2.04E-06
cg24937768	-0.016	-0.024; - 0.008	3.67E-05	-0.00016	-0.00024; - 0.00009	2.24E-05
cg03308706	-0.057	-0.084; - 0.03	3.68E-05	-0.00060	-0.00088; - 0.00033	1.27E-05

Table 6.13. Table of characteristics of probes found to be significantly associated with combined air and dietary PAH8 exposure and dietary PAH8 exposure only at the FDR significance levels ($q < 0.05$).

Probe ID	Chromosome	Position	UCSC RefGene Name	Gene Location	Relation to CpG Island	Methylation Change Direction
cg24937768	1	2092853	<i>PRKCZ</i>	Body		-
cg06182121	1	3080723	<i>PRDM16</i>	Body	North Shore	+
cg26913155	1	3128175	<i>PRDM16</i>	Body		+
cg05262877	1	42631835	<i>GUCA2A</i>	TSS1500		-
cg18936620	1	43811019	<i>MPL</i>	Body	North Shelf	+
cg23759710	2	42990957	<i>OXER1</i>	1stExon		-
cg19697911	2	241080057	<i>OTOS</i>	5'UTR		-
cg21548131	3	173639566	<i>NLGN1</i>	Body		+
cg00686197	6	31733619	<i>C6orf27</i>	Body		-
cg09214099	7	72791740			Island	-
cg03308706	7	91763433	<i>CYP51A1</i>	5'UTR	Island	-
cg20585869	8	24772333	<i>NEFM</i>	TSS200	Island	-
cg15706250	8	41583321	<i>ANK1</i>	Body	Island	-
cg14027524	9	140120587	<i>C9orf169</i>	3'UTR	South Shelf	-
cg03246584	10	134663467				+
cg14494090	10	134972969	<i>KNDC1</i>	TSS1500	North Shore	-
cg15659420	11	20034979	<i>NAV2</i>	Body		+
cg24413662	11	122311293				-
cg05419385	12	27352945				+
cg18827332	12	103344506			Island	-
cg05881436	14	62331619			Island	-
cg27158340	14	105603389			Island	-
cg26780022	16	1336537				-
cg12187586	17	2627661			Island	-
cg25930644	17	8531915	<i>MYH10</i>	5'UTR	North Shore	+
cg12550399	17	19482275	<i>SLC47A1</i>	3'UTR	North Shore	+
cg16548154	17	74565757	<i>ST6GALNAC2</i>	Body		-
cg14307471	18	31432117	<i>NOL4</i>	3'UTR		+
cg04351156	19	10562415	<i>PDE4A</i>	Body		-
cg00910067	19	33717545	<i>SLC7A10</i>	TSS1500	Island	-
cg12610917	19	46387992	<i>IRF2BP1</i>	1stExon	Island	-

cg17583504	19	49669542	<i>TRPM4</i>	Body	Island	-
cg00086493	19	51535348	<i>KLK12</i>	Body	Island	-
cg08073527	21	43256581	<i>PRDM15</i>	Body	South Shore	+
cg19312314	21	44473962	<i>CBS</i>	3'UTR	Island	+

6.3 Discussion

In this chapter, the combined air and dietary PAH8 exposures were calculated for a total of 833 subjects in three datasets: Training (N = 493 subjects from EPIC-Italy cohort), Testing (N = 208 subjects from EPIC-Italy cohort), and EPIC-NL (N = 132) datasets. The combined exposure was calculated by converting the air exposures and dietary exposures from the previous two chapters to Z-scores, and then the air Z-scores were added to the dietary Z-scores for each individual. The Z-scores of combined air and dietary PAH8 exposures were used to carry out an EWAS which found 274 CpG probes associated with exposure to PAHs. The genomic distribution of these probes followed the expected pattern for all genomic regions with the exception of promoter regions where less changes than expected based on the underlying distribution of probes tested, and exon regions where more changes than expected were observed. A methylation index was developed using the probes, and as expected, the performance of the index was very good in the Training dataset in which the model was developed, but model performance declined in the Testing dataset. Comparison of the results from this chapter to those in the preceding two chapters showed that 58 CpG probes were also associated with air PAH8 exposure and 35 probes were also associated with dietary PAH8 exposure. Interestingly, these overlaps were not only the probes with the largest effect sizes (β -coefficients). In all three sets of EWAS results, significantly less changes than expected occurred at promoter regions. Taken together, the results show that combined air and PAH8 exposure may induce changes in DNA methylation.

6.3.1 Using Z-scores to Represent Combined Air and Dietary PAH8 Exposure

Z-scores were used as a proxy for combined air and dietary PAH8 exposure because the air exposure calculated in Chapter 4, and the dietary exposure calculated in Chapter 5 could not be combined due to having different units. The dietary exposure of PAH8 was in ng of PAH8 ingested/day, while air PAH8 exposure was in ng of PAH8 /m³ of air inhaled. To combine the two exposures, the mean

respiration volume per day for all subjects would be required, then this could be multiplied by the air PAH8 exposure to determine the ng of PAH8 inhaled/ day.

Because Z-scores were used, the model β -coefficients are not interpretable as in previous chapters. Z-scores represent the number of standard deviations away from the mean a value is, with positive and negative Z-scores indicating values above and below the mean respectively, and a Z-score close to 0 indicating a value close to the mean. Using the method described above to combine air and dietary PAH8 exposures would mean that the β -coefficients represent the % change in methylation per standard deviation of exposure.

Finally, smoking is an important source of exposure to PAHs, second only to diet, and PAH8 exposure from smoking has not been included in this chapter. A large number of EWAS looking at smoking have been published in recent years, and another was considered to be outside the scope of this project. Tobacco smoke contains a number of carcinogens in addition to PAHs, therefore, to fit in with the models described here, the quantity of PAHs, specifically PAH8, inhaled per cigarette smoked would be required and this could be multiplied by the number of cigarettes smoked per day. Some studies have looked at the amount of PAHs present in various cigarette brands, however cohort studies do not always collect detailed smoking information like brand, strength and filter. This means that any estimates of smoking PAH8 exposure would be open to misclassification errors, in the same way as the air PAH8 exposures, and dietary PAH8 exposures used here as discussed in Chapters 4 and 5.

6.3.2 Comparison of EWAS Results to Previously Published Findings

Of the 274 FDR-significant CpG sites found to be differentially methylated in association with combined air and dietary PAH8 exposure, none of the probes at the CpG level have been previously reported to be associated specifically with PAH exposure. One of the CpG probes (cg14677612) located in the promoter of the *MGMT* gene was found to be hypermethylated in the results presented above, and hypomethylation of this gene has previously been reported in association with urinary PAH metabolites in coke-oven workers³⁴³ and diesel engine exhaust particle exposed

workers³⁴⁹. None of the CpG probes reported above or the genes in which they are located have been reported to be associated with other known air pollutants. No studies have been published to date investigating the associations between DNA methylation and dietary intake of PAHs in humans.

Over 6500 differentially methylated regions were identified in human liver cells (HepaRG cell line) exposed to B[a]P³²⁵. Of the differentially methylated genes reported in this study by Tryndyak *et al.* (2018)³²⁵, 6 were also identified in the combined PAH8 exposure EWAS. Despite the overlaps, the location of the differentially methylated sites and the direction of change were not consistent between the two studies. The results are shown in Appendix 4 Table 9.17.

One of the CpG probes identified to be significantly associated with both combined air and dietary PAH8 exposure, and air PAH exposure as reported in Chapter 4, has also been reported to be associated with smoking in an EWAS carried out by Besingi *et al.* (2014)³⁸⁰. In all instances, probe cg02583546 located in the promoter region of C14orf4 was found to be hypermethylated. Four other CpG probes were found to be differentially methylated in association with both smoking and combined PAH8 exposure: cg04042861 (*HTR2B* promoter) was reported to be hypomethylated by Joehanes *et al.* (2016)⁴¹⁵ and in the current chapter, and probes cg14875327 (open sea), cg26590603 (*C6orf154* promoter), cg23432930 (*CHFR* promoter) were all found to be hypomethylated in the EWAS reported above, and the latter was also reported to be hypomethylated by Joubert *et al.* (2016)⁴¹⁶, but the former two were reported to be hypermethylated in the same study. While no other specific CpG probes identified in the EWAS presented here have been previously associated with smoking, an important source of PAH exposure, 85 of the genes in which significant CpG probes were located did overlap. These findings are summarised in Appendix 4 in Table 9.18.

As discussed in previous chapters, the Comparative Toxicogenomics Database reports 12,723 unique gene interactions associated with at least one of the PAHs that make up PAH8: B[a]A, B[b]Fl, B[k]Fl, B[ghi]P, B[a]P, Chr, DB[a,h]A, and I[cd]P¹⁸⁸. Of the genes associated with the 274 FDR-significant CpG sites identified from the EWAS (N = 208 genes), 116 had unique gene interactions listed in the

Comparative Toxicogenomics Database associated with one or more of the PAH8. It is important to note that many of these studies report findings from animal models or human cell lines, and DNA methylation patterns tend to tissue-specific, with blood being the source tissue of the DNA methylation analysed above. The human genome contains between 19,000 – 20,000 genes, meaning that the CTD has reports of PAHs interacting with over half the genome (approximately 63%), and the gene overlaps found here are about 42% (116 of 208 genes). This may suggest that the overlaps identified are possibly due to chance.

6.3.3 Statistical and Other Considerations

In the discussion of Chapter 4, several points were made about the EWAS results which were also relevant to the chapter on dietary PAH8 exposure (Chapter 5) and the current chapter on combined air and dietary PAH8 exposure. These will not be discussed in detail here to avoid repetition, but a summary of the relevant discussion points as they pertain to all relevant chapters will be included in the next chapter. These points include the tendency towards bias and inflation of test statistics in EWAS, the trade-off between using P-values over effect sizes to determine the associations of interest, the lack of statistical power due to low number of subjects, the lack of replication across the three datasets used, underlying population differences or measurement errors, residual confounding, and the probable over-fitting of the methylation index.

6.3.4 Conclusions

In this chapter, 274 CpG sites were found to be differentially methylated as a consequence of combined air and dietary PAH8 exposure. As expected, a number of these CpG probes overlapped with the findings from the EWASs on air and dietary PAH8 exposure separately. The methylation index developed using these results performed poorly in the Testing dataset, and in general, the results did not validate well across the three datasets analysed. The reasons for this need to be investigated further but some possible explanations are underlying methylation differences in the population that

are not related to PAH8 exposure, lack of statistical power, residual confounding, and model overfitting.

7 Chapter 7 – General Discussion, Future Work, and Conclusions

7.1 General Discussion

The overarching aims of this project were two-fold: the first was to measure DNA methylation changes in mice exposed to different doses of B[a]P to identify any dose-dependent changes and also link the observed changes to the gene expression data that was also available for the same mice. The second aim was to identify and understand the effects of environmentally-relevant exposures to PAHs on DNA methylation in humans. This was done by assessing the impacts of air inhalation and dietary PAH8 exposures separately, and as a single combined exposure.

7.1.1 Influence of PAH Exposure on DNA Methylation in Animals and Humans

Both the mouse and human studies found that DNA methylation was altered as a consequence of B[a]P and PAH8 exposure respectively. One finding was consistent across all human models and the window analysis in mice: when compared to distribution of 500 b.p windows, and probes located in promoter regions of the RRBS data and the Illumina Infinium HumanMethylation450 BeadChip array, the number of methylation changes significantly associated with B[a]P and PAH exposure located in promoter regions was significantly lower than expected by chance. In both the sites and windows analyses in mice, significantly more changes occurred at intron and intergenic regions than expected, although this was not observed in the human analyses, possibly due to the relative under-representation of these regions on the Illumina Infinium HumanMethylation450 BeadChip array compared to the RRBS data. The overlap between differentially methylated genes in both mice and humans was very low, with only a handful of genes occurring across analyses. There are several possible reasons for this: the difference in exposure dose between the mice and human subjects resulted in different genes being affected; the small methylation changes observed in the human EWAS are more difficult to validate than larger changes; the differences between mouse and human genomes result in different responses to exposure; and the effects of B[a]P-only exposure compared to a mixture (PAH8) may have different epigenetic responses in the same way that gene expression

changes have previously been reported to be different for individual PAH compounds compared to mixtures.

Another possible explanation is the one offered in the discussion of Chapter 3 which proposed that PAH-induced DNA methylation changes are dependent on the sites of PAH-DNA adduct formation which preferentially form at guanine bases adjacent to methylated cytosines, at least for BPDE-DNA adducts. This would suggest that adduct formation, and consequently, DNA methylation are driven by the genomic landscape rather than preference for particular genes. This would also go some way to explain why more methylation changes than expected occur at intergenic and intronic regions, while less changes than expected occurred at promoter regions since it is more likely that DNA adducts would be repaired at important regulatory regions like promoters compared to intergenic regions. Further data would be required to support this hypothesis which is discussed further in section 7.3 below.

7.1.2 Effect of Route of PAH Exposure on DNA Methylation

Investigating the potential differing effects of two major sources of PAH8 exposure, air and the diet, was one of the overarching aims of this project. This was done by using LUR models to estimate air PAH8 exposure, and using FFQ data in conjunction with a dataset synthesised from published concentrations of PAHs in food for dietary PAH8 exposure. These methods are discussed in a subsequent section, here the results from the two chapters and those from the chapter on the combined air and dietary PAH8 exposure are compared. There were no overlaps between the air and dietary PAH8 exposure EWAS results at the probe level, and only one gene was common to both, *GPR77* which is known to play a role in the complement system of the innate immune response, for which one probe in each set of results was found to be hypomethylated. This distinctiveness in the results could suggest that the consequences of air and dietary PAH8 exposure have different effects on DNA methylation in blood. This is further supported by the results from the EWAS of Z-score of combined air and dietary PAH8 exposure where the air and dietary exposures were combined by

converting each to Z-scores and then adding the two Z-scores for each subject. The differentially methylated probes identified in this EWAS included a subset that were identified in the air PAH8 exposure EWAS, a subset of probes identified in the dietary PAH8 exposure EWAS, and a third set of probes that did not overlap with either set of results. Taken together, the results do suggest that the effects of PAH exposure on DNA methylation may be different depending on the route of exposure. This is consistent with findings from previous studies which suggest that air exposure to PAHs includes inhalation of PAHs adsorbed on particles which increases pulmonary retention time and may also have consequences on the resulting metabolite pattern and the downstream metabolite-adduct formation^{44,45}. Additionally, it has been shown that the bioavailability of inhaled PAHs is higher than of those ingested through food²¹ which could account for the smaller effect sizes observed in the dietary PAH8 exposure EWAS compared to the air PAH8 exposure EWAS. However, they also suggest that the combination of exposures, possibly due to the different composition of the underlying mixtures, is associated with additional methylation changes. Further work would be required to confirm this hypothesis, and these analyses should also include PAH exposure from tobacco smoke as described in the future work section below.

7.2 Strengths and Limitations

This thesis reports novel findings from analyses that have not been previously conducted. No epigenome-wide studies have been previously published in mouse models that have been exposed to B[a]P or any other PAH. Additionally, as highlighted in the relevant chapters, while some human studies relating DNA methylation and PAH exposure have been published, these were focussed on either subjects known to have high exposure through their occupation for example as shown in the introduction of Chapter 4, or did not look at the epigenome-wide effects of exposure in adults.

7.2.1 Power and Other Statistical Considerations

The lack of statistical power due to the small sample sizes of both the mouse and human studies is one of the main limitations of the studies presented here. Mouse studies always tend to use as few

animals as possible due to the feasibility and ethical considerations around conducting studies with large numbers of mice. In human studies, it is common knowledge that in comparison to genome-wide association studies (GWAS) which often include tens of thousands of subjects, EWAS studies tend to be under-powered. Large, collaborative DNA methylation studies have been published for BMI⁴⁵¹ and alcohol consumption⁴⁵², however these covariates are routinely collected as part of prospective cohorts and so in these instances the limiting factor tends to be the number of subjects for which DNA methylation is measured. Exposures such as PAH exposure are much more difficult to capture in prospective studies as personal monitoring and duplicate plate methods, for example, are costly to apply to a large number of subjects. Prospective cohorts are set up to capture as many variables as possible, however the traffic and related variables required for many LUR models developed to estimate PAH exposure are specific and not routinely collected limiting the number of subjects that can be included in such studies. Moreover, as discussed above and in previous chapters, proxies such as LUR models and FFQs are subject to misclassification errors. Methods for assessing PAH exposure are discussed in more detail in a subsequent section, however this is directly linked to the issue of statistical power. As previously mentioned, the Training (N = 493) and Testing (N = 208) datasets which were both subsets of the EPIC-Italy cohort could have been maintained as a single dataset to maximise power. However, this would have limited the potential for validation in an independent dataset as the EPIC-NL dataset had too many underlying differences to reasonably expect results to validate.

Despite the lack of statistical power, several DMWs, DMCs, and CpG probes were identified to be associated with B[a]P and PAH8 exposures across the multiple studies carried out during this project. The mouse RRBS data were interrogated using the methylkit R package¹⁸⁷, the output of which does not provide the underlying test statistics, and so inflation of the test statistics could not be assessed. This package was used as it has been designed to handle data generated from RRBS and other similar methods. For the human studies, beta regression was used which is a method designed for handling rates and proportions which does not have the same assumptions of normally distributed

homoscedastic data as linear regression ²⁸⁶. A more detailed discussion of this can be found in Chapter 2. Of studies which have been published so far that have employed beta regression ^{191,192,287,288}, none have reported the results of any tests for inflation of test statistics. For the human studies in this thesis, inflation of the test statistics was carried out using the 'bacon' R package ²⁹². The results from the air PAH8 exposure EWAS had the lowest level of inflation ($\lambda = 1.14$) and the dietary and combined PAH8 exposure EWASs having similar levels of inflation (dietary PAH8 exposure EWAS: $\lambda = 1.19$; combined PAH8 exposure EWAS: $\lambda = 1.18$). This shows that some inflation did occur and might suggest residual confounding. As mentioned previously, although not shown in this thesis, early analyses carried out during the course of this project showed that inflation tended to increase when additional covariates were added to the model. Further investigation into the influence of inflation and the correct methods by which to measure this are required as currently no comparisons can be made.

Inflation of the test statistics and, by extension p values, is one reason why p values should be used with caution when filtering EWAS results to find the most interesting results. Currently, the most widely employed and accepted method is correction for multiple testing either using FDR or Bonferroni methods. While these methods do provide a strict threshold, if the p values themselves are inflated, then the subset of probes that pass this threshold may still include a number of false-positives. An alternative may be to use a combination of effect size and statistical significance when identifying probes of interest. This is how the DMWs and DMCs from the mouse RRBS study were identified. Knowing that the statistical power of the models was low, with 9 vs 3 mice in the treated vs untreated model, and 3 vs 3 mice in each of control vs dose models, regions and sites of interest were selected based on $p < 0.05$ and methylation differences of $> 25\%$ in either direction. Such methods, however, will need to be investigated further in studies which are more highly powered.

The results of the inflation tests may indeed suggest residual confounding however, this could not be confirmed. For the human studies, the models were defined *a priori* using covariates that have been

well-established to be associated with DNA methylation in the literature, such as WBC distributions, smoking status, sex, and the technical covariates array chip and position on array chip. Other covariates such as cancer case status were included because work done by colleagues on the EPIC-Italy cohort suggested that this needed to be adjusted for ¹⁹¹. Finally, due to the cohort sizes and the large number of array chips and the twelve positions on the array chip which has to be included to account for batch effects, limited covariates could be included in the model. In fact, array chip was not included in the models run on the Testing and EPIC-NL datasets due to the large number of unique array chips which would have resulted in significant over-fitting. If larger cohorts were available, more rigorous model building techniques could be applied to minimise the effects of residual confounding. In EWAS studies, similar to GWAS studies, one of the primary sources of confounding is technical confounding which has been accounted for as far as statistically possible in the models used in the studies presented in this thesis. However, there are additional factors that confound EWAS studies which do not apply to GWAS. Given the extent of the effect of environmental factors on DNA methylation, these factors introduce an additional level of complexity in EWAS studies. Such environmental confounders have been reported to inflate type I errors, and therefore the effect sizes reported in such studies (reference to be included in clean version). Confounding due to environmental factors increases the complexity of EWAS studies as the effects of the wide range of such exposures should be taken into account when building EWAS models. However, this is further complicated by EWAS studies being generally under-powered compared to GWAS studies, which inherently limits the number of confounders that can be reasonably considered.

In each of the three human EWAS chapters, the results were used to build a methylation index of PAH exposure as an alternative validation method. The hypothesis was that, if the CpG sites identified by the EWAS were indeed associated with PAH exposure, then these same sites should be able to reasonably predict PAH exposure. Additionally, if these sites were generally associated with PAH exposure, i.e. independently of the dataset in which they were identified, then the methylation index should also be able to predict the PAH exposure of an independent cohort. This however, was found

not to be the case. Over-fitting is one of the most likely explanations and this was due to the number of CpG sites included in the model (Air PAH8 exposure: N = 204; Dietary PAH8 exposure: N = 97; Combined air and dietary PAH8 exposure: N = 274). The performance of the models in the Training dataset on which they were trained was very good, however this dramatically decreased when the model was applied to a new dataset (Testing dataset) which is often a sign of over-fitting the model to the data on which it was trained. To mitigate this, the number of probes included in the model could have been reduced through an additional filtering step, such as including only those probes with the largest effect sizes, however setting a threshold for this would be somewhat arbitrary. If more datasets with the required data were available, meta-analyses could be conducted to identify those changes that occur in multiple datasets to improve generalisability.

7.2.2 Technical Considerations and Underlying Differences in DNA Methylation

While the Illumina Infinium HumanMethylation450 BeadChip array and RRBS are two established methods by which DNA methylation is assessed and analysed, it is important to consider the coverage of these methods in relation to the human and mouse genomes. The Illumina Infinium HumanMethylation450 BeadChip array covers less than 2% of the CpG sites in the human genome, while the RRBS libraries prepared and sequenced in this PhD project covered between 5.5-7.3% of the CpG sites in the mouse genome. This is important to consider when interpreting the results presented here because the representation levels of different genomic regions is different between the two methods and in comparison to the genome. Interestingly, the genomic distributions of the CpG sites and 500 b.p. windows common to all mouse samples were also different, with the former showing more similarity to the distribution of all CpGs in the mouse genome. Whole genome sequencing would cover the majority of CpG sites of the genome however, the costs of this method make it prohibitive particularly for large human studies. The methods employed during this project were designed to be as representative as possible, however much information is missing.

A further experimental consideration is that the DNA extracted and used in both the animal and human studies originated from a heterogeneous cell population (lung tissue and blood respectively). This was accounted for in the human studies by adjusting the models for WBC proportions calculated using the Houseman method ²⁸⁵, but such methods have not yet been established for other tissues or methods that do not employ the Illumina methylation array. Such adjustments, while required, further contribute to the problem of statistical power in DNA methylation studies as discussed above.

In both the animal and human studies, significant differences in the underlying methylation patterns were observed. The PCA carried out on the RRBS data showed that the methylation differences between the mice were not necessarily due to B[a]P exposure, and the differences between the control mice were larger than some of the exposed mice. Some possible reasons for this were explored in Chapter 3. As shown in Chapter 4, there were statistically significant differences between the EPIC-Italy subjects (Training and Testing datasets) and the EPIC-NL subjects. There are number of reasons that could explain this such as population differences, lifestyle, or biological factors, however further investigation would be required to identify the correct explanation for both the mouse and human observations.

7.2.3 Methods for Assessing PAH Exposure

In the human studies, air PAH8 exposure was estimated using models developed for Rome in the ESCAPE study ²⁵, and dietary PAH8 exposures were estimated using a combination of FFQ results and the collated results of previously published studies measuring PAHs in food. While the limitations of these methods have been discussed in the relevant results chapters, here the considerations common to both methods will be emphasised. Both methods only provide a single snapshot of exposure, which does not account for seasonality, with the LUR calculating exposure at a single address, and the FFQ only accounting for reported diet over a short period of time. One of the major limitations of FFQs is recall bias which exacerbates the misclassification errors to which both FFQ and LUR methods are prone. Other, more accurate and reliable methods have been used in previous studies looking at air

and dietary PAH8 exposure: personal exposure monitoring equipment for the former and the duplicate plate method for the latter. Use of personal exposure monitoring equipment allows for the real-time assessment of PAHs present in the air that the subject is breathing in throughout the day, both inside and outside the home and/or work locations. However, this method has a few drawbacks which explain why it is not more commonly used. The first is that the number of subjects that can be included in such a study is limited due the cost of the equipment itself. This could be overcome if the study is conducted over time, however the time of year during which each subject participated would need to be accounted for. Additionally, the amount of data captured by personal air monitoring equipment over the course of the study period would require particular statistical methods in order for the time-series aspects to be included with the beta regression methods required for the EWAS. One method to overcome this is the use of silicone wristbands developed by a group at Oregon State University specifically to measure PAHs⁴⁵³. These wristbands are much cheaper than traditional methods, are non-intrusive to the subject making them easy to integrate into studies, and have been shown to have a better correlation with urinary hydroxy-PAH metabolites than personal air monitors⁴⁵³.

Cost is also a limiting factor for the duplicate plate method, and while this method accounts for cooking processes and differences in the PAH profiles of the constituent foods, it does not necessarily represent the food that the subject would eat “normally” and thus would still be subject to a degree of misclassification errors. Therefore, unless urinary or blood biomarkers are used, the results of the available methods for measuring human PAH exposure need to be interpreted with these limitations in mind. Some of these limitations may be overcome by transforming the continuous exposure variables to categorical variable however, this would greatly reduce the granularity of the data. The use of urinary and blood biomarkers are also flawed in that the inter-individual variation in the metabolism of PAHs is not accounted for. Additionally, urinary biomarkers have a relatively short life and therefore do not reflect bioaccumulation.

7.3 Future Work

The results reported in this thesis indicate that PAH exposure does affect DNA methylation, however, the downstream consequences of these changes is unclear. Additionally, the methods used in the human studies to measure exposure to PAHs can be improved to give more reliable results. Finally, further mouse experiments should be carried out that could help to further understand the effects of exposure source, the effects of other PAHs and groups of PAHs, tissue-specific changes, and the effects of exposure on other epigenetic mechanisms such as histone modifications and the chromatin landscape.

7.3.1 Animal Studies

Given the degree of variation in the DNA methylation profiles of the mice sequenced in this project which was not related to B[a]P exposure as observed in the controls, the first requirement would be to carry out the RRBS experiment again with deeper sequencing. This would also help to address the heterogeneity in the sequenced sites and increase overlap in common regions across the different samples. More mice per exposure dose should be sequenced in order to maximise statistical power as much as possible, however it is unusual for a large number of mice to be included in experiments, so power will always be a limiting factor when the effect sizes are small. While the design of the initial study for which the mice were used¹⁴² might have had sufficient power, it was not powered with the analyses presented here in mind. To maximise the potential of extrapolation from mouse to human studies, environmentally-relevant concentrations of B[a]P should be used, as the mouse B[a]P exposures used in this project were much higher. Moreover, mixtures of PAHs with similar compositions to those to which humans are exposure should be used as humans are never exposed to a single PAH compound. The effects of PAH mixtures compared to individual compounds was discussed in the introduction to this thesis, but briefly, the effects of mixtures can be more- or less-than-additive when considering the individual effects of the constituent compounds, however a recent study supports the additive effects of complex mixtures³⁶. The analysis of tissues in addition to

lung tissue from the same mice would allow for comparison of DNA methylation profiles across multiple tissues from both target and non-target organs which have already been shown to have different gene expression profiles in response to B[a]P exposure¹⁴³. Additionally, the analysis of DNA methylation in blood from these mice would allow for direct comparison with results from human studies where blood is often used as a surrogate tissue.

In addition to the DNA methylation experiments described above, other experiments using the same mice could be carried out which would allow for further understanding of the downstream consequences of the DNA methylation changes taking place. In this project, the relationship between gene expression and DNA methylation was investigated for those DNA methylation changes occurring at genic regions, although correlation was only observed for a subset of these genes. This would also need to be repeated by carrying out RNA sequencing. The effects of the changes happening at intergenic and intronic regions which were enriched for methylation changes however, are yet to be understood, with potential explanations being altered chromatin structure or regulation of enhancers. Additionally, the relationship between DNA methylation and PAH-DNA adducts needs to be further understood. In this thesis, this relationship could not be interrogated as only the absolute number of BPDE-DNA adducts for the mouse samples was available. Members of our lab have been working on a sequencing-based technique to identify the exact genomic loci at which cisplatin DNA adducts form (Gallon *et al.*, unpublished data). Such sequencing-based techniques would allow for the mapping of the DNA adducts which, in combination with RRBS data, would allow for the confirmation of the hypotheses that BPDE-DNA adducts preferentially form at guanines adjacent to methylated cytosines. Furthermore, a direct comparison of the two sets of results would go a way to understand the relationship between methylation changes and DNA adduct formation, however, understanding whether DNA adducts influence DNA methylation, or vice versa would require a time series experiment using cell lines. To shed more light on the genomic landscape, histone modifications could be mapped using chromatin immunoprecipitation with sequencing (ChIP-seq) to further increase understanding of the effects of PAH exposure on chromatin structure. Finally, the assay for

transposase-accessible chromatin using sequencing (ATAC-seq) would identify whether the chromatin structure of exposed vs non-exposed mice is different and could suggest whether any observed changes in DNA methylation and histone modifications lead to these changes to the chromatin landscape.

The large number of experiments described above highlight how much more work is required to fully understand the effects of PAH exposure on epigenetic mechanisms and the consequences of any modifications on gene expression or the genomic landscape. Such work would help to further understand the complex consequences of PAH exposure and the relationships between them. In turn, this increased understanding would help to support observations made in human epidemiological studies where such analyses may be difficult to complete.

7.3.2 Human Epidemiological Studies

The biggest limitations of the human studies carried out in this project are low statistical power, the methods employed to determine PAH8 exposure, and the lack of validation of observed results. Collaborative studies using cohorts where PAH exposure has already been measured or where the necessary variables and/or LUR models along with FFQ data are required in order to carry out sufficiently-powered studies. One important aspect of such a collaboration is that a subset of the included subjects would have air exposure measured using personal air monitoring equipment, and dietary exposure measured using the duplicate plate method, in addition to the LUR model and FFQ exposure estimates to assess the correlation between these different measures and perhaps develop a penalty that can be applied to account for the variation between the exposure assessment methods, and this is known as calibration. In this way, two of the limitations of the studies presented in this thesis could be addressed directly. By increasing cohort size, the range and distribution of PAH exposure across the subjects can also be improved which together may help to improve validation of results in independent cohorts. Future analyses should also take PAH exposure from smoking into account. Smoking status is routinely collected in prospective cohorts, and in some instances

biomarkers of smoking such as urinary cotinine levels are also available. By including all three major sources of PAH exposure, their individual and combined effects could be analysed, and the observations made in this thesis could be validated. The EWAS results should be validated in a subset of the subjects included in the EWAS using an alternative method to ensure that the differences identified are genuine and not technical or statistical artefacts. Additionally, the top probes should be validated in a least one independent cohort which would support the generalisability of the results. Pyrosequencing is one method by which this validation could be carried out, however this may not be the most cost- or time-efficient method for large numbers of probes and samples. Targeted sequencing based methods such as the Fluidigm 48.48 Access Array would ease the burden of validating results in independent samples. Several of the experiments described in the previous section on future work in the mouse model would contribute much knowledge if these could also be applied in at least a small number of human subjects if only to validate that the observations in mice still hold in humans.

7.4 Final Conclusions

In this thesis, the effects of PAH exposure on DNA methylation were analysed in mice and humans. Several associations were identified, with less changes than expected occurring at promoter regions in both mice and humans in response to B[a]P and PAH8 exposures respectively. The results of the human studies suggest that air and dietary PAH8 exposures induce separate DNA methylation responses, and when these exposures are combined, the methylation changes observed represent both the separate exposures. While the site- and gene-specific changes require further validation, the results presented here suggest that PAH-induced DNA methylation changes may occur as a consequence of PAH-DNA adduct formation, which do not necessarily occur in a gene-specific manner. The downstream consequences of these DNA methylation changes need to be investigated in future studies. The work presented in this thesis will aim to be published as two separate publications: the first will cover the animal study, and the second will cover the human studies.

8 References

1. Wenzl, T., Simon, R., Kleiner, J. & Anklam, E. Analytical methods for polycyclic aromatic hydrocarbons (PAHs) in food and the environment needed for new food legislation in the European Union. *Trends Anal. Chem.* **25**, 716–725 (2006).
2. Ravindra, K., Sokhi, R. & Van Grieken, R. Atmospheric polycyclic aromatic hydrocarbons: Source attribution, emission factors and regulation. *Atmos. Environ.* **42**, 2895–2921 (2008).
3. Abdel-Shafy, H. I. & Mansour, M. S. M. A review on polycyclic aromatic hydrocarbons: Source, environmental impact, effect on human health and remediation. *Egypt. J. Pet.* **25**, 107–123 (2016).
4. Jarvis, I. *et al.* Persistent activation of DNA damage signaling in response to complex mixtures of PAHs in air particulate matter. *Toxicol. Appl. Pharmacol.* **266**, 408–418 (2013).
5. Bai, H. & Zhang, H. Characteristics, sources, and cytotoxicity of atmospheric polycyclic aromatic hydrocarbons in urban roadside areas of Hangzhou, China. *J Environ. Sci. Heal. Part A* **0**, 1–10 (2016).
6. World Health Organization, W. Air quality guidelines for Europe. *WHO Reg. Publ. Eur. Ser. No. 91* 1–288 (2000). doi:10.1007/BF02986808
7. Srogi, K. Monitoring of environmental exposure to polycyclic aromatic hydrocarbons: a review. *Environ. Chem. Lett.* **5**, 169–195 (2007).
8. Achten, C. & Andersson, J. T. Overview of Polycyclic Aromatic Compounds (PAC). *Polycycl. Aromat. Compd.* **35**, 177–186 (2015).
9. Moorthy, B., Chu, C. & Carlin, D. J. Polycyclic Aromatic Hydrocarbons: From Metabolism to Lung Cancer. *Toxicol. Sci.* **145**, 5–15 (2015).
10. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Chemical agents and related occupations. *IARC Monogr. Eval. Carcinog. Risks Hum.* **100**, 9–562 (2012).
11. IARC. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Some Non-heterocyclic Polycyclic Aromatic Hydrocarbons and Some Related Exposures. *Iarc Monogr. Eval. Carcinog. Risks To Humans* **92**, 1–868 (2010).
12. Ewa, B. & Danuta, M.-Ś. Polycyclic aromatic hydrocarbons and PAH-related DNA adducts. *J. Appl. Genet.* **58**, 321–330 (2017).
13. Xue, W. & Warshawsky, D. Metabolic activation of polycyclic and heterocyclic aromatic hydrocarbons and DNA damage: A review. *Toxicol. Appl. Pharmacol.* **206**, 73–93 (2005).
14. Shen, H. *et al.* Global Atmospheric Emissions of Polycyclic Aromatic Hydrocarbons from 1960 to 2008 and Future Predictions. *Environ. Sci. Technol.* **47**, 6415–6424 (2013).
15. Prevedouros, K. *et al.* Seasonal and long-term trends in atmospheric PAH concentrations: evidence and implications. *Environ. Pollut.* **128**, 17–27 (2004).
16. Villar-Vidal, M. *et al.* Air Polycyclic Aromatic Hydrocarbons (PAHs) associated with PM2.5 in a North Cantabric coast urban environment. *Chemosphere* **99**, 233–8 (2014).
17. Bartrons, M., Catalan, J. & Penuelas, J. Spatial And Temporal Trends Of Organic Pollutants In Vegetation From Remote And Rural Areas. *Sci. Rep.* **6**, 25446 (2016).
18. Yu, Y. *et al.* Risk of human exposure to polycyclic aromatic hydrocarbons: A case study in

- Beijing, China. *Environ. Pollut.* **205**, 70–77 (2015).
19. Tarantini, A. *et al.* Relative contribution of DNA strand breaks and DNA adducts to the genotoxicity of benzo[a]pyrene as a pure compound and in complex mixtures. *Mutat. Res.* **671**, 67–75 (2009).
 20. Boström, C.-E. *et al.* Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. *Environ. Health Perspect.* **110 Suppl**, 451–88 (2002).
 21. Menzie, C. A., Potocki, B. B. & Santodonato, J. Exposure to carcinogenic PAHs in the environment. *Environ. Sci. Technol.* **26**, 1278–1284 (1992).
 22. Boström, C. *et al.* Cancer Risk Assessment, Indicators, and Guidelines for Polycyclic Aromatic Hydrocarbons in the Ambient Air. *Environ. Health Perspect.* **110**, 451–489 (2002).
 23. Nadeau, K. *et al.* Ambient air pollution impairs regulatory T-cell function in asthma. *J. Allergy Clin. Immunol.* **126**, 845–852.e10 (2010).
 24. Jedynska, A. *et al.* Spatial variations of PAH, hopanes/steranes and EC/OC concentrations within and between European study areas. *Atmos. Environ.* **87**, 239–248 (2014).
 25. Jedynska, A. *et al.* Development of Land Use Regression Models for Elemental, Organic Carbon, PAH, and Hopanes/Steranes in 10 ESCAPE/TRANSPHORM European Study Areas. *Environ. Sci. Technol.* **48**, 14435–44 (2014).
 26. Dat, N.-D. & Chang, M. B. Review on characteristics of PAHs in atmosphere, anthropogenic sources and control technologies. *Sci. Total Environ.* **609**, 682–693 (2017).
 27. Kieth, L. H. The Source of U.S. EPA's Sixteen PAH Priority Pollutants. *Polycycl. Aromat. Compd.* **35**, 147–160 (2015).
 28. Andersson, J. & Achten, C. Time to say goodbye to the 16 EPA PAHs? Toward an up-to-date use of PACs for environmental purposes. *Polycycl. Aromat. Compd.* **35**, 330–354 (2015).
 29. European Parliament & Council of the European Union. *Directive 2004/107/EC of the European Parliament and of the Council of 15/12/2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air.* 3–17 (2004).
 30. Baird, W. M., Hooven, L. A. & Mahadevan, B. Carcinogenic polycyclic aromatic hydrocarbon-DNA adducts and mechanism of action. *Environ. Mol. Mutagen.* **45**, 106–14 (2005).
 31. Jarvis, I. W. H., Dreij, K., Mattsson, Å., Jernström, B. & Stenius, U. Interactions between polycyclic aromatic hydrocarbons in complex mixtures and implications for cancer risk assessment. *Toxicology* **321**, 27–39 (2014).
 32. Bauer, A. K. *et al.* Environmentally prevalent polycyclic aromatic hydrocarbons can elicit co-carcinogenic properties in an in vitro murine lung epithelial cell model. *Arch. Toxicol.* **92**, 1311–1322 (2018).
 33. Tarantini, A. *et al.* Polycyclic aromatic hydrocarbons in binary mixtures modulate the efficiency of benzo[a]pyrene to form DNA adducts in human cells. *Toxicology* **279**, 36–44 (2011).
 34. Samburova, V., Zielinska, B. & Khlystov, A. Do 16 Polycyclic Aromatic Hydrocarbons Represent PAH Air Toxicity? *Toxics* **5**, 17 (2017).
 35. Tilton, S. C. *et al.* Mechanism-Based Classification of PAH Mixtures to Predict Carcinogenic Potential. *Toxicol. Sci.* **146**, 135–45 (2015).
 36. Long, A. S., Lemieux, C. L., Gagné, R., Lambert, I. B. & White, P. A. Genetic Toxicity of Complex

- Mixtures of Polycyclic Aromatic Hydrocarbons: Evaluating Dose-Additivity in a Transgenic Mouse Model. *Environ. Sci. Technol.* **51**, 8138–8148 (2017).
37. Coluci, V. R., Vendrame, R., Braga, R. S. & Galvão, D. S. Identifying Relevant Molecular Descriptors Related to Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons (PAHs) Using Pattern Recognition Methods. *J Chem Inf Comput Sci* **42**, 1479–1489 (2002).
 38. Butler, J. P., Post, G. B., Lioy, P. J., Waldman, J. M. & Greenberg, A. Assessment of Carcinogenic Risk from Personal Exposure to Benzo(a)pyrene in the Total Human Environmental Exposure Study (THEES). *Air Waste Air Waste Manag. Assoc* **43**, 970–977 (1993).
 39. Phillips, D. H. Polycyclic aromatic hydrocarbons in the diet. *Mutat. Res.* **443**, 139–147 (1999).
 40. Ding, Y. S., Ashley, D. L. & Watson, C. H. Determination of 10 Carcinogenic Polycyclic Aromatic Hydrocarbons in Mainstream Cigarette Smoke. *J. Agric. Food Chem.* **5**, 5966–5973 (2007).
 41. Vu, A. T. *et al.* Polycyclic Aromatic Hydrocarbons in the Mainstream Smoke of Popular U.S. Cigarettes. *Chem. Res. Toxicol.* **28**, 1616–26 (2015).
 42. International Agency for Research on Cancer. *Polynuclear aromatic compounds, part 1 : chemical, environmental and experimental data.* (International Agency for Research on Cancer, 1983).
 43. Weyand, E. H. & Bevan, D. R. Benzo(a)pyrene Disposition and Metabolism in Rats following Intratracheal Instillation. *Cancer Res.* **46**, 5655–5661 (1986).
 44. Sun, J. D. *et al.* Lung Retention and Metabolic Fate of Inhaled Benzo(a)pyrene Associated with Diesel Exhaust Particles. *Toxicol. Appl. Pharmacol.* **73**, 48–59 (1984).
 45. Tornquist, S. T., Wiklund, L. & Toftgard, R. Investigation of Absorption, Metabolism Kinetics and DNA-Binding of Intratracheally Administered Benzo[a]pyrene in the Isolated, Perfused Rat Lung: A Comparative Study Between Microcrystalline and Particulate Adsorbed Benzo[a]pyrene. *Chem.-Biol. Interact.* **54**, 185–198 (1985).
 46. Foth, H., Kahl, R. & Kahl, G. F. Pharmacokinetics of Low Doses of Benzo[a]pyrene in the Rat. *Fd Chem. Toxic* **26**, 45–51 (1988).
 47. Withey, J. R., Shedden, J., Law, F. C. P. & Abedini, S. Distribution of benzo[a]pyrene in pregnant rats following inhalation exposure and a comparison with similar data obtained with pyrene. *J. Appl. Toxicol.* **13**, 193–202 (1993).
 48. Neubert, D. & Tapken, S. Transfer of benzo(a)pyrene into mouse embryos and fetuses. *Arch Toxicol* **62**, 236–239 (1988).
 49. Lee, J. *et al.* Prenatal airborne polycyclic aromatic hydrocarbon exposure, LINE1 methylation and child development in a Chinese cohort. *Environ. Int.* **99**, 315–320 (2016).
 50. Van De Wiel, J. A. G. *et al.* Excretion of benzo[a]pyrene and metabolites in urine and feces of rats: influence of route of administration, sex and long-term ethanol treatment. *Toxicology* **80**, 103–115 (1993).
 51. Pott, P. Surgical Observations Relative to... Cancer of the Scrotum.. *CA. Cancer J. Clin.* **24**, 110–116 (1974).
 52. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
 53. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

54. Bai, H., Wu, M., Zhang, H. & Tang, G. Chronic polycyclic aromatic hydrocarbon exposure causes DNA damage and genomic instability in lung epithelial cells. *Oncotarget* **8**, 79034–79045 (2017).
55. Sartor, M. A. *et al.* Genomewide analysis of aryl hydrocarbon receptor binding targets reveals an extensive array of gene clusters that control morphogenetic and developmental programs. *Environ. Health Perspect.* **117**, 1139–46 (2009).
56. Syed, A., Hew, K., Kohli, A., Knowlton, G. & Nadeau, K. Air Pollution and Epigenetics: Recent Findings. *Curr. Environ. Heal. Reports* **1**, 35–45 (2014).
57. Kerley-Hamilton, J. S. *et al.* Inherent and benzo[a]pyrene-induced differential aryl hydrocarbon receptor signaling greatly affects life span, atherosclerosis, cardiac gene expression, and body and heart growth in mice. *Toxicol. Sci.* **126**, 391–404 (2012).
58. Chou, W.-C. *et al.* Development of an in Vitro-Based Risk Assessment Framework for Predicting Ambient Particulate Matter-Bound Polycyclic Aromatic Hydrocarbon-Activated Toxicity Pathways. *Env. Sci Technol* **51**, 14262–14272 (2017).
59. Whitlock, J. P. Induction of cytochrome P4501A1. *Annu. Rev. Pharmacol. Toxicol.* **39**, 103–25 (1999).
60. Shimada, T. Xenobiotic-metabolizing enzymes involved in activation and detoxification of carcinogenic polycyclic aromatic hydrocarbons. *Drug Metab. Pharmacokinet.* **21**, 257–276 (2006).
61. Tekpli, X. *et al.* DNA methylation of the CYP1A1 enhancer is associated with smoking-induced genetic alterations in human lung. *Int. J. Cancer* **131**, 1509–16 (2012).
62. Uppstad, H., Øvrebø, S., Haugen, A. & Møllerup, S. Importance of CYP1A1 and CYP1B1 in bioactivation of benzo[a]pyrene in human lung cell lines. *Toxicol. Lett.* **192**, 221–228 (2010).
63. Beland, F. A. & Poirier, M. C. in *Methods to Assess DNA Damage and Repair: Interspecies Comparisons* 29–55 (1994). doi:10.1046/j.1365-2648.1994.20050975-11.x
64. Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ. Mol. Mutagen.* **58**, 235–263 (2017).
65. Geacintov, N. E., Shahbaz, M., Ibanez, V., Moussaoui, K. & Harvey, R. G. Base-sequence dependence of noncovalent complex formation and reactivity of benzo[a]pyrenediol epoxide with polynucleotides. *Biochemistry* **27**, 8380–8387 (1988).
66. Käfferlein, H. U., Marczynski, B., Mensing, T., Brüning, T. & Käfferlein, H. U. Albumin and hemoglobin adducts of benzo[a]pyrene in humans-Analytical methods, exposure assessment, and recommendations for future directions Albumin and hemoglobin adducts of benzo[a]pyrene in humans-Analytical methods, exposure assessment, and recommend. *Crit. Rev. Toxicol.* **40**, 126–150 (2010).
67. Braithwaite, E., Wu, X. & Wang, Z. Repair of DNA lesions induced by polycyclic aromatic hydrocarbons in human cell-free extracts: involvement of two excision repair mechanisms in vitro. *Carcinogenesis* **19**, 1239–1246 (1998).
68. Briedé, J. J. *et al.* In vitro and in vivo studies on oxygen free radical and DNA adduct formation in rat lung and liver during benzo[a]pyrene metabolism. *Free Radic. Res.* **38**, 995–1002 (2004).
69. Yoon, J.-H. *et al.* Methylated CpG Dinucleotides Are the Preferential Targets for G-to-T Transversion Mutations Induced by Benzo[a]pyrene Diol Epoxide in Mammalian Cells: Similarities with the p53 Mutation Spectrum in Smoking-associated Lung Cancers. *Cancer Res.*

- 61, 7110–7117 (2001).
70. Severson, P. L., Vrba, L., Stampfer, M. R. & Futscher, B. W. Exome-wide mutation profile in benzo[a]pyrene-derived post-stasis and immortal human mammary epithelial cells. *Mutat. Res. Genet. Toxicol. Environ. Mutagen.* **775–776**, 48–54 (2014).
 71. Hainaut, P. & Pfeifer, G. P. Patterns of p53 G → T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. *Carcinogenesis* **22**, 367–374 (2001).
 72. Huan Dong *et al.* Mutagenic Potential of Benzo[a]pyrene-Derived DNA Adducts Positioned in Codon 273 of the Human P53 Gene. *Biochemistry* **43**, 15922–15928 (2004).
 73. Prahalad, A. *et al.* Dibenzo[a,l]pyrene-induced DNA adduction, tumorigenicity, and Ki-ras oncogene mutations in strain A/J mouse lung. *Carcinogenesis* **18**, 1955–1963 (1997).
 74. Hu, W., Feng, Z. & Tang, M.-S. Preferential Carcinogen-DNA Adduct Formation at Codons 12 and 14 in the Human K-ras Gene and Their Possible Mechanisms †. *Biochemistry* **42**, 10012–10023 (2003).
 75. Georgiadis, P. *et al.* Impact of phase I or phase II enzyme polymorphisms on lymphocyte DNA adducts in subjects exposed to urban air pollution and environmental tobacco smoke. *Toxicol. Lett.* **149**, 269–280 (2004).
 76. Mollerup, S., Ryberg, D., Hewer, A., Phillips, D. H. & Haugen, A. Sex differences in lung CYP1A1 expression and DNA adduct levels among lung cancer patients. *Cancer Res.* **59**, 3317–3320 (1999).
 77. Rihs, H.-P. *et al.* Modulation of urinary polycyclic aromatic hydrocarbon metabolites by enzyme polymorphisms in workers of the German Human Bitumen Study. *Arch. Toxicol.* **85**, S73–S79 (2011).
 78. Ravegnini, G. *et al.* Key Genetic and Epigenetic Mechanisms in Chemical Carcinogenesis. *Toxicol. Sci.* **148**, 2–13 (2015).
 79. Salama, S. A., Sierra-Torres, C. H., Young Oh, H., Hamada, F. A. & Au, W. W. Variant metabolizing gene alleles determine the genotoxicity of benzo[a]pyrene. *Environ. Mol. Mutagen.* **37**, 17–26 (2001).
 80. Xiao, M. *et al.* Genetic polymorphisms in 19q13.3 genes associated with alteration of repair capacity to BPDE-DNA adducts in primary cultured lymphocytes. *Mutat. Res.* **812**, 39–47 (2016).
 81. Marston, C. P. *et al.* Effect of a complex environmental mixture from coal tar containing polycyclic aromatic hydrocarbons (PAH) on the tumor initiation, PAH-DNA binding and metabolic activation of carcinogenic PAH in mouse epidermis. *Carcinogenesis* **22**, 1077–1086 (2001).
 82. Bjelogrić, N. M., Mäkinen, M., Stenbäck, F. & Vähäkangas, K. Benzo[a]pyrene-7,8-diol-9,10-epoxide-DNA adducts and increased p53 protein in mouse skin. *Carcinogenesis* **15**, 771–774 (1994).
 83. Lloyd, D. R. & Hanawalt, P. C. P53-Dependent Global Genomic Repair of Benzo [a] Pyrene-7, 8-Diol-9, 10-Epoxyde Adducts in Human Cells. *Cancer Res.* **60**, 517 (2000).
 84. Franco, R., Schoneveld, O., Georgakilas, A. G. & Panayiotidis, M. I. Oxidative stress, DNA methylation and carcinogenesis. *Cancer Lett.* **266**, 6–11 (2008).

85. Liu, J. *et al.* Epigenetically mediated pathogenic effects of phenanthrene on regulatory T cells. *J. Toxicol.* **2013**, (2013).
86. Lin, Y. *et al.* Urinary Metabolites of Polycyclic Aromatic Hydrocarbons and the Association with Lipid Peroxidation: A Biomarker-Based Study between Los Angeles and Beijing. *Environ. Sci. Technol.* acs.est.5b04629 (2016). doi:10.1021/acs.est.5b04629
87. Shi, Q., Godschalk, R. W. L. & Van Schooten, F. J. Inflammation and the chemical carcinogen benzo[a]pyrene: Partners in crime. *Mutat. Res. - Rev. Mutat. Res.* **774**, 12–24 (2017).
88. Kropachev, K. *et al.* Adenine-DNA adducts derived from the highly tumorigenic dibenzo[a,l]pyrene are resistant to nucleotide excision repair while guanine adducts are not. *Chem Res Toxicol.* **26**, 783–793 (2013).
89. van Kesteren, P. C. E. *et al.* Deregulation of Cancer-Related Pathways in Primary Hepatocytes Derived from DNA Repair-Deficient Xpa^{-/-}p53^{+/-} Mice upon Exposure to Benzo[a]pyrene. *Toxicol. Sci.* **123**, 123–132 (2011).
90. Wang, F. *et al.* Genetic variants of nucleotide excision repair genes are associated with DNA damage in coke oven workers. *Cancer Epidemiol. biomarkers Prev.* **19**, 211–8 (2010).
91. Sharovskaja, J. J., Vaiman, A. V., Solomatina, N. A. & Kobliakov, V. A. Inhibition of Gap junction intercellular communications in cell culture by polycyclic aromatic hydrocarbons (PAH) in the absence of PAH metabolism. *Biochem.* **69**, 511–518 (2004).
92. Weis, L. M., Rummel, A. M., Masten, S. J., Trosko, J. E. & Upham, B. L. Bay or baylike regions of polycyclic aromatic hydrocarbons were potent inhibitors of Gap junctional intercellular communication. *Environ. Health Perspect.* **106**, 17–22 (1998).
93. Osgood, R. S. *et al.* Polycyclic Aromatic Hydrocarbon-Induced Signaling Events Relevant to Inflammation and Tumorigenesis in Lung Cells Are Dependent on Molecular Structure. *PLoS One* **8**, e65150 (2013).
94. Blaha, L., Kapplova, P., Vondracek, J., Upham, B. & Machala, M. Inhibition of gap-junctional intercellular communication by environmentally occurring polycyclic aromatic hydrocarbons. *Toxicol. Sci.* **65**, 43–51 (2002).
95. Ehrenhauser, F. S. PAH and IUPAC Nomenclature. *Polycycl. Aromat. Hydrocarb.* **35**, 161–176 (2015).
96. Hockley, S. L., Arlt, V. M., Brewer, D., Giddings, I. & Phillips, D. H. Time- and concentration-dependent changes in gene expression induced by benzo(a)pyrene in two human cell lines, MCF-7 and HepG2. *BMC Genomics* **7**, 1–23 (2006).
97. Hockley, S. L. *et al.* AHR- and DNA-Damage-Mediated Gene Expression Responses Induced by Benzo(a)pyrene in Human Cell Lines. *Chem. Res. Toxicol.* **20**, 1797–1810 (2007).
98. Souza, T. *et al.* New insights into BaP-induced toxicity: role of major metabolites in transcriptomics and contribution to hepatocarcinogenesis. *Arch Toxicol* **90**, 1449–1458 (2016).
99. Lizarraga, D. *et al.* Benzo[a]pyrene-Induced Changes in MicroRNA–mRNA Networks. *Chem. Res. Toxicol.* **25**, 838–849 (2012).
100. Sparfel, L. *et al.* Transcriptional Signature of Human Macrophages Exposed to the Environmental Contaminant Benzo(a)pyrene. *Toxicol. Sci.* **114**, 247–259 (2010).
101. Castorena-Torres, F. *et al.* Changes in gene expression induced by polycyclic aromatic hydrocarbons in the human cell lines HepG2 and A549. *Toxicol. Vitro.* **22**, 411–421 (2008).

102. Sadikovic, B. & Rodenhiser, D. I. Benzopyrene exposure disrupts DNA methylation and growth dynamics in breast cancer cells. *Toxicol. Appl. Pharmacol.* **216**, 458–468 (2006).
103. Guo, J. *et al.* Effects of exposure to benzo[a]pyrene on metastasis of breast cancer are mediated through ROS-ERK-MMP9 axis signaling. *Toxicol. Lett.* **234**, 201–210 (2015).
104. Malik, A. I. *et al.* Hepatic genotoxicity and toxicogenomic responses in MutaTMMouse males treated with dibenz[a,h]anthracene. *Mutagenesis* **28**, 543–54 (2013).
105. Malik, A. I., Williams, A., Lemieux, C. L., White, P. A. & Yauk, C. L. Hepatic mRNA, microRNA, and miR-34a-target responses in mice after 28 days exposure to doses of benzo(a)pyrene that elicit DNA damage and mutation. *Environ. Mol. Mutagen.* **53**, 10–21 (2012).
106. Zuo, J., Brewer, D. S., Arlt, V. M., Cooper, C. S. & Phillips, D. H. Benzo pyrene-induced DNA adducts and gene expression profiles in target and non-target organs for carcinogenesis in mice. *BMC Genomics* **15**, 880 (2014).
107. Jung, K. H. *et al.* Characteristic molecular signature for the early detection and prediction of polycyclic aromatic hydrocarbons in rat liver. *Toxicol. Lett.* **216**, 1–8 (2013).
108. Maikawa, C. L. *et al.* Murine precision-cut lung slices exhibit acute responses following exposure to gasoline direct injection engine emissions. *Sci. Total Environ.* (2016). doi:10.1016/j.scitotenv.2016.06.173
109. Beranek, M. *et al.* Genetic polymorphisms in biotransformation enzymes for benzo[a]pyrene and related levels of benzo[a]pyrene-7,8-diol-9,10-epoxide-DNA adducts in Goeckerman therapy. *Toxicol. Lett.* **255**, 47–51 (2016).
110. Roth, M. J. *et al.* Aryl hydrocarbon receptor expression is associated with a family history of upper gastrointestinal tract cancer in a high-risk population exposed to aromatic hydrocarbons. *Cancer Epidemiol. Biomarkers Prev.* **18**, 2391–6 (2009).
111. White, A. J. *et al.* Sources of polycyclic aromatic hydrocarbons are associated with gene-specific promoter methylation in women with breast cancer. *Environ. Res.* **145**, 93–100 (2016).
112. Mordukhovich, I. *et al.* Polymorphisms in DNA repair genes, traffic-related polycyclic aromatic hydrocarbone exposure and breast cancer incidence. *Int. J. cancer* **124**, 30–38 (2016).
113. Niehoff, N. *et al.* Polycyclic aromatic hydrocarbons and postmenopausal breast cancer: An evaluation of effect measure modification by body mass index and weight change. *Environ. Res.* **152**, 17–25 (2016).
114. Parikh, P. V. & Wei, Y. PAHs and PM_{2.5} emissions and female breast cancer incidence in metro Atlanta and rural Georgia. *Int. J. Environ. Health Res.* **3123**, 1–9 (2016).
115. Sinha, R., Kulldorff, M., Gunter, M. J., Strickland, P. & Rothman, N. Dietary benzo[a]pyrene intake and risk of colorectal adenoma. *Cancer Epidemiol. Biomarkers Prev.* **14**, 2030–4 (2005).
116. Gunter, M. J. *et al.* Meat intake, cooking-related mutagens and risk of colorectal adenoma in a sigmoidoscopy-based case-control study. *Carcinogenesis* **26**, 637–642 (2004).
117. Tabatabaei, S. M., Heyworth, J. S., Knuiiman, M. W. & Fritschi, L. Dietary benzo[a]pyrene intake from meat and the risk of colorectal cancer. *Cancer Epidemiol. Biomarkers Prev.* **19**, 3182–4 (2010).
118. Shin, A. *et al.* Meat and meat-mutagen intake, doneness preference and the risk of colorectal polyps: The Tennessee colorectal polyp study. *Int. J. Cancer* **121**, 136–142 (2007).

119. Olsson, A. C. *et al.* Occupational exposure to polycyclic aromatic hydrocarbons and lung cancer risk: a multicenter study in Europe. *Occup. Environ. Med.* **67**, 98–103 (2010).
120. Helena Guerra Andersen, M. *et al.* Association between polycyclic aromatic hydrocarbon exposure and peripheral blood mononuclear cell DNA damage in human volunteers during fire extinction exercises. *Mutagenesis* **00**, 1–11 (2017).
121. Ricceri, F. *et al.* Bulky DNA adducts in white blood cells: a pooled analysis of 3,600 subjects. *Cancer Epidemiol. Biomarkers Prev.* **19**, 3174–81 (2010).
122. Pavanello, S. *et al.* Shorter telomere length in peripheral blood lymphocytes of workers exposed to polycyclic aromatic hydrocarbons. *Carcinogenesis* **31**, 216–221 (2010).
123. Pavanello, S. *et al.* Mitochondrial DNA copy number and exposure to polycyclic aromatic hydrocarbons. *Cancer Epidemiol. Biomarkers Prev.* **22**, 1722–9 (2013).
124. Perera, F., Tang, D., Whyatt, R., Lederman, S. A. & Jedrychowski, W. DNA Damage from Polycyclic Aromatic Hydrocarbons Measured by Benzo[a]pyrene-DNA Adducts in Mothers and Newborns from Northern Manhattan, The World Trade Center Area, Poland, and China. *Cancer Epidemiol. Biomarkers Prev.* **14**, 709–714 (2005).
125. Choi, H. *et al.* International studies of prenatal exposure to polycyclic aromatic hydrocarbons and fetal growth. *Environ. Health Perspect.* **114**, 1744–50 (2006).
126. Deng, Q. *et al.* Polycyclic aromatic hydrocarbon exposure, miR-146a rs2910164 polymorphism, and heart rate variability in coke oven workers. *Environ. Res.* **148**, 277–284 (2016).
127. Burchiel, S. W. & Luster, M. I. Signaling by Environmental Polycyclic Aromatic Hydrocarbons in Human Lymphocytes. *Clin. Immunol.* **98**, 2–10 (2001).
128. Yin, W. *et al.* Obesity mediated the association of exposure to polycyclic aromatic hydrocarbon with risk of cardiovascular events. *Sci. Total Environ.* **616–617**, 841–854 (2017).
129. Burstyn, I. *et al.* Polycyclic Aromatic Hydrocarbons and Fatal Ischemic Heart Disease. *Epidemiology* **16**, 744–750 (2005).
130. Choi, Y.-H., Kim, J. H. & Hong, Y.-C. CYP1A1 genetic polymorphism and polycyclic aromatic hydrocarbons on pulmonary function in the elderly: Haplotype-based approach for gene–environment interaction. *Toxicol. Lett.* **221**, 185–190 (2013).
131. Hou, J. *et al.* Combined effect of urinary monohydroxylated polycyclic aromatic hydrocarbons and impaired lung function on diabetes. *Environ. Res.* **148**, 467–474 (2016).
132. Burstyn, I. *et al.* Mortality from Obstructive Lung Diseases and Exposure to Polycyclic Aromatic Hydrocarbons among Asphalt Workers. *Am. J. Epidemiol.* **158**, 468–478 (2003).
133. Jurewicz, J. *et al.* Exposure to widespread environmental endocrine disrupting chemicals and human sperm sex ratio. *Environ. Pollut.* **213**, 732–740 (2016).
134. Ling, X. *et al.* TERT regulates telomere-related senescence and apoptosis through DNA damage response in male germ cells exposed to BPDE in vitro and to B[a]P in vivo. *Environ. Pollut.* **235**, 836–849 (2018).
135. Bolden, A. L., Rochester, J. R., Schultz, K. & Kwiatkowski, C. F. Polycyclic aromatic hydrocarbons and female reproductive health: A scoping review. *Reprod. Toxicol.* **73**, 61–74 (2017).
136. Fang, X. *et al.* Transcriptomic Changes in Zebrafish Embryos and Larvae Following Benzo[a]pyrene Exposure. *Toxicol. Sci.* **146**, 395–411 (2015).

137. Gao, D. *et al.* Early-Life Benzo[a]Pyrene Exposure Causes Neurodegenerative Syndromes in Adult Zebrafish (*Danio rerio*) and the Mechanism Involved. *Toxicol. Sci.* **157**, 74–84 (2017).
138. Knecht, A. L. *et al.* Transgenerational inheritance of neurobehavioral and physiological deficits from developmental exposure to benzo[a]pyrene in zebrafish. *Toxicol. Appl. Pharmacol.* **329**, 148–157 (2017).
139. Qiu, C. *et al.* Effects of subchronic benzo(a)pyrene exposure on neurotransmitter receptor gene expression in the rats hippocampus related with spatial learning and memory change. *Toxicology* **289**, 83–90 (2011).
140. Chengzhi, C. *et al.* New candidate proteins for Benzo(a)pyrene-induced spatial learning and memory deficits. *J. Toxicol. Sci.* **36**, 163–171 (2011).
141. Chepelev, N. L. *et al.* Transcriptional profiling of the mouse hippocampus supports an NMDAR-mediated neurotoxic mode of action for benzo[a]pyrene. *Environ. Mol. Mutagen.* **57**, 350–363 (2016).
142. Labib, S. *et al.* Subchronic Oral Exposure to Benzo(a)pyrene Leads to Distinct Transcriptomic Changes in the Lungs That Are Related to Carcinogenesis. *Toxicol. Sci.* **129**, 213–224 (2012).
143. Labib, S. *et al.* Comparative transcriptomic analyses to scrutinize the assumption that genotoxic PAHs exert effects via a common mode of action. *Arch. Toxicol.* (2015). doi:10.1007/s00204-015-1595-5
144. Perera, F. P. *et al.* Effect of prenatal exposure to airborne polycyclic aromatic hydrocarbons on neurodevelopment in the first 3 years of life among inner-city children. *Environ. Health Perspect.* **114**, 1287–1292 (2006).
145. Edwards, S. C. *et al.* Prenatal Exposure to Airborne Polycyclic Aromatic Hydrocarbons and Children’s Intelligence at 5 Years of Age in a Prospective Cohort Study in Poland. *Environ. Health Perspect.* **118**, 1326–1331 (2010).
146. Perera, F. P. *et al.* Prenatal airborne polycyclic aromatic hydrocarbon exposure and child IQ at age 5 years. *Pediatrics* **124**, e195-202 (2009).
147. Perera, F. P. *et al.* Combined effects of prenatal exposure to polycyclic aromatic hydrocarbons and material hardship on child ADHD behavior problems. *Environ. Res.* **160**, 506–513 (2017).
148. Margolis, A. E. *et al.* Longitudinal effects of prenatal exposure to air pollutants on self-regulatory capacities and social competence. *J. Child Psychol. Psychiatry* n/a-n/a (2016). doi:10.1111/jcpp.12548
149. Barth, T. K. & Imhof, A. Fast signals and slow marks: the dynamics of histone modifications. *Trends Biochem. Sci.* **35**, 618–626 (2010).
150. Baylin, S. B. DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.* **2**, S4–S11 (2005).
151. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
152. Cannell, I. G., Kong, Y. W. & Bushell, M. How do microRNAs regulate gene expression? *Biochem. Soc. Trans.* **36**, 1224–31 (2008).
153. Cheung, P., David Allis, C. & Sassone-Corsi, P. Signaling to Chromatin Through Histone Modifications. *Cell* **103**, 263–271 (2000).
154. Dean, W., Lucifero, D. & Santos, F. DNA methylation in mammalian development and disease.

- Birth Defects Res. Part C - Embryo Today Rev.* **75**, 98–111 (2005).
155. Goldberg, A. D., Allis, D. & Bernstein, E. Epigenetics: A Landscape Takes Shape. *Cell* **128**, 635–638 (2007).
 156. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007).
 157. Santos-Rosa, H. & Caldas, C. Chromatin modifier enzymes, the histone code and cancer. *Eur. J. Cancer* **41**, 2381–2402 (2005).
 158. Virani, S., Virani, S., Colacino, J. A., Kim, J. H. & Rozek, L. S. Cancer epigenetics: a brief review. *ILAR J.* **53**, 359–69 (2012).
 159. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–20 (2013).
 160. Breiling, A. & Lyko, F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics {&} Chromatin* **8**, 24 (2015).
 161. Baccarelli, A. & Bollati, V. Epigenetics and environmental chemicals. *Curr. Opin. Pediatr.* **21**, 243–251 (2009).
 162. Chia, N. *et al.* Hypothesis: Environmental regulation of 5-hydroxymethylcytosine by oxidative stress. *Epigenetics* **6**, 853–856 (2011).
 163. Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.* **13**, 97–109 (2012).
 164. Ho, S.-M. *et al.* Environmental epigenetics and its implication on disease risk and health outcomes. *ILAR J.* **53**, 289–305 (2012).
 165. Ruiz-Hernandez, A. *et al.* Environmental chemicals and DNA methylation in adults: a systematic review of the epidemiologic evidence. *Clin. Epigenetics* **7**, 55 (2015).
 166. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191 (2010).
 167. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
 168. Gardiner-Garden, M. & Frommer, M. CpG Islands in Vertebrate Genomes. *J.Mol.Biol.* **196**, 261–282 (1987).
 169. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–86 (2009).
 170. Yang, X. *et al.* Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell* **26**, 577–590 (2014).
 171. Clapier, C. R. & Cairns, B. R. The Biology of Chromatin Remodeling Complexes. *Annu. Rev. Biochem.* **78**, 273–304 (2009).
 172. Stein, R. A. Epigenetics and environmental exposures. *J. Epidemiol. Community Health* **66**, 8–13 (2012).
 173. Michels, K. B. The promises and challenges of epigenetic epidemiology. *Exp. Gerontol.* **45**, 297–301 (2010).
 174. Chen, J. X., Zheng, Y., West, M. & Tang, M. S. Carcinogens preferentially bind at methylated CpG in the p53 mutational hot spots. *Cancer Res.* **58**, 2070–2075 (1998).

175. Pfeifer, G. P., Grunberger, D. & Drahovsky, D. Impaired enzymatic methylation of BPDE-modified DNA. **5**, 931–935 (1984).
176. Wojciechowskis, M. F. & Meehan, T. Inhibition of DNA Methyltransferases in Vitro by Benzo [a] pyrene Diol Epoxide-modified Substrates *. *J. Biol. Chem.* **259**, 9711–9716 (1984).
177. Gromova, E. S., Subach, O. M., Baskunov, V. B. & Geacintov, N. E. in *Structural Biology of DNA Damage and Repair* 103–116 (American Chemical Society, 2010).
178. Huang, X., Colgate, K. C., Kolbanovskiy, A., Amin, S. & Geacintov, N. E. Conformational Changes of a Benzo[a]pyrene Diol Epoxide-N 2-dG Adduct Induced by a 5'-Flanking 5-Methyl-Substituted Cytosine in a Me CG Double-Stranded Oligonucleotide Sequence Context. *Chem Res Toxicol* **15**, 438–444 (2002).
179. Cai, Y. *et al.* A bulky DNA lesion derived from a highly potent polycyclic aromatic tumorigen stabilizes nucleosome core particle structure. *Biochemistry* **49**, 9943–5 (2010).
180. Fu, I., Cai, Y., Geacintov, N. E., Zhang, Y. & Broyde, S. Nucleosome Histone Tail Conformation and Dynamics: Impacts of Lysine Acetylation and a Nearby Minor Groove Benzo[a]pyrene-Derived Lesion. *Biochemistry* **56**, 1963–1973 (2017).
181. Teneng, I., Montoya-Durango, D. E., Quertermous, J. L., Lacy, M. E. & Ramos, K. S. Reactivation of L1 retrotransposon by benzo(a)pyrene involves complex genetic and epigenetic regulation. *Epigenetics* **6**, 355–367 (2011).
182. Sadikovic, B., Andrews, J., Carter, D., Robinson, J. & Rodenhiser, D. I. Genome-wide H3K9 Histone Acetylation Profiles Are Altered in Benzopyrene-treated MCF7 Breast Cancer Cells. *J. Biol. Chem.* **28**, 4051–4060 (2008).
183. Laknaur, A. *et al.* Altered expression of histone deacetylases, inflammatory cytokines and contractile-associated factors in uterine myometrium of Long Evans rats gestationally exposed to benzo[a]pyrene. *J. Appl. Toxicol.* **36**, 827–35 (2016).
184. Schnekenburger, M., Peng, L. & Puga, A. HDAC1 bound to the Cyp1a1 promoter blocks histone acetylation associated with Ah receptor-mediated trans-activation. *Biochim. Biophys. Acta* **1769**, 569–78 (2007).
185. Lemieux, C. L. *et al.* Simultaneous Measurement of Benzo[a]pyrene- induced Pig-a and lacZ Mutations, Micronuclei and DNA Adducts in Muta TM Mouse. *Environ. Mol. Mutagen.* **52**, 756–765 (2011).
186. Veillard, A.-C., Datlinger, P., Laczik, M., Squazzo, S. & Bock, C. Diagenode® Premium RRBS technology: cost-effective DNA methylation mapping with superior coverage. *Nat. Publ. Gr.* **13**, (2016).
187. A, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, 87 (2012).
188. Davis, A. P. *et al.* The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res.* **45**, D972–D978 (2017).
189. Riboli, E. & Kaaks, R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* **26 Suppl 1**, S6–S14 (1997).
190. Beulens, J. W. J. *et al.* Cohort Profile: The EPIC-NL study. *Int. J. Epidemiol.* **39**, 1170–1178 (2010).
191. Plusquin, M. *et al.* DNA methylation and exposure to ambient air pollution in two prospective

- cohorts. *Environ. Int.* **108**, 127–136 (2017).
192. Campanella, G. *et al.* Epigenome-wide association study of adiposity and future risk of obesity-related diseases. *Int. J. Obes.* (2018). doi:10.1038/s41366-018-0064-7
193. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–9 (2013).
194. Price, M. E. *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6**, 4 (2013).
195. Martorell, I. *et al.* Polycyclic aromatic hydrocarbons (PAH) in foods and estimated PAH intake by the population of Catalonia, Spain: Temporal trend. *Environ. Int.* **36**, 424–432 (2010).
196. Martí-Cid, R., Llobet, J. M., Castell, V. & Domingo, J. L. Evolution of the dietary exposure to polycyclic aromatic hydrocarbons in Catalonia, Spain. *Food Chem. Toxicol.* **46**, 3163–3171 (2008).
197. Falcó, G. *et al.* Polycyclic aromatic hydrocarbons in foods: human exposure through the diet in Catalonia, Spain. *J. Food Prot.* **66**, 2325–2331 (2003).
198. de Vos, R. H., van Dokkum, W., Schouten, A. & de Jong-Berkhout, P. Polycyclic aromatic hydrocarbons in Dutch total diet samples (1984–1986). *Food Chem. Toxicol.* **28**, 263–268 (1990).
199. Duedahl-Olesen, L., White, S. & Binderup, M.-L. Polycyclic Aromatic Hydrocarbons (Pah) in Danish Smoked Fish and Meat Products. *Polycycl. Aromat. Compd.* **26**, 163–184 (2006).
200. Yurchenko, S. & Mölder, U. The determination of polycyclic aromatic hydrocarbons in smoked fish by gas chromatography mass spectrometry with positive-ion chemical ionization. *J. Food Compos. Anal.* **18**, 857–869 (2005).
201. Dennis, M. J., Massey, R. C., McWeeny, D. J., Knowles, M. E. & Watson, D. Analysis of polycyclic aromatic hydrocarbons in UK total diets. *Food Chem. Toxicol.* **21**, 569–74 (1983).
202. Karl, H. & Leinemann, M. Determination of polycyclic aromatic hydrocarbons in smoked fishery products from different smoking kilns. *Z. Lebensm. Unters. Forsch.* **202**, 458–64 (1996).
203. Mottier, P., Parisod, V. & Turesky, R. J. Quantitative Determination of Polycyclic Aromatic Hydrocarbons in Barbecued Meat Sausages by Gas Chromatography Coupled to Mass Spectrometry. *J. Agr. Food Chem.* **48**, 1160–1166 (2000).
204. García Falcón, M. S., González Amigo, S., Lage Yusty, M. a & Simal Lozano, J. Determination of benzo[a]pyrene in some Spanish commercial smoked products by HPLC-FL. *Food Addit. Contam.* **16**, 9–14 (1999).
205. Barranco, A. *et al.* Solid-phase clean-up in the liquid chromatographic determination of polycyclic aromatic hydrocarbons in edible oils. *J. Chromatogr. A* **988**, 33–40 (2003).
206. Guillén, M. D., Sopelana, P. & Palencia, G. Polycyclic Aromatic Hydrocarbons and Olive Pomace Oil. *J. Agric. Food Chem.* **52**, 2123–2132 (2004).
207. Moreda, W., Rodríguez-Acuña, R., Pérez-Camino, M. D. C. & Cert, A. Determination of high molecular mass polycyclic aromatic hydrocarbons in refined olive pomace and other vegetable oils. *J. Sci. Food Agric.* **84**, 1759–1764 (2004).
208. Purcaro, G., Morrison, P., Moret, S., Conte, L. S. & Marriott, P. J. Determination of polycyclic

- aromatic hydrocarbons in vegetable oils using solid-phase microextraction-comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry. *J. Chromatogr. A* **1161**, 284–91 (2007).
209. Aguinaga, N., Campillo, N., Viñas, P. & Hernández-Córdoba, M. A headspace solid-phase microextraction procedure coupled with gas chromatography-mass spectrometry for the analysis of volatile polycyclic aromatic hydrocarbons in milk samples. *Anal. Bioanal. Chem.* **391**, 753–8 (2008).
210. Grova, N. *et al.* Detection of polycyclic aromatic hydrocarbon levels in milk collected near potential contamination sources. *J. Agric. Food Chem.* **50**, 4640–4642 (2002).
211. Viñas, P., Campillo, N., Aguinaga, N., Pérez-Cánovas, E. & Hernández-Córdoba, M. Use of headspace solid-phase microextraction coupled to liquid chromatography for the analysis of polycyclic aromatic hydrocarbons in tea infusions. *J. Chromatogr. A* **1164**, 10–7 (2007).
212. Djinic, J., Popovic, A. & Jira, W. Polycyclic aromatic hydrocarbons (PAHs) in different types of smoked meat products from Serbia. *Meat Sci.* **80**, 449–456 (2008).
213. Djinic, J., Popovic, A. & Jira, W. Polycyclic aromatic hydrocarbons (PAHs) in traditional and industrial smoked beef and pork ham from Serbia. *Eur. Food Res. Technol.* **227**, 1191–1198 (2008).
214. Perugini, M. *et al.* Polycyclic aromatic hydrocarbons in marine organisms from the Gulf of Naples, Tyrrhenian Sea. *J. Agric. Food Chem.* **55**, 2049–54 (2007).
215. Ramalhosa, M. J., Paíga, P., Morais, S., Delerue-Matos, C. & Oliveira, M. B. P. P. Analysis of polycyclic aromatic hydrocarbons in fish: Evaluation of a quick, easy, cheap, effective, rugged, and safe extraction method. *J. Sep. Sci.* **32**, 3529–3538 (2009).
216. Aguinaga, N., Campillo, N., Viñas, P. & Hernández-Córdoba, M. Evaluation of solid-phase microextraction conditions for the determination of polycyclic aromatic hydrocarbons in aquatic species using gas chromatography. *Anal. Bioanal. Chem.* **391**, 1419–24 (2008).
217. Anastasio, A., Mercogliano, R., Vollano, L., Pepe, T. & Cortesi, M. L. Levels of benzo[a]pyrene (BaP) in ‘mozzarella di bufala campana’ cheese smoked according to different procedures. *J. Agric. Food Chem.* **52**, 4452–5 (2004).
218. Suchanová, M., Hajšlová, J., Tomaniová, M., Kocourek, V. & Babička, L. Polycyclic aromatic hydrocarbons in smoked cheese. *J. Sci. Food Agric.* **88**, 1307–1317 (2008).
219. Pagliuca, P. *et al.* Determination of High Molecular Mass Polycyclic Aromatic Hydrocarbons in a Typical Italian Smoked Cheese by HPLC-FL. *J. Agr. Food Chem.* **51**, 5111–5115 (2003).
220. Bordajandi, L. R. *et al.* Survey of Persistent Organochlorine Contaminants (PCBs, PCDD/Fs, and PAHs), Heavy Metals (Cu, Cd, Zn, Pb, and Hg), and Arsenic in Food Samples from Huelva (Spain): Levels and Health Implications. *J. Agric. Food Chem.* **52**, 992–1001 (2004).
221. Reinik, M. *et al.* Polycyclic aromatic hydrocarbons (PAHs) in meat products and estimated PAH intake by children and the general population in Estonia. *Food Addit. Contam.* **24**, 429–37 (2007).
222. Stumpe-Viksna, I., Bartkevičs, V., Kukare, A. & Morozovs, A. Polycyclic aromatic hydrocarbons in meat smoked with different types of wood. *Food Chem.* **110**, 794–797 (2008).
223. Wretling, S., Eriksson, a., Eskhult, G. a. & Larsson, B. Polycyclic aromatic hydrocarbons (PAHs) in Swedish smoked meat and fish. *J. Food Compos. Anal.* **23**, 264–272 (2010).

224. Larsson, B. K., Sahlberg, G. P., Eriksson, A. T. & Busk, L. a. Polycyclic aromatic hydrocarbons in grilled food. *J. Agric. Food Chem.* **31**, 867–873 (1983).
225. Purcaro, G., Moret, S. & Conte, L. S. Optimisation of microwave assisted extraction (MAE) for polycyclic aromatic hydrocarbon (PAH) determination in smoked meat. *Meat Sci.* **81**, 275–280 (2009).
226. Perelló, G., Martí-Cid, R., Castell, V., Llobet, J. M. & Domingo, J. L. Concentrations of polybrominated diphenyl ethers, hexachlorobenzene and polycyclic aromatic hydrocarbons in various foodstuffs before and after cooking. *Food Chem. Toxicol.* **47**, 709–15 (2009).
227. Orecchio, S., Ciotti, V. P. & Culotta, L. Polycyclic aromatic hydrocarbons (PAHs) in coffee brew samples: Analytical method by GC-MS, profile, levels and sources. *Food Chem. Toxicol.* **47**, 819–826 (2009).
228. Jira, W., Ziegenhals, K. & Speer, K. A GC/MS method for the determination of 16 European priority polycyclic aromatic hydrocarbons in smoked meat products and edible oils. *Food Addit. Contam.* **25**, 704–713 (2008).
229. Janoszka, B., Warzecha, L., Błaszczuk, U. & Bodzek, D. Organic compounds formed in thermally treated high-protein food. Part II: Azaarenes. *Acta Chromatogr.* 129–141 (2004). at <<http://www.scopus.com/inward/record.url?eid=2-s2.0-3142653689&partnerID=tZOtx3y1>>
230. Ciecierska, M. & Obiedziński, M. Influence of Smoking Process on Polycyclic Aromatic Hydrocarbons ' Content. *Acta Sci. Pol., Technol. Aliment.* **6**, 17–28 (2007).
231. Fasano, E., Esposito, F., Scognamiglio, G. & Cirillo, T. Detection of Polycyclic Aromatic Hydrocarbons in smoked buffalo mozzarella cheese produced in Campania Region (Italy). *J. Sci. Food Agric.* (2015). doi:10.1002/jsfa.7275
232. Cirillo, T., Milano, N. & Cocchieri, R. A. Polycyclic aromatic hydrocarbons (PAHs) in traditional smoked dairy products from Campania (Italy). *Ital. J. Public Health* **1**, 51–53 (2004).
233. Larsson, B. K., Eriksson, A. T. & Cervenka, M. Polycyclic aromatic hydrocarbons in crude and deodorized vegetable oils. *J. Am. Oil Chem. Soc.* **64**, 365–370 (1987).
234. Moret, S., Piani, B., Bortolomeazzi, R. & Contel, L. S. HPLC determination of polycyclic aromatic hydrocarbons in olive oils. *Z Leb. Unters Forsch A* **205**, 116–120 (1997).
235. García-Falcón, M. S. & Simal-Gándara, J. Determination of polycyclic aromatic hydrocarbons in alcoholic drinks and the identification of their potential sources. *Food Addit. Contam.* **22**, 791–797 (2005).
236. Fasano, E., Yebra-pimentel, I. & Martínez-carballo, E. Profiling , distribution and levels of carcinogenic polycyclic aromatic hydrocarbons in traditional smoked plant and animal foods. *Food Control* **59**, 581–590 (2016).
237. Moreda, W., Pérez-Camino, M. . & Cert, A. Gas and liquid chromatography of hydrocarbons in edible vegetable oils. *J. Chromatogr. A* **936**, 159–171 (2001).
238. Houessou, J. K. *et al.* Effect of roasting conditions on the polycyclic aromatic hydrocarbon content in ground Arabica coffee and coffee brew. *J. Agric. Food Chem.* **55**, 9719–9726 (2007).
239. Pincemaille, J., Schummer, C., Heinen, E. & Moris, G. Determination of polycyclic aromatic hydrocarbons in smoked and non-smoked black teas and tea infusions. *Food Chem.* **145**, 807–13 (2014).
240. Drabova, L. *et al.* Rapid determination of polycyclic aromatic hydrocarbons (PAHs) in tea using

- two-dimensional gas chromatography coupled with time of flight mass spectrometry. *Talanta* **100**, 207–16 (2012).
241. Drabova, L. *et al.* Application of solid phase extraction and two-dimensional gas chromatography coupled with time-of-flight mass spectrometry for fast analysis of polycyclic aromatic hydrocarbons in vegetable oils. *Food Control* **33**, 489–497 (2013).
 242. Jira, W. A GC/MS method for the determination of carcinogenic polycyclic aromatic hydrocarbons (PAH) in smoked meat products and liquid smokes. *Eur. Food Res. Technol.* **218**, 208–212 (2004).
 243. Naccari, C. *et al.* PAHs concentration in heat-treated milk samples. *Food Res. Int.* **44**, 716–724 (2011).
 244. Voutsas, D. & Samara, C. Dietary intake of trace elements and polycyclic aromatic hydrocarbons via vegetables grown in an industrial Greek area. *Sci. Total Environ.* **218**, 203–216 (1998).
 245. FSA (Food Standards Agency). Organic Environmental Contaminants in the 2012 Total Diet Study Samples Report to the Food Standards Agency. **44**, 1–83 (2012).
 246. Veyrand, B. *et al.* Human dietary exposure to polycyclic aromatic hydrocarbons: results of the second French Total Diet Study. *Environ. Int.* **54**, 11–7 (2013).
 247. Akdoğan, A., Buttinger, G. & Wenzl, T. Single-laboratory validation of a saponification method for the determination of four polycyclic aromatic hydrocarbons in edible oils by HPLC-fluorescence detection. *Food Addit. Contam. Part A* **33**, 215–224 (2016).
 248. Zelinkova, Z. & Wenzl, T. EU marker polycyclic aromatic hydrocarbons in food supplements: analytical approach and occurrence. *Food Addit. Contam. - Part A Chem. Anal. Control. Expo. Risk Assess.* **32**, 1914–1926 (2015).
 249. Rose, M. *et al.* Investigation into the formation of PAHs in foods prepared in the home to determine the effects of frying, grilling, barbecuing, toasting and roasting. *Food Chem. Toxicol.* **78**, 1–9 (2015).
 250. Raters, M. & Matissek, R. Quantitation of polycyclic aromatic hydrocarbons (PAH4) in cocoa and chocolate samples by an HPLC-FD method. *J. Agric. Food Chem.* **62**, 10666–71 (2014).
 251. Battisti, C., Girelli, A. M. & Tarola, A. M. Polycyclic aromatic hydrocarbons (PAHs) in yogurt samples. *Food Addit. Contam. Part B* **8**, 50–55 (2015).
 252. Schulz, C. M., Fritz, H. & Ruthenschrör, A. Occurrence of 15 + 1 EU priority polycyclic aromatic hydrocarbons (PAH) in various types of tea (*Camellia sinensis*) and herbal infusions. *Food Addit. Contam. - Part A Chem. Anal. Control. Expo. Risk Assess.* **31**, 1723–1735 (2014).
 253. Vieira Madureira, T. *et al.* A step forward using QuEChERS (Quick, Easy, Cheap, Effective, Rugged, and Safe) based extraction and gas chromatography-tandem mass spectrometry—levels of priority polycyclic aromatic hydrocarbons in wild and commercial mussels. *Env. Dci Pollut Res* **21**, 6089–6098 (2014).
 254. Girelli, A. M., Sperati, D. & Tarola, A. M. Determination of polycyclic aromatic hydrocarbons in Italian milk by HPLC with fluorescence detection. *Food Addit. Contam. Part A* **31**, 703–710 (2014).
 255. Ciemiński, A., Witczak, A. & Mocek, K. Assessment of honey contamination with polycyclic aromatic hydrocarbons. *J. Environ. Sci. Heal. - Part B Pestic. Food Contam. Agric. Wastes* **48**, 993–998 (2013).

256. Purcaro, G., Picardo, M., Barp, L., Moret, S. & Conte, L. S. Direct-immersion solid-phase microextraction coupled to fast gas chromatography mass spectrometry as a purification step for polycyclic aromatic hydrocarbons determination in olive oil. *J. Chromatogr. A* **1307**, 166–71 (2013).
257. Sadowska-Rociek, A., Surma, M. & Cieřlik, E. Comparison of different modifications on QuEChERS sample preparation method for PAHs determination in black, green, red and white tea. *Environ. Sci. Pollut. Res. Int.* **21**, 1326–38 (2014).
258. Gomes, F., Oliveira, M., Ramalhosa, M. J., Delerue-Matos, C. & Morais, S. Polycyclic aromatic hydrocarbons in commercial squids from different geographical origins: levels and risks for human consumption. *Food Chem. Toxicol.* **59**, 46–54 (2013).
259. Hitzel, A., Pöhlmann, M., Schwägele, F., Speer, K. & Jira, W. Polycyclic aromatic hydrocarbons (PAH) and phenolic substances in meat products smoked with different types of wood and smoking spices. *Food Chem.* **139**, 955–62 (2013).
260. Aaslyng, M. D., Duedahl-Olesen, L., Jensen, K. & Meinert, L. Content of heterocyclic amines and polycyclic aromatic hydrocarbons in pork, beef and chicken barbecued at home by Danish consumers. *Meat Sci.* **93**, 85–91 (2013).
261. Gosetti, F. *et al.* Simultaneous determination of thirteen polycyclic aromatic hydrocarbons and twelve aldehydes in cooked food by an automated on-line solid phase extraction ultra high performance liquid chromatography tandem mass spectrometry. *J. Chromatogr. A* **1218**, 6308–18 (2011).
262. Duedahl-Olesen, L., Christensen, J. H., Højgaard, A., Granby, K. & Timm-Heinrich, M. Influence of smoking parameters on the concentration of polycyclic aromatic hydrocarbons (PAHs) in Danish smoked fish. *Food Addit. Contam. - Part A Chem. Anal. Control. Expo. Risk Assess.* **27**, 1294–1305 (2010).
263. Kuhn, K., Nowak, B., Behnke, A., Seidel, A. & Lampen, A. Effect-based and chemical analysis of polycyclic aromatic hydrocarbons in smoked meat: A practical food-monitoring approach. *Food Addit. Contam. - Part A Chem. Anal. Control. Expo. Risk Assess.* **26**, 1104–1112 (2009).
264. Danyi, S. *et al.* Analysis of EU priority polycyclic aromatic hydrocarbons in food supplements using high performance liquid chromatography coupled to an ultraviolet, diode array or fluorescence detector. *Anal. Chim. Acta* **633**, 293–9 (2009).
265. Orecchio, S. & Papuzza, V. Levels, fingerprint and daily intake of polycyclic aromatic hydrocarbons (PAHs) in bread baked using wood as fuel. *J. Hazard. Mater.* **164**, 876–83 (2009).
266. Rodríguez-Acuna, R., Pérez-Camino, M. D. C., Cert, A. & Moreda, W. Polycyclic aromatic hydrocarbons in Spanish olive oils: Relationship between benzo(a)pyrene and total polycyclic aromatic hydrocarbon content. *J. Agric. Food Chem.* **56**, 10428–10432 (2008).
267. Hernández-Poveda, G. F., Morales-Rubio, A., Pastor-García, A. & De La Guardia, M. Extraction of polycyclic aromatic hydrocarbons from cookies: A comparative study of ultrasound and microwave-assisted procedures. *Food Addit. Contam. - Part A Chem. Anal. Control. Expo. Risk Assess.* **25**, 356–363 (2008).
268. Jáněká, M., Hajřlová, J., Tomaniová, M., Kocourek, V. & Vávrová, M. Polycyclic aromatic hydrocarbons in fruits and vegetables grown in the Czech Republic. *Bull. Environ. Contam. Toxicol.* **77**, 492–499 (2006).
269. Moret, S., Conte, L. & Dean, D. Assessment of Polycyclic Aromatic Hydrocarbon Content of Smoked Fish by Means of a Fast HPLC/HPLC Method. *J. Agric. Food Chem.* **47**, 1367–1371

- (1999).
270. EFSA. Findings of the EFSA Data Collection on Polycyclic Aromatic Hydrocarbons in Food. *Efsa J* **724**, 1–55 (2008).
271. Abramsson-Zetterberg, L., Darnerud, P. O. & Wretling, S. Low intake of polycyclic aromatic hydrocarbons in Sweden: results based on market basket data and a barbecue study. *Food Chem. Toxicol.* **74**, 107–11 (2014).
272. FSA (Food Standards Agency). PAHs in the UK diet: 2000 total diet study samples. *Brand* (2002). at <http://tna.europarchive.org/20110116113217/http://www.food.gov.uk/multimedia/pdfs/31pah.pdf>
273. Wiekstrom, K., Pyysalo, H., Plaami-heikkilii, S. & Tuominen, J. Polycyclic aromatic compounds (PAC) in leaf lettuce. *Z Leb. Unters Forsch* **183**, 182–185 (1986).
274. Larsson, B. K. Polycyclic Aromatic Hydrocarbons in Smoked Fish. *Z Leb. Unters Forsch* **174**, 101–107 (1982).
275. Crosby, N. T., Hunt, D. C. & Philp, L. A. Polynuclear Aromatic Hydrocarbons in Food, Water and Smoke Using High-performance Liquid Chromatography. **106**, 135–145 (1981).
276. Sannino, A. Polycyclic aromatic hydrocarbons in Italian preserved food products in oil. *Food Addit. Contam. Part B* **3210**, 19393210.2016.1145148 (2016).
277. Rozentāle, I. *et al.* Assessment of dietary exposure to polycyclic aromatic hydrocarbons from smoked meat products produced in Latvia. *Food Control* **54**, 16–22 (2015).
278. Martena, M. J., Grutters, M. M. P., De Groot, H. N., Konings, E. J. M. & Rietjens, I. M. C. M. Monitoring of polycyclic aromatic hydrocarbons (PAH) in food supplements containing botanicals and other ingredients on the Dutch market. *Food Addit. Contam. Part A* **28**, 925–942 (2011).
279. Duedahl-Olesen, L., Navaratnam, M. A., Jewula, J. & Jensen, A. H. PAH in Some Brands of Tea and Coffee. *Polycycl. Aromat. Compd.* **35**, 74–90 (2014).
280. Rodríguez-Hernández, A. *et al.* Daily intake of anthropogenic pollutants through yogurt consumption in the Spanish population. *J Appl Anim. Res.* **2119**, (2014).
281. Mercogliano, R. *et al.* Occurrence and distribution of polycyclic aromatic hydrocarbons in mussels from the gulf of Naples, Tyrrhenian Sea, Italy. *Mar. Pollut. Bull.* 1–5 (2016). doi:10.1016/j.marpolbul.2016.01.015
282. Ramalhosa, M. J. *et al.* Polycyclic aromatic hydrocarbon levels in three pelagic fish species from Atlantic Ocean: Inter-specific and inter-season comparisons and assessment of potential public health risks. *Food Chem. Toxicol.* **50**, 162–167 (2012).
283. CONTAM (EFSA Panel on Contaminants in the Food). Scientific opinion of the Panel on Contaminants in the Food Chain on a request from the European Commission on polycyclic aromatic hydrocarbons. *EFSA J.* **724**, 2–114 (2008).
284. Price, E. M. & Robinson, W. P. Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. *Front. Genet.* **9**, 83 (2018).
285. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

286. Ferrari, S. L. P. & Cribari-Neto, F. Beta Regression for Modelling Rates and Proportions. *J. Appl. Stat.* **31**, 799–815 (2004).
287. Campanella, G. *et al.* Epigenetic signatures of internal migration in Italy. *Int. J. Epidemiol.* **44**, 1442–1449 (2015).
288. Triche, T. J., Laird, P. W. & Siegmund, K. D. Beta regression improves the detection of differential DNA methylation for epigenetic epidemiology. *bioRxiv* (2016). doi:10.1101/054643
289. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
290. Saadati, M. & Benner, A. Statistical challenges of high-dimensional methylation data. *Stat. Med.* **33**, 5347–5357 (2014).
291. Yee, T. *Package 'VGAM'*. (2018). doi:10.1007/978-1-4939-2818-7
292. van Iterson, M. *Package 'bacon'*. (2019). at <<https://git.bioconductor.org/packages/bacon>>
293. Marzese, D. M. & Hoon, D. S. B. Emerging technologies for studying DNA methylation for the molecular diagnosis of cancer HHS Public Access. *Expert Rev Mol Diagn* **15**, 647–664 (2015).
294. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology (Basel)*. **5**, doi:10.3390/biology5010003 (2016).
295. Soozangar, N. *et al.* Comparison of genome-wide analysis techniques to DNA methylation analysis in human cancer. *J. Cell. Physiol.* **233**, 3968–3981 (2018).
296. Schumacher, A. *et al.* Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res.* **34**, 528–42 (2006).
297. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1827–31 (1992).
298. Bibikova, M. *et al.* High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.* **16**, 383–93 (2006).
299. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).
300. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. (2016). doi:10.1186/s13059-016-1066-1
301. Olova, N. *et al.* Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.* **19**, 33 (2018).
302. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–70 (2008).
303. Meissner, A. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).
304. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* **6**, (2011).
305. Gu, H. *et al.* Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat. Methods* **7**, 133–6 (2010).

306. Zhang, N. *et al.* Methylation of Cytosine at C5 in a CpG Sequence Context Causes a Conformational Switch of a Benzo[a]pyrene diol epoxide-N2-guanine Adduct in DNA from a Minor Groove Alignment to Intercalation with Base Displacement. *J. Mol. Biol.* **346**, 951–965 (2005).
307. Wang, H.-T. *et al.* Effect of CpG methylation at different sequence context on acrolein- and BPDE-DNA binding and mutagenesis. *Carcinogenesis* **34**, 220–7 (2013).
308. Huang, H. *et al.* Role of poly(ADP-ribose) glycohydrolase silencing in DNA hypomethylation induced by benzo(a)pyrene. *Biochem. Biophys. Res. Commun.* **452**, 708–714 (2014).
309. Minero, A. S. *et al.* Probing murine methyltransferase Dnmt3a interactions with benzo[a]pyrene-modified DNA by fluorescence methods. *FEBS J.* **279**, 3965–80 (2012).
310. Baskunov, V. B. *et al.* Effects of Benzo[a]pyrene-Deoxyguanosine Lesions on DNA Methylation Catalyzed by EcoRII DNA Methyltransferase and on DNA Cleavage Effected by EcoRII Restriction Endonuclease †. *Biochemistry* **44**, 1054–1066 (2005).
311. Subach, O. M. *et al.* The stereochemistry of benzo[a]pyrene-2'-deoxyguanosine adducts affects DNA methylation by SssI and HhaI DNA methyltransferases. *FEBS J.* **274**, 2121–2134 (2007).
312. Ye, F. & Xu, X.-C. Benzo[a]pyrene diol epoxide suppresses retinoic acid receptor-beta2 expression by recruiting DNA (cytosine-5-)-methyltransferase 3A. *Mol. Cancer* **9**, 93 (2010).
313. He, Z. *et al.* TRIM36 hypermethylation is involved in polycyclic aromatic hydrocarbons-induced cell transformation. *Environ. Pollut.* **2017**, 93–103 (2017).
314. Corrales, J. *et al.* Effects on specific promoter DNA methylation in zebrafish embryos and larvae following benzo[a]pyrene exposure. *Comp. Biochem. Physiol. C. Toxicol. Pharmacol.* **163**, 37–46 (2014).
315. Zhang, W., Tian, F., Zheng, J., Li, S. & Qiang, M. Chronic Administration of Benzo(a)pyrene Induces Memory Impairment and Anxiety-Like Behavior and Increases of NR2B DNA Methylation. *PLoS One* **11**, e0149574 (2016).
316. Tommasi, S., Zheng, A., Yoon, J.-I. & Besaratinia, A. Epigenetic targeting of the Nanog pathway and signaling networks during chemical carcinogenesis. *Carcinogenesis* **35**, 1726–1736 (2014).
317. Fish, T. J. & Benninghoff, A. D. Data on the effect of in utero exposure to polycyclic aromatic hydrocarbons on genome-wide patterns of DNA methylation in lung tissues. *Data Br.* **13**, 498–513 (2017).
318. Fish, T. J. & Benninghoff, A. D. DNA methylation in lung tissues of mouse offspring exposed in utero to polycyclic aromatic hydrocarbons. *Food Chem. Toxicol.* **109**, 703–713 (2017).
319. Cocci, P. *et al.* Investigating the potential impact of polycyclic aromatic hydrocarbons (PAHs) and polychlorinated biphenyls (PCBs) on gene biomarker expression and global DNA methylation in loggerhead sea turtles (*Caretta caretta*) from the Adriatic Sea. *Sci. Total Environ.* **619–620**, 49–57 (2018).
320. Brandenburg, J. & Head, J. A. Effects of in ovo exposure to benzo[k]fluoranthene (BkF) on CYP1A expression and promoter methylation in developing chicken embryos. *Comp. Biochem. Physiol. Part C Toxicol. Pharmacol.* **204**, 88–96 (2018).
321. Carmona, J. J. *et al.* Empirical comparison of reduced representation bisulfite sequencing and Infinium BeadChip reproducibility and coverage of DNA methylation in humans. *npj Genomic Med.* **2**, (2017).

322. McCormick, H. *et al.* Isogenic mice exhibit sexually-dimorphic DNA methylation patterns across multiple tissues. *BMC Genomics* **18**, DOI 10.1186/s12864-017-4350-x (2017).
323. Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).
324. Geng, Y. *et al.* Folate deficiency impairs decidualization and alters methylation patterns of the genome in mice. *Mol. Hum. Reprod.* **21**, 844–856 (2015).
325. Tryndyak, V. *et al.* Effect of aflatoxin B₁, benzo[a]pyrene, and methapyrilene on transcriptomic and epigenetic alterations in human liver HepaRG cells. *Food Chem. Toxicol.* **PII: S0278**, DOI: 10.1016/j.fct.2018.08.034 (2018).
326. Herlofsen, S. R. *et al.* Genome-wide map of quantified epigenetic changes during in vitro chondrogenic differentiation of primary human mesenchymal stem cells. *BMC Genomics* **14**, 105 (2013).
327. Zhou, Y. *et al.* Reduced representation bisulphite sequencing of ten bovine somatic tissues reveals DNA methylation patterns and their impacts on gene expression. *BMC Genomics* **17**, 779 (2016).
328. Mathijs, K. *et al.* Discrimination for Genotoxic and Nongenotoxic Carcinogens by Gene Expression Profiling in Primary Mouse Hepatocytes Improves with Exposure Time. *Toxicol. Sci.* **112**, 374–384 (2009).
329. Long, A. S., Lemieux, C. L., Arlt, V. M. & White, P. A. Tissue-specific in vivo genetic toxicity of nine polycyclic aromatic hydrocarbons assessed using the MutaTMMouse transgenic rodent assay. *Toxicol. Appl. Pharmacol.* **290**, 31–42 (2016).
330. Guerreiro, C. B. B., Horálek, J., de Leeuw, F. & Couvidat, F. Benzo(a)pyrene in Europe: Ambient air concentrations, population exposure and health effects. *Environ. Pollut.* **214**, 657–667 (2016).
331. Šrám, R. J. *et al.* Impact of air pollution to genome of newborns. *Cent Eur J Public Heal.* **24**, 40–44 (2016).
332. DeMarini, D. M. Genotoxicity biomarkers associated with exposure to traffic and near-road atmospheres: a review. *Mutagenesis* **28**, 485–505 (2013).
333. Desai, G., Chu, L., Guo, Y., Myneni, A. A. & Mu, L. Biomarkers used in studying air pollution exposure during pregnancy and perinatal outcomes: a review. *Biomarkers* **22**, 489–501 (2017).
334. Castano-Vinyals, G., D'Errico, A., Malats, N. & Kogevinas, M. Biomarkers of exposure to polycyclic aromatic hydrocarbons from environmental air pollution. *Occup Env. Med* e12 (2004). doi:10.1136/oem.2003.008375
335. Nilsson, R. *et al.* Exposure to polycyclic aromatic hydrocarbons in women from Poland, Serbia and Italy – relation between PAH metabolite excretion, DNA damage, diet and genotype (the EU DIEPHY project). *Biomarkers* **18**, 165–173 (2013).
336. Barbeau, D. *et al.* Urinary trans-anti-7,8,9,10-tetrahydroxy-7,8,9,10-tetrahydrobenzo(a)pyrene as the most relevant biomarker for assessing carcinogenic polycyclic aromatic hydrocarbons exposure. *Environ. Int.* **2018**, 147–155 (2017).
337. Grova, N. *et al.* Identification of new tetrahydroxylated metabolites of Polycyclic Aromatic Hydrocarbons in hair as biomarkers of exposure and signature of DNA adduct levels. (2017). doi:10.1016/j.aca.2017.10.002

338. Gunier, R. B. *et al.* Estimating Exposure to Polycyclic Aromatic Hydrocarbons: A Comparison of Survey, Biological Monitoring, and Geographic Information System–Based Methods. *Cancer Epidemiol Biomarkers Prev* **15**, 1376–81 (2006).
339. Oliveira, M. *et al.* Polycyclic aromatic hydrocarbons at fire stations: firefighters' exposure monitoring and biomonitoring, and assessment of the contribution to total internal dose. *J. Hazard. Mater.* **5**, 184–194 (2017).
340. Pacchierotti, F. & Spanò, M. Environmental Impact on DNA Methylation in the Germline: State of the Art and Gaps of Knowledge. *Biomed Res. Int.* **2015**, 1–23 (2015).
341. Alegría-Torres, J. A. *et al.* Epigenetic mechanisms of exposure to polycyclic aromatic hydrocarbons in Mexican brickmakers: A pilot study. *Chemosphere* **91**, 475–480 (2013).
342. Alhamdow, A. *et al.* DNA methylation of the cancer-related genes F2RL3 and AHRR is associated with occupational exposure to polycyclic aromatic hydrocarbons. *Carcinogenesis* (2018). doi:10.1093/carcin/bgy059
343. Duan, H. *et al.* Global and MGMT promoter hypomethylation independently associated with genomic instability of lymphocytes in subjects exposed to high-dose polycyclic aromatic hydrocarbon. *Arch. Toxicol.* **87**, 2013–2022 (2013).
344. Li, J. *et al.* Exposure to polycyclic aromatic hydrocarbons and accelerated DNA methylation ageing: an observational study. *Lancet* **386**, S20 (2015).
345. Herbstman, J. B. *et al.* Prenatal exposure to polycyclic aromatic hydrocarbons, benzo[a]pyrene-DNA adducts, and genomic DNA methylation in cord blood. *Env. Heal. Perspect* **120**, 733–738 (2012).
346. Pavanello, S. *et al.* Global and gene-specific promoter methylation changes are related to anti-B[a]PDE-DNA adduct levels and influence micronuclei levels in polycyclic aromatic hydrocarbon-exposed individuals. *Int. J. Cancer* **125**, 1692–1697 (2009).
347. Yang, J. *et al.* Urinary 1-hydroxypyrene and smoking are determinants of LINE-1 and AhRR promoter methylation in coke oven workers. *Mutat. Res. Toxicol. Environ. Mutagen.* **826**, 33–40 (2018).
348. Zhang, H. *et al.* Methylation of CpG island of p14(ARK), p15(INK4b) and p16(INK4a) genes in coke oven workers. *Hum. Exp. Toxicol.* **34**, 191–197 (2015).
349. Zhang, X. *et al.* Associations between DNA methylation in DNA damage response-related genes and cytokinesis-block micronucleus cytome index in diesel engine exhaust-exposed workers. *Arch. Toxicol.* **90**, 1997–2008 (2015).
350. Yang, P. *et al.* CpG Site-Specific Hypermethylation of p16INK4 in Peripheral Blood Lymphocytes of PAH-Exposed Workers. *Cancer Epidemiol. Biomarkers Prev.* **21**, 182–190 (2011).
351. He, Z. *et al.* CpG site-specific RASSF1a hypermethylation is associated with occupational PAH exposure and genomic instability. *Toxicol. Res.* **4**, 848–857 (2015).
352. Li, J. *et al.* Exposure to Polycyclic Aromatic Hydrocarbons and Accelerated DNA Methylation Aging. *Environ. Health Perspect.* **126**, 067005 (2018).
353. Perera, F. *et al.* Relation of DNA methylation of 5'-CpG island of ACSL3 to transplacental exposure to airborne polycyclic aromatic hydrocarbons and childhood asthma. *PLoS One* **4**, e4488 (2009).
354. Tian, M. *et al.* Association of environmental benzo[a]pyrene exposure and DNA methylation

- alterations in hepatocellular carcinoma: A Chinese case–control study. *Sci. Total Environ.* **541**, 1243–1252 (2016).
355. White, A. J. *et al.* Polycyclic aromatic hydrocarbon (PAH)-DNA adducts and breast cancer: modification by gene promoter methylation in a population-based study. *Cancer Causes Control* **26**, 1791–802 (2015).
356. Ouyang, B. *et al.* Hypomethylation of dual specificity phosphatase 22 promoter correlates with duration of service in firefighters and is inducible by low-dose benzo[a]pyrene. *J. Occup. Environ. Med.* **54**, 774–80 (2012).
357. Tang, W. *et al.* Maternal Exposure to Polycyclic Aromatic Hydrocarbons and 5'-CpG Methylation of Interferon- γ in Cord White Blood Cells. *Environ. Health Perspect.* **120**, 1195–1200 (2012).
358. Lee, J. *et al.* Prenatal airborne polycyclic aromatic hydrocarbon exposure, LINE1 methylation and child development in a Chinese cohort. *Environ. Int.* **99**, 315–320 (2017).
359. Hew, K. M. *et al.* Childhood exposure to ambient polycyclic aromatic hydrocarbons is linked to epigenetic modifications and impaired systemic immunity in T cells. *Clin Exp Allergy* **45**, 238–248 (2015).
360. Alvarado-Cruz, I. *et al.* Increased methylation of repetitive elements and DNA repair genes is associated with higher DNA oxidation in children in an urbanized, industrial environment. *Mutat. Res. - Genet. Toxicol. Environ. Mutagen.* **813**, 27–36 (2017).
361. Alhamdow, A. *et al.* DNA methylation of the cancer-related genes F2RL3 and AHRR is associated with occupational exposure to polycyclic aromatic hydrocarbons. *Carcinogenesis* **1–10** (2018). doi:10.1093/carcin/bgy059
362. Yang, P. *et al.* Prenatal urinary polycyclic aromatic hydrocarbon metabolites, global DNA methylation in cord blood, and birth outcomes: A cohort study in China. *Environ. Pollut.* 396–405 (2017). doi:10.1016/j.envpol.2017.11.082
363. White, A. J. *et al.* Exposure to multiple sources of polycyclic aromatic hydrocarbons and breast cancer incidence. *Environ. Int.* **XX**, XX–XX (2016).
364. Kim, Y. H., Lee, Y. S., Lee, D. H. & Kim, D. S. Polycyclic aromatic hydrocarbons are associated with insulin receptor substrate 2 methylation in adipose tissues of Korean women. *Environ. Res.* **150**, 47–51 (2016).
365. Kim, Y. H., Lee, Y. S., Lee, D. H. & Kim, D. S. Polycyclic aromatic hydrocarbons are associated with insulin receptor substrate 2 methylation in adipose tissues of Korean women. *Environ. Res.* (2016). doi:10.1016/j.envres.2016.05.043
366. Madrigano, J. *et al.* Prolonged Exposure to Particulate Pollution, Genes Associated with Glutathione Pathways, and DNA Methylation in a Cohort of Older Men. *Environ. Health Perspect.* **119**, 977–982 (2011).
367. Tarantini, A. *et al.* Relative contribution of DNA strand breaks and DNA adducts to the genotoxicity of benzo[a]pyrene as a pure compound and in complex mixtures. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **671**, 67–75 (2009).
368. Byun, H.-M. *et al.* Evolutionary age of repetitive element subfamilies and sensitivity of DNA methylation to airborne pollutants. *Part. Fibre Toxicol.* **10**, 28 (2013).
369. Hou, L. *et al.* Ambient PM exposure and DNA methylation in tumor suppressor genes: a cross-sectional study. *Part Fibre Toxicol* **8**, 25 (2011).

370. Callahan, C. L. *et al.* Lifetime exposure to ambient air pollution and methylation of tumor suppressor genes in breast tumors. *Environ. Res.* **161**, 418–424 (2018).
371. Bind, M.-A. *et al.* Air pollution and gene-specific methylation in the Normative Aging Study View supplementary material. *Epigenetics* **9**, 448–458 (2014).
372. Kohli, A. *et al.* Secondhand smoke in combination with ambient air pollution exposure is associated with increased CpG methylation and decreased expression of IFN- γ in T effector cells and Foxp3 in T regulatory cells in children. *Clin. Epigenetics* **4**, 17 (2012).
373. De Prins, S. *et al.* Influence of ambient air pollution on global DNA methylation in healthy adults: A seasonal follow-up. *Environ. Int.* **59**, 418–424 (2013).
374. Peluso, M. *et al.* DNA methylation differences in exposed workers and nearby residents of the Ma Ta phut industrial estate, rayong, Thailand. *Int. J. Epidemiol.* **41**, 1753–1760 (2012).
375. Salam, M. T. *et al.* Genetic and Epigenetic Variations in Inducible Nitric Oxide Synthase Promoter, Particulate Pollution and Exhaled Nitric Oxide in Children. *J Allergy Clin Immunol* **129**, 232–239 (2012).
376. Baccarelli, A. *et al.* Rapid DNA methylation changes after exposure to traffic particles. *Am. J. Respir. Crit. Care Med.* **179**, 572–8 (2009).
377. Guo, L. *et al.* Effects of short-term exposure to inhalable particulate matter on DNA methylation of tandem repeats. *Environ. Mol. Mutagen.* **55**, 322–335 (2014).
378. Breton, C. V. *et al.* Particulate Matter, DNA Methylation in Nitric Oxide Synthase, and Childhood Respiratory Disease. *Child. Heal.* **120**, 1320–1326 (2012).
379. Lepeule, J. *et al.* Epigenetic Influences on Associations Between Air Pollutants and Lung Function in Elderly Men: The Normative Aging Study. *Environ. Health Perspect.* **122**, 566–572 (2014).
380. Besingi, W. & Johansson, A. Smoke-related DNA methylation changes in the etiology of human disease. *Hum. Mol. Genet.* **23**, 2290–2297 (2014).
381. Guillen, M. D., Sopelana, P. & Partearroyo, M. A. Food as a Source of Polycyclic Aromatic Carcinogens. *Rev. Environ. Health* (1997). doi:10.1515/REVEH.1997.12.3.133
382. SCF (Scientific Committee on Food). Polycyclic Aromatic Hydrocarbons – Occurrence in foods, dietary exposure and health effects. (2002). at <http://europa.eu.int/comm/food/fs/sc/scf/index_en.html>
383. FAO/WHO. Dietary exposure assessment of chemicals in food. Report of a Joint FAO/WHO Consultation. 1–88 (2005).
384. Lijinsky, W. & Ross, a E. Production of carcinogenic polynuclear hydrocarbons in the cooking of food. *Food Cosmet. Toxicol.* **5**, 343–347 (1967).
385. Paris, A., Ledauphin, J., Poinot, P. & Gaillard, J.-L. Polycyclic aromatic hydrocarbons in fruits and vegetables: Origin, analysis, and occurrence. *Environ. Pollut.* **234**, 96–106 (2018).
386. SCF (Scientific Committee on Food). Opinion of the Scientific Committee on Food on the risks to human health of Polycyclic Aromatic Hydrocarbons in food. 1–84 (2002).
387. Zelinkova, Z. & Wenzl, T. The Occurrence of 16 EPA PAHs in Food – A Review. *Polycycl. Aromat. Compd.* **6638**, 1–37 (2015).
388. Yebra-Pimentel, I., Fernandez-Gonzalez, R., Martinez-Carballo, E. & Simal-Gandara, J. A Critical

- Review about the Health Risk Assessment of PAHs and Their Metabolites in Foods. *Crit. Rev. Food Sci. Nutr.* **55**, 1383–1405 (2013).
389. Stołyhwo, A. & Sikorski, Z. E. Polycyclic aromatic hydrocarbons in smoked fish - A critical review. *Food Chem.* **91**, 303–311 (2005).
390. Bansal, V. & Kim, K.-H. Review of PAH contamination in food products and their health hazards. *Environ. Int.* **84**, 26–38 (2015).
391. Fähnrich, K. A., Pravda, M. & Guilbault, G. G. Immunochemical detection of polycyclic aromatic hydrocarbons (PAHs). *Anal. Lett.* **35**, 1269–1300 (2002).
392. Domingo, J. L. & Nadal, M. Human dietary exposure to polycyclic aromatic hydrocarbons: A review of the scientific literature. *Food Chem. Toxicol.* **86**, 144–153 (2015).
393. EC (European Commission). COMMISSION REGULATION (EC) No 208/2005 of 4 February 2005 amending Regulation (EC) No 466/2001 as regards polycyclic aromatic hydrocarbons. *Off. J. Eur. Union* **L 34**, 3–5 (2005).
394. FAO/WHO. Summary and conclusions of the sixty-fourth meeting of the Joint FAO/WHO expert Committee on Food Additives (JECFA). *Evaluation* 1–47 (2005). at <http://www.who.int/ipcs/food/jecfa/summaries/summary_report_64_final.pdf>
395. EC (European Commission). 2005/108 of 4 February 2005 on the further investigation into the levels of polycyclic aromatic hydrocarbons in certain foods. *Off. J. Eur. Union* **L34**, 43–45 (2006).
396. EC (European Commission). Commission regulation (EU) No 835/2011 of 19 August 2011 amending Regulation (EC) No 1881/2006 as regards maximum levels for polycyclic aromatic hydrocarbons in foodstuffs. *Off. J. Eur. Union* 4–8 (2011). doi:10.3000/17252555.L_2011.006.eng
397. FAO/WHO. *Safety evaluation of certain contaminants in food. Safety evaluation of certain contaminants in food. WHO Food additives series: 55* (2006). doi:10.1016/j.ijfoodmicro.2007.01.001
398. EC (European Commission). *Commission Regulation (EC) No 1881/2006 of 19 December 2006 setting maximum levels for certain contaminants in foodstuffs. Official Journal* (2006).
399. IARC & Group, T. W. Other data relevant to an evaluation of carcinogenicity and its mechanisms. *IARC Monogr. Eval. Carcinog. Risks to Humans* **92**, 512–753 (2010).
400. Traoré, T. *et al.* To which mixtures are French pregnant women mainly exposed? A combination of the second French total diet study with the EDEN and ELFE cohort studies. *Food Chem. Toxicol.* **111**, 310–328 (2017).
401. Lamichhane, D. K. *et al.* Impact of prenatal exposure to polycyclic aromatic hydrocarbons from maternal diet on birth outcomes: a birth cohort study in Korea. *Public Health Nutr.* 1–10 (2016). doi:10.1017/S1368980016000550
402. Withey, J. R., Shedden, J., Law, F. C. P. & Abedini, S. Distribution of benzo[a]pyrene in pregnant rats following inhalation exposure and a comparison with similar data obtained with pyrene. *J. Appl. Toxicol.* **13**, 193–202 (1993).
403. Çok, I. *et al.* Analysis of human milk to assess exposure to PAHs, PCBs and organochlorine pesticides in the vicinity Mediterranean city Mersin, Turkey. *Environ. Int.* **40**, 63–69 (2012).
404. Duan, X. *et al.* Dietary intake polycyclic aromatic hydrocarbons (PAHs) and associated cancer

- risk in a cohort of Chinese urban adults: Inter- and intra-individual variability. *Chemosphere* **144**, 2469–2475 (2016).
405. Li, J. *et al.* Quantitatively assessing the health risk of exposure to PAHs from intake of smoked meats. *Ecotoxicol. Environ. Saf.* **124**, 91–95 (2016).
406. Rothman, N. *et al.* Association of PAH-DNA Adducts in Peripheral White Blood Cells with Dietary Exposure to Polyaromatic Hydrocarbons. *Environ. Health Perspect.* **99**, 265–267 (1993).
407. Ma, Y. & Harrad, S. Spatiotemporal analysis and human exposure assessment on polycyclic aromatic hydrocarbons in indoor air, settled house dust, and diet: A review. *Environ. Int.* **84**, 7–16 (2015).
408. Bansal, V., Kumar, P., Kwon, E. E. & Kim, K.-H. Review of the quantification techniques for polycyclic aromatic hydrocarbons (PAHs) in food products. *Crit. Rev. Food Sci. Nutr.* **57**, 3297–3312 (2017).
409. Food and Agriculture Organization, European Food Safety Authority & World Health Organization. Towards a harmonised Total Diet Study approach: a guidance document. *EFSA J.* **9**, 1–66 (2011).
410. Kazerouni, N., Sinha, R., Hsu, C.-H., Greenberg, A. & Rothman, N. Analysis of 200 food items for benzo[a]pyrene and estimation of its intake in an epidemiologic study. *Food Chem. Toxicol.* **39**, 423–436 (2001).
411. Palli, D. *et al.* Biomarkers of dietary intake of micronutrients modulate DNA adduct levels in healthy adults. *Carcinogenesis* **24**, 739–746 (2003).
412. Palli, D. *et al.* Diet, metabolic polymorphisms and dna adducts: the EPIC-Italy cross-sectional study. *Int. J. cancer* **87**, 444–51 (2000).
413. Sram, R. J. *et al.* Effect of vitamin levels on biomarkers of exposure and oxidative damage-The EXPAH study. *Mutat. Res. - Genet. Toxicol. Environ. Mutagen.* **672**, 129–134 (2009).
414. Luzardo, O. P. *et al.* Influence of the method of production of eggs on the daily intake of polycyclic aromatic hydrocarbons and organochlorine contaminants: An independent study in the Canary Islands (Spain). *Food Chem. Toxicol.* **60**, 455–462 (2013).
415. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* **9**, 436–447 (2016).
416. Joubert, B. R. *et al.* DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am. J. Hum. Genet.* **98**, 680–96 (2016).
417. Allione, A. *et al.* Novel epigenetic changes unveiled by monozygotic twins discordant for smoking habits. *PLoS One* **10**, e0128265 (2015).
418. Breitling, L. P., Yang, R., Korn, B., Burwinkel, B. & Brenner, H. Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. *Am. J. Hum. Genet.* **88**, 450–457 (2011).
419. Breton, C. V *et al.* Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. *PLoS One* **9**, e99716 (2014).
420. Chhabra, D. *et al.* Fetal lung and placental methylation is associated with in utero nicotine exposure. *Epigenetics* **9**, 1473–84 (2014).
421. Dogan, M. V *et al.* The effect of smoking on DNA methylation of peripheral blood mononuclear

- cells from African American women. *BMC Genomics* **15**, 151 (2014).
422. Elliott, H. R. *et al.* Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin. Epigenetics* **6**, 4 (2014).
423. Fasanelli, F. *et al.* Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun.* **6**, 10192 (2015).
424. Freeman, J. R., Chu, S., Hsu, T. & Huang, Y. Epigenome-wide association study of smoking and DNA methylation in non-small cell lung neoplasms. *Oncotarget* (2016).
425. Guida, F. *et al.* Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum. Mol. Genet.* **24**, 2349–2359 (2015).
426. Harlid, S., Xu, Z., Panduri, V., Sandler, D. P. & Taylor, J. A. CpG sites associated with cigarette smoking: Analysis of epigenome-wide data from the Sister Study. **122**, 673–678 (2014).
427. Ivorra, C. *et al.* DNA methylation patterns in newborns exposed to tobacco in utero. *J. Transl. Med.* **13**, 25 (2015).
428. Joubert, B. R. *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* **120**, 1425–1431 (2012).
429. Kupers, L. K. *et al.* DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *Int. J. Epidemiol.* 1–14 (2015). doi:10.1093/ije/dyv048
430. Lee, K. W. K. *et al.* Prenatal Exposure to Maternal Cigarette Smoking and DNA Methylation : Epigenome-Wide Association in a Discovery Sample of Adolescents and Replication in an Independent Cohort at Birth through 17 Years of Age. *Env. Heal. Perspect* **123**, 193–199 (2015).
431. Li, S. *et al.* Causal effect of smoking on DNA methylation in peripheral blood: a twin and family study. *Clin. Epigenetics* **10**, 18 (2018).
432. Maccani, J. Z. J., Koestler, D. C., Houseman, E. A., Marsit, C. J. & Kelsey, K. T. Placental DNA methylation alterations associated with maternal tobacco smoking at the RUNX3 gene are also associated with gestational age. *Epigenomics* **5**, 619–30 (2013).
433. Markunas, C. A. *et al.* Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* (2014). doi:10.1289/ehp.1307892
434. Monick, M. M. *et al.* Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **159 B**, 141–151 (2012).
435. Morales, E. *et al.* Genome-wide DNA methylation study in human placenta identifies novel loci associated with maternal smoking during pregnancy. *Int. J. Epidemiol.* dyw196 (2016). doi:10.1093/ije/dyw196
436. Philibert, R. A., Beach, S. R. H. & Brody, G. H. Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. *Epigenetics* **7**, 1331–8 (2012).
437. Philibert, R. A., Beach, S. R. H., Lei, M.-K. & Brody, G. H. Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. *Clin. Epigenetics* **5**, 19 (2013).

438. Richmond, R. C. *et al.* Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum. Mol. Genet.* **24**, 2201–2217 (2015).
439. Richmond, R. C., Suderman, M., Langdon, R., Relton, C. L. & Davey Smith, G. DNA methylation as a marker for prenatal smoke exposure in adults. *Int. J. Epidemiol.* **47**, 1120–1130 (2018).
440. Shenker, N. S. *et al.* DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology* **24**, 712–6 (2013).
441. Stueve, T. R. *et al.* Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. *Hum. Mol. Genet.* **26**, 3014–3027 (2017).
442. Sun, Y. V. *et al.* Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. *Hum. Genet.* **132**, 1027–37 (2013).
443. Suter, M. *et al.* Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. *Epigenetics* **6**, 1284–94 (2011).
444. Tzaprouni, L. G. *et al.* Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* **9**, 1382–1396 (2014).
445. Wan, E. S. *et al.* Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum. Mol. Genet.* **21**, 3073–3082 (2012).
446. Witt, S. H. *et al.* Impact on birth weight of maternal smoking throughout pregnancy mediated by DNA methylation. *BMC Genomics* **19**, 290 (2018).
447. Zaghlool, S. B. *et al.* Association of DNA methylation with age, gender, and smoking in an Arab population. (2011). doi:10.1186/s13148-014-0040-6
448. Zeilinger, S. *et al.* Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *PLoS One* **8**, e63812 (2013).
449. Zhu, X. *et al.* Genome-wide analysis of DNA methylation and cigarette smoking in a Chinese population. *Env. Heal. Perspect* **124**, 966–973 (2016).
450. Lee, M. K., Hong, Y., Kim, S.-Y., London, S. J. & Kim, W. J. DNA methylation and smoking in Korean adults: epigenome-wide association study. *Clin. Epigenetics* **8**, 103 (2016).
451. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
452. Liu, C. *et al.* A DNA methylation biomarker of alcohol consumption. *Mol. Psychiatry* **23**, 422–433 (2018).
453. Dixon, H. M. *et al.* Silicone wristbands compared with traditional polycyclic aromatic hydrocarbon exposure assessment methods. *Anal. Bioanal. Chem.* **410**, 3059–3071 (2018).

9 Appendices

9.1 Appendix 1 - Chapter 3 Supporting Tables and Figures

9.1.1 Model Results: Control mice vs mice exposed to low dose of B[a]P (25 mg/kg b.w.)

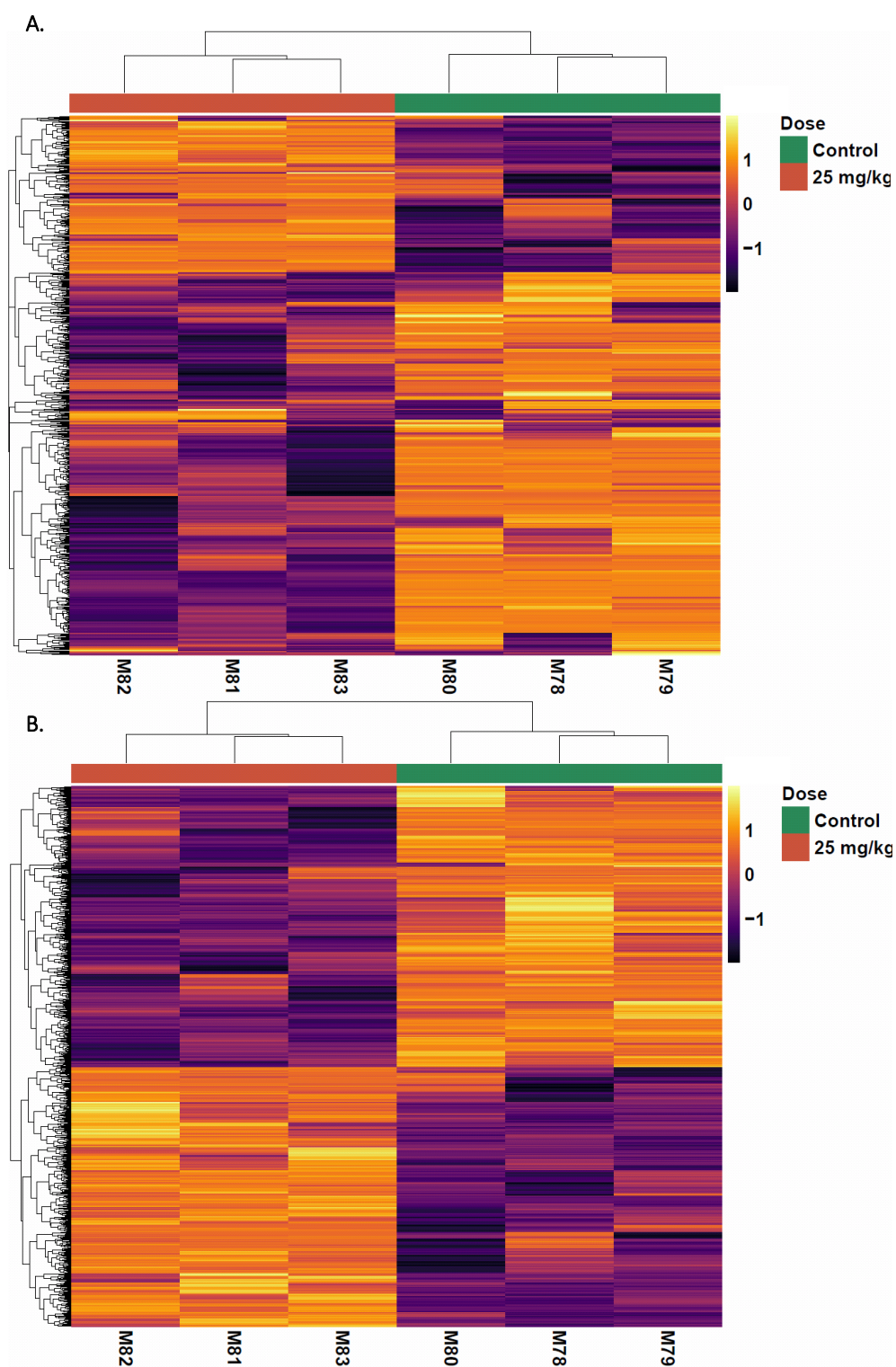


Figure 9.1 Heatmaps of the 699 differentially methylated CpG sites (A) and the 1910 differentially methylated 500 bp windows (B) from the control vs low dose model.

Table 9.1 Table of Fisher’s test results comparing the number of DMWs (N = 1910) and DMCs (N = 699) to all tested windows (N = 152,720) and CpG sites (N = 38,874) at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred.

Genomic Region	Differentially Methylated Windows			Differentially Methylated CpG Sites		
	Odds Ratio	Confidence Interval	P Value	Odds Ratio	Confidence Interval	P Value
3' UTR	1.29	0.86 – 1.88	0.19	1.56	0.86 – 2.61	0.10
5' UTR	0.16	0.064 – 0.33	1.69×10^{-11}	2.06	0.05 – 12.55	0.39
Exon	0.47	0.36 – 0.61	4.50×10^{-11}	0.88	0.67 – 1.14	0.38
Intergenic	1.82	1.65 – 2.01	2.92×10^{-31}	1.42	1.19 – 1.68	8.20×10^{-5}
Intron	1.43	1.30 – 1.58	1.29×10^{-12}	1.46	1.23 – 1.73	1.40×10^{-5}
Non-coding	0.36	0.098 – 0.93	0.03	1.50	0.40 – 3.93	0.35
Promoter	0.17	0.12 – 0.22	1.04×10^{-71}	0.82	0.33 – 1.71	0.73
TTS	1.63	1.19 – 2.18	2.07×10^{-3}	0.66	0.21 – 1.56	0.46
CpG Island	0.07	0.028 – 0.13	7.65×10^{-40}	1.38	0.28 – 4.15	0.48
LINE	1.04	0.89 – 1.22	0.60	0.71	0.56 – 0.89	2.53×10^{-3}
SINE	1.56	1.28 – 1.88	1.20×10^{-5}	0.72	0.47 – 1.06	0.10
LTR	1.12	0.94 – 1.32	0.21	0.59	0.47 – 0.74	1.33×10^{-6}
Other	1.48	1.09 – 1.96	0.01	0.90	0.39 – 1.80	1.00

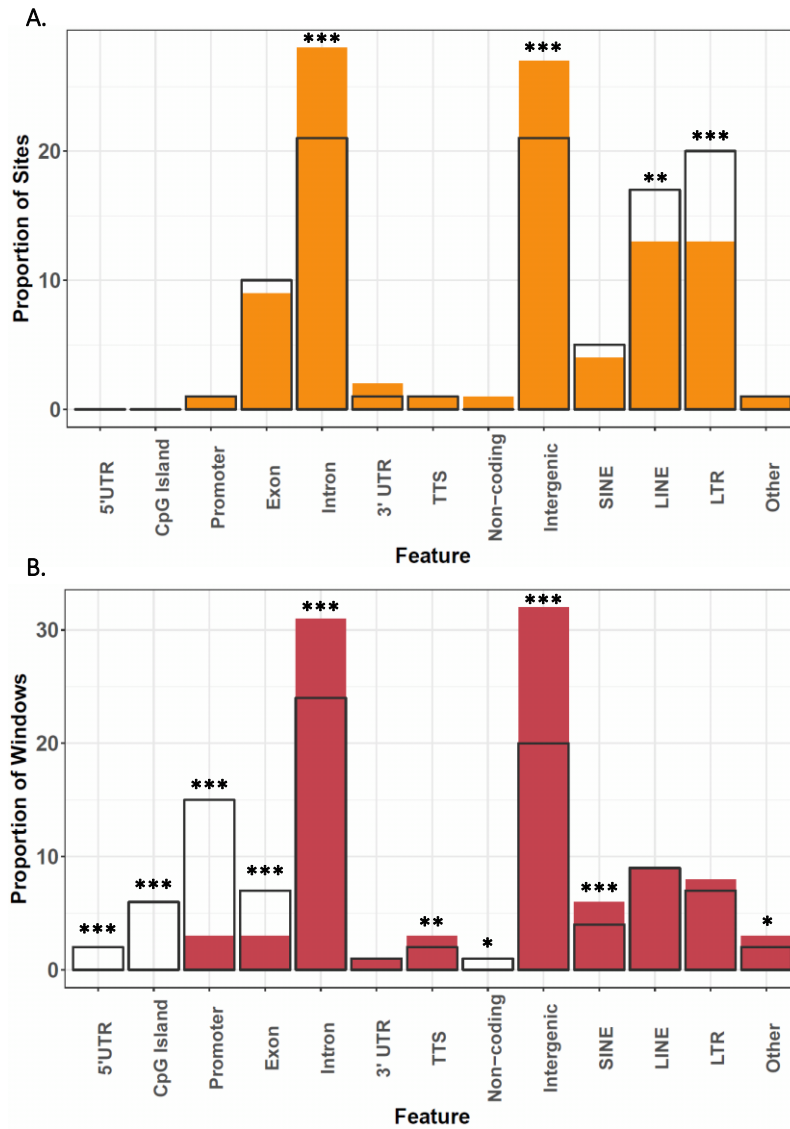


Figure 9.2 Comparison of the genomic distribution of differentially methylated sites (N = 699) and all tested sites (N = 38,874) (A) and of differentially methylated windows (N = 1910) and all tested windows (N = 152,691) (B). The coloured bars show the proportion of significant sites/windows, the grey outline bars show the proportion of all sites/windows tested, i.e. the expected distribution. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.005$ following Fisher's Exact test.

Table 9.2 Table of Fisher’s test results comparing the number of hypermethylation changes (DMWs: N = 993; DMCs: N = 478) to hypomethylation changes (DMWS: N = 917; DMCs: N = 221) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected.

Genomic Region	Differentially Methylated Windows			Differentially Methylated CpGs		
	Odds Ratio	Confidence Interval	P Value	Odds Ratio	Confidence Interval	P Value
3' UTR	1.97	0.84 – 4.96	0.13	1.97	0.28 – 3.49	0.13
5' UTR	0.69	0.1 – 4.10	0.72	0.69	0.01 – Inf	0.72
Exon	0.95	0.56 – 1.61	0.90	0.95	0.55 – 1.76	0.90
Intergenic	0.99	0.82 – 1.21	0.96	0.99	1.12 – 2.48	0.96
Intron	1.02	0.84 – 1.25	0.84	1.02	0.72 – 1.52	0.84
Non-coding	0.92	0.07 – 12.76	1.00	0.92	0.11 – 73.25	1.00
Promoter	0.92	0.52 – 1.63	0.79	0.92	0.34 – 129.15	0.79
TTS	1.19	0.64 – 2.24	0.56	1.19	0.08 – 8.34	0.56
CpG Island	1.54	0.30 – 9.96	0.73	1.54	0.05 – 54.78	0.73
LINE	0.57	0.41 – 0.80	6.50×10^{-4}	0.57	0.39 – 1.02	0.05
SINE	1.23	0.83 – 1.83	0.30	1.23	0.24 – 1.35	0.30
LTR	1.35	0.95 – 1.93	0.09	1.35	0.41 – 1.09	0.09
Other	1.00	0.55 – 1.83	1.00	1.00	0.79 - Inf	1.00

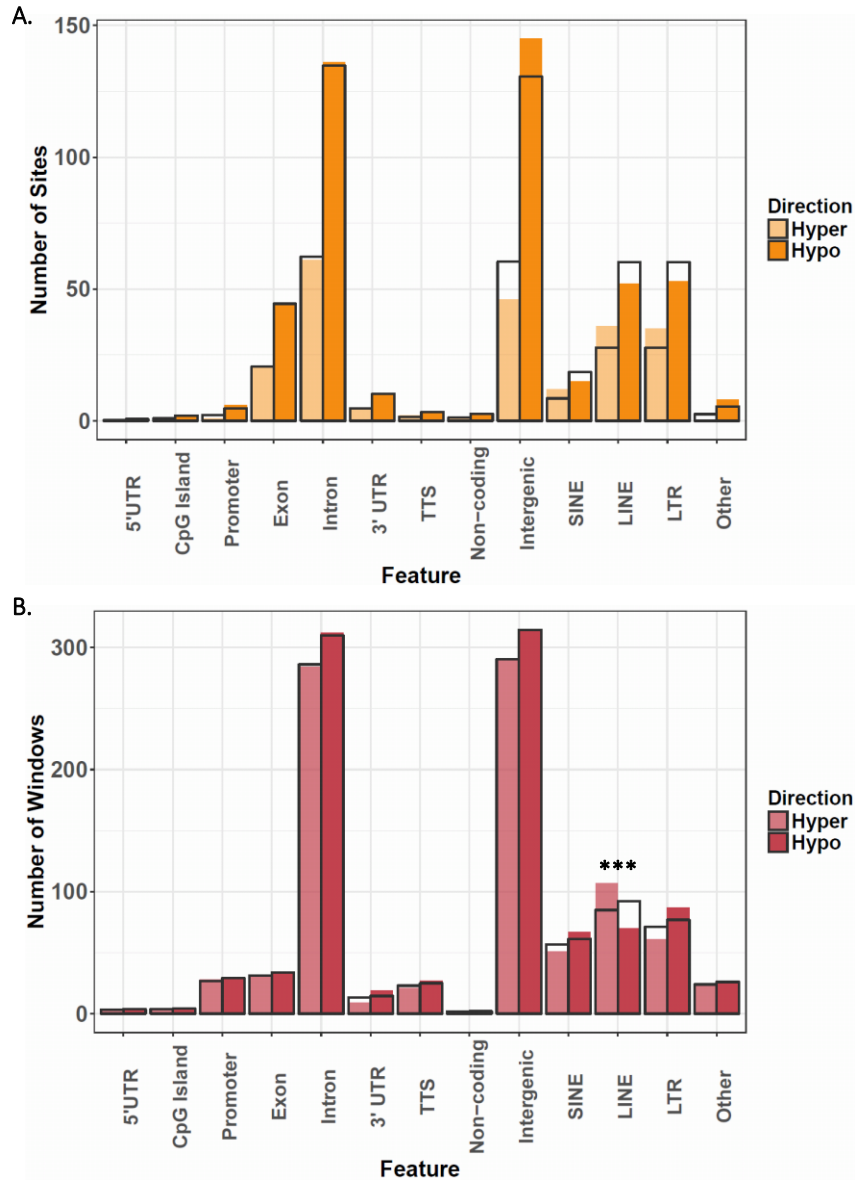


Figure 9.3 Comparison of the genomic distribution of hypermethylated (N = 145) and hypomethylated (N = 285) sites (A), and hypermethylated (N = 660) and hypomethylated (N = 1120) windows (B). The coloured bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. In all cases, these results are for the treated vs untreated models. For all plots, *** indicates $p < 0.005$ following Fisher's Exact test.

9.1.2 Model Results: Control mice vs mice exposed to medium dose of B[a]P (50 mg/kg b.w.)

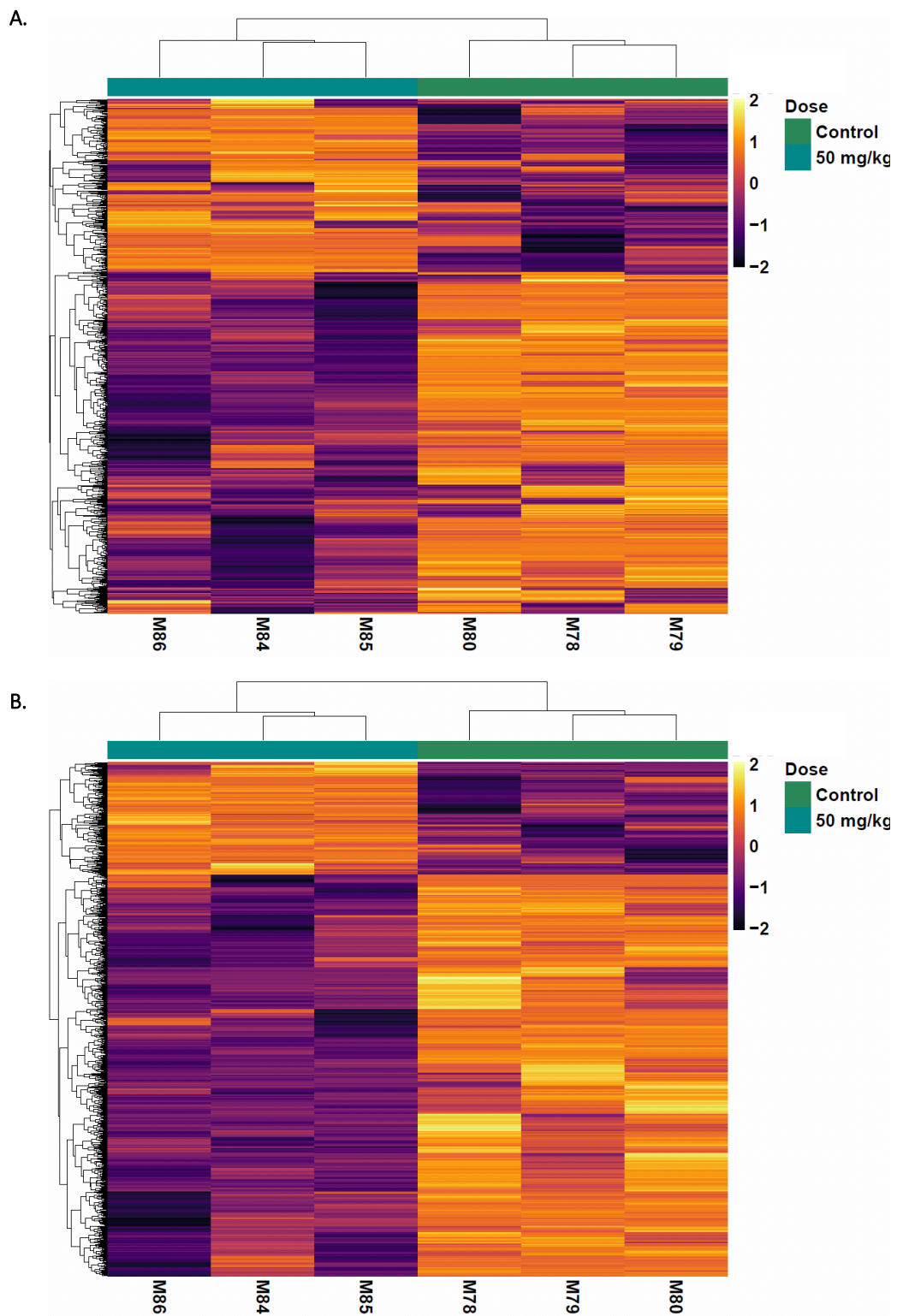


Figure 9.4 Heatmaps of the 768 differentially methylated CpG sites (A) and the 2671 differentially methylated 500 bp windows (B) from the control vs medium dose model.

Table 9.3 Table of Fisher’s test results comparing the number of DMWs (N = 2671) and DMCs (N = 768) to all tested windows (N = 152,720) and CpG sites (N = 38,874) at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred.

Genomic Region	Differentially Methylated Windows			Differentially Methylated CpG Sites		
	Odds Ratio	Confidence Interval	P Value	Odds Ratio	Confidence Interval	P Value
3' UTR	1.26	0.88 – 1.74	0.17	0.94	0.45 – 1.75	1.00
5' UTR	0.08	0.03 – 0.19	5.70×10^{-20}	1.88	0.05 – 11.41	0.42
Exon	0.61	0.50 – 0.73	2.05×10^{-8}	0.84	0.64 – 1.08	0.19
Intergenic	1.65	1.51 – 1.79	2.57×10^{-29}	1.33	1.12 – 1.56	9.31×10^{-4}
Intron	1.54	1.42 – 1.67	7.58×10^{-24}	1.24	1.05 – 1.47	0.01
Non-coding	0.39	0.14 – 0.85	0.01	1.36	0.37 – 3.57	0.55
Promoter	0.14	0.11 – 0.18	2.78×10^{-109}	1.18	0.58 – 2.14	0.51
TTS	1.21	0.89 – 1.60	0.21	1.21	0.57 – 2.26	0.48
CpG Island	0.04	0.02 – 0.08	3.38×10^{-60}	2.52	0.90 – 5.68	0.04
LINE	1.14	1.00 – 1.29	0.05	0.66	0.53 – 0.83	1.66×10^{-4}
SINE	1.56	1.32 – 1.83	2.72×10^{-7}	0.93	0.65 – 1.30	0.74
LTR	1.17	1.01 – 1.35	0.03	0.79	0.64 – 0.96	0.02
Other	1.52	1.18 – 1.93	9.30×10^{-4}	1.34	0.71 – 2.33	0.33

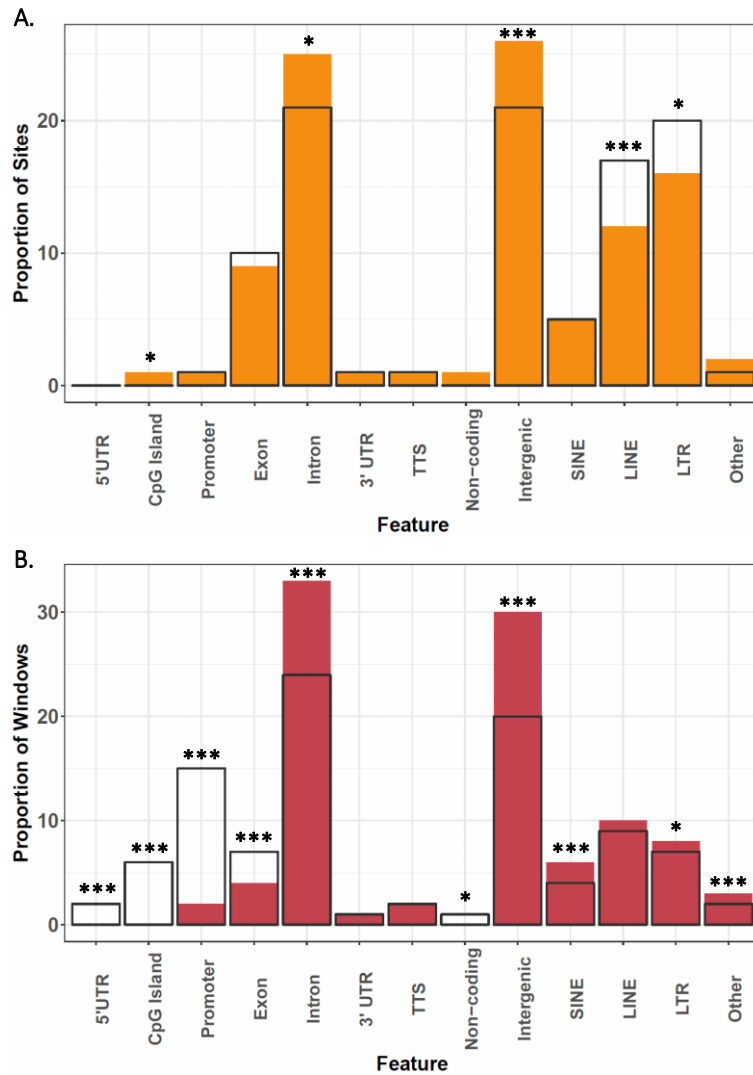


Figure 9.5 Comparison of the genomic distribution of differentially methylated sites (N = 768) and all tested sites (N = 38,874) (A) and of differentially methylated windows (N = 2671) and all tested windows (N = 152,691) (B). The coloured bars show the proportion of significant sites/windows, the grey outline bars show the proportion of all sites/windows tested, i.e. the expected distribution. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.005$ following Fisher's Exact test.

Table 9.4 Table of Fisher’s test results comparing the number of hypermethylation changes (DMWS: N = 584; DMCs: N = 263) to hypomethylation changes (DMWs: N = 2087; DMCs: N = 505) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected.

Genomic Region	Differentially Methylated Windows			Differentially Methylated CpGs		
	Odds Ratio	Confidence Interval	P Value	Odds Ratio	Confidence Interval	P Value
3' UTR	1.97	0.84 – 4.96	0.13	1.97	0.28 – 3.49	0.13
5' UTR	0.69	0.1 – 4.10	0.72	0.69	0.01 – Inf	0.72
Exon	0.95	0.56 – 1.61	0.90	0.95	0.55 – 1.76	0.90
Intergenic	0.99	0.82 – 1.21	0.96	0.99	1.12 – 2.48	0.96
Intron	1.02	0.84 – 1.25	0.84	1.02	0.72 – 1.52	0.84
Non-coding	0.92	0.07 – 12.76	1.00	0.92	0.11 – 73.25	1.00
Promoter	0.92	0.52 – 1.63	0.79	0.92	0.34 – 129.15	0.79
TTS	1.19	0.64 – 2.24	0.56	1.19	0.08 – 8.34	0.56
CpG Island	1.54	0.30 – 9.96	0.73	1.54	0.05 – 54.78	0.73
LINE	0.57	0.41 – 0.80	6.50×10^{-4}	0.57	0.39 – 1.02	0.05
SINE	1.23	0.83 – 1.83	0.30	1.23	0.24 – 1.35	0.30
LTR	1.35	0.95 – 1.93	0.09	1.35	0.41 – 1.09	0.09
Other	1.00	0.55 – 1.83	1.00	1.00	0.79 - Inf	1.00

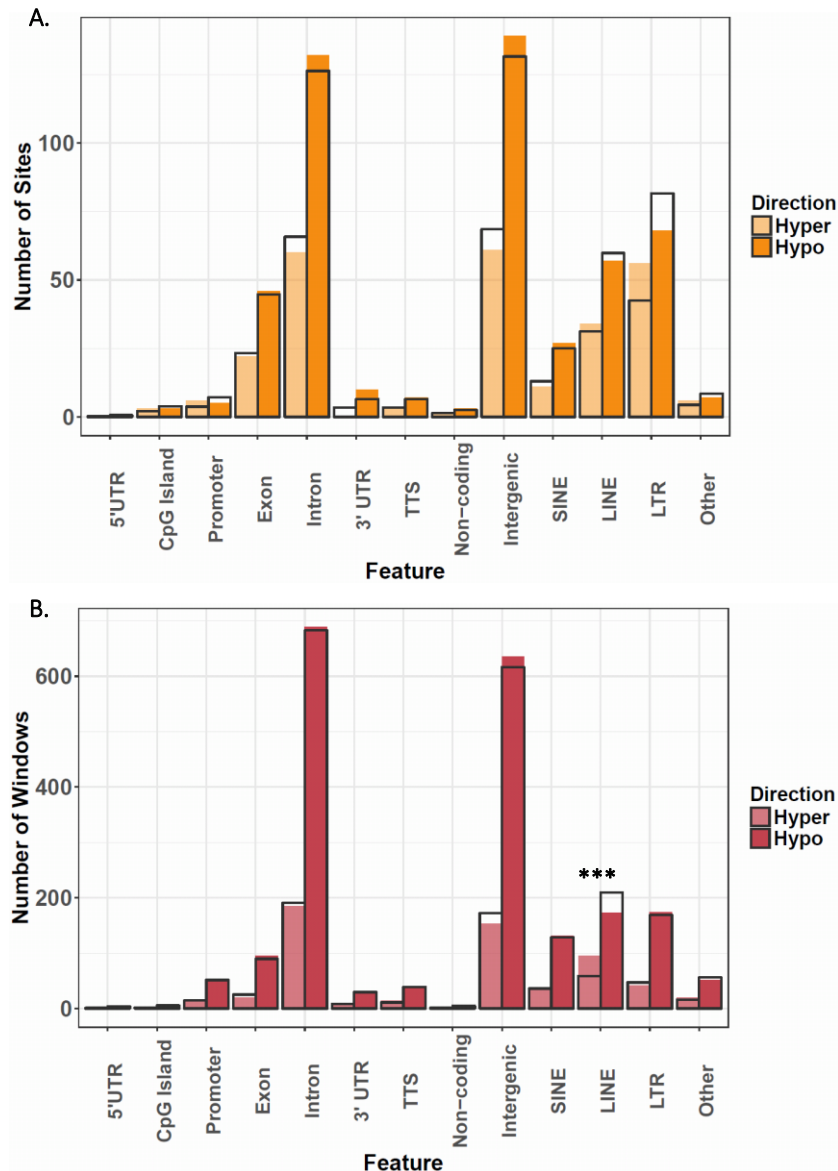


Figure 9.6 Comparison of the genomic distribution of hypermethylated (N = 263) and hypomethylated (N = 505) sites (A), and hypermethylated (N = 584) and hypomethylated (N = 2087) windows (B). The coloured bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.005$ following Fisher's Exact test.**

9.1.3 Model Results: Control mice vs mice exposed to a high dose of B[a]P (75 mg/kg b.w.)

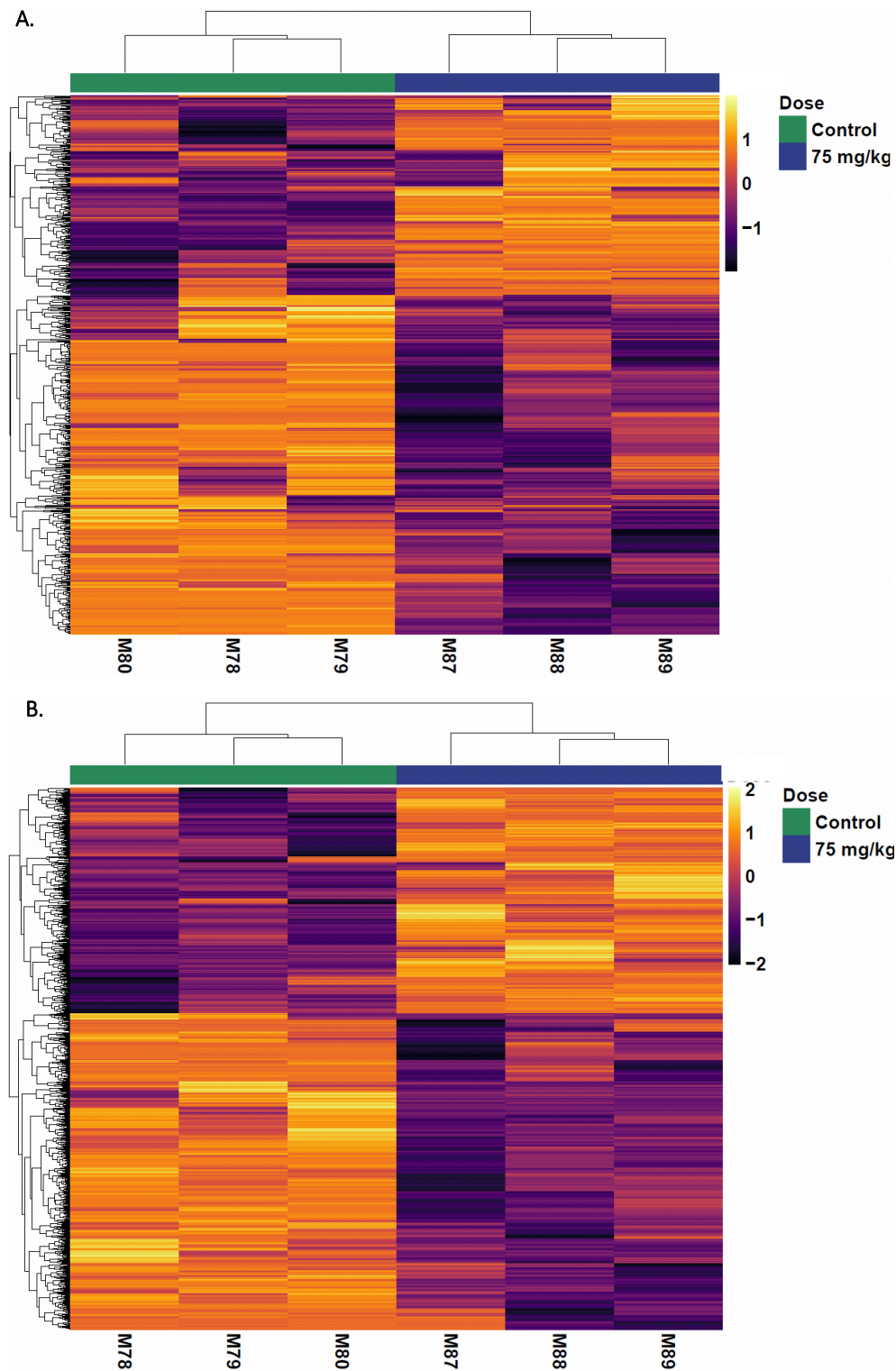


Figure 9.7 Heatmaps of the 664 differentially methylated CpG sites (A) and the 1952 differentially methylated 500 bp windows (B) from the control vs high dose model.

Table 9.5 Table of Fisher’s test results comparing the number of DMWs (N = 1952) and DMCs (N = 664) to all tested windows (N = 152,720) and CpG sites (N = 38,874) at various genomic regions. An OR < 1 indicates that less methylation changes than expected occurred at a given genomic region given the underlying distribution of all tested probes, while an OR > 1 indicates that more changes than expected occurred.

Genomic Region	Differentially Methylated Windows			Differentially Methylated CpG Sites		
	Odds Ratio	Confidence Interval	P Value	Odds Ratio	Confidence Interval	P Value
3' UTR	0.81	0.48 – 1.29	0.45	1.87	1.07 – 3.04	0.02
5' UTR	0.07	0.01 – 0.19	1.41×10^{-15}	0	0 – 8.59	1.00
Exon	0.59	0.47 – 0.74	5.55×10^{-7}	0.73	0.54 – 0.97	0.03
Intergenic	1.69	1.53 – 1.87	1.82×10^{-24}	1.52	1.28 – 1.81	2.46×10^{-6}
Intron	1.43	1.29 – 1.57	1.81×10^{-12}	1.36	1.14 – 1.62	6.71×10^{-4}
Non-coding	0.80	0.36 – 1.52	0.65	1.18	0.24 – 3.53	0.74
Promoter	0.16	0.12 – 0.21	4.05×10^{-74}	0.86	0.34 – 1.81	0.86
TTS	0.99	0.66 – 1.42	1.00	0.69	0.22 – 1.64	0.57
CpG Island	0.05	0.02 – 0.10	2.44×10^{-43}	1.94	0.52 – 5.12	0.16
LINE	1.38	1.19 – 1.58	1.11×10^{-5}	0.67	0.52 – 0.85	5.29×10^{-4}
SINE	1.48	1.21 – 1.79	1.11×10^{-5}	0.76	0.50 – 1.12	0.19
LTR	1.18	1.00 – 1.39	0.05	0.67	0.53 – 0.83	2.07×10^{-4}
Other	1.36	0.99 – 1.81	0.05	0.95	0.41 – 1.90	1.00

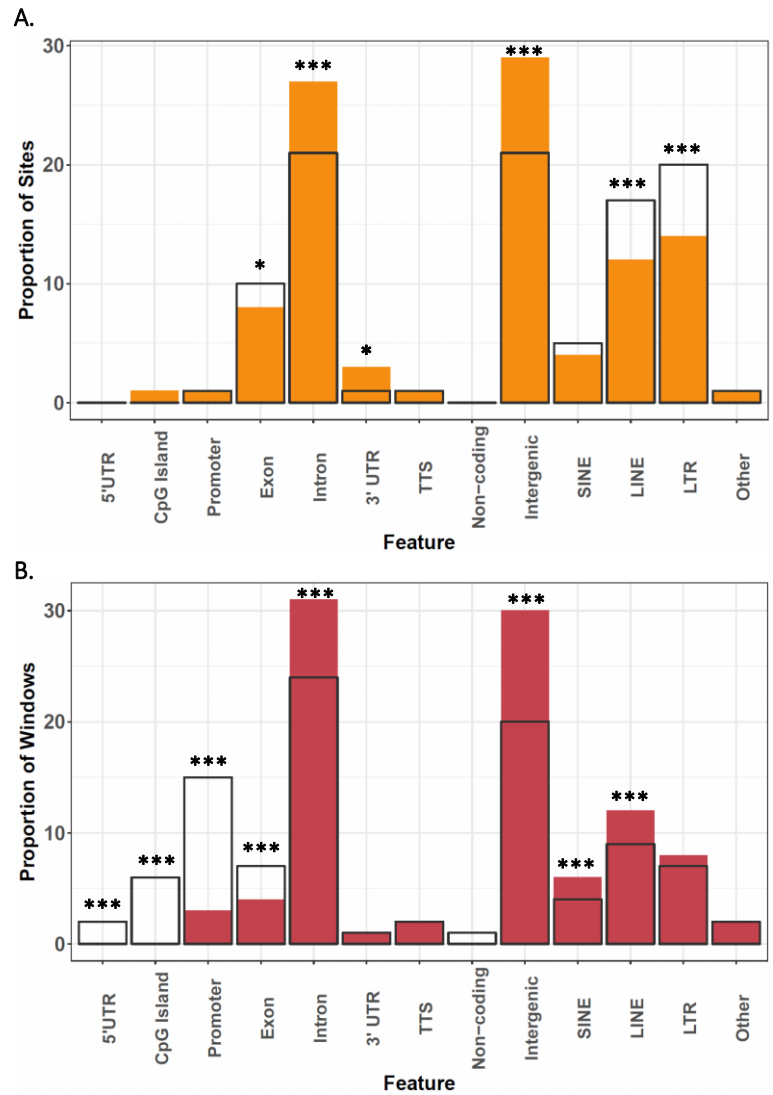


Figure 9.8 Comparison of the genomic distribution of differentially methylated sites (N = 664) and all tested sites (N = 38,874) (A) and of differentially methylated windows (N = 1952) and all tested windows (N = 152,691) (B). The coloured bars show the proportion of significant sites/windows, the grey outline bars show the proportion of all sites/windows tested, i.e. the expected distribution. In all cases, these results are for the treated vs untreated models. For all plots, * indicates $p < 0.05$ and *** indicates $p < 0.005$ following Fisher's Exact test.

Table 9.6 Table of Fisher’s test results comparing the number of hypermethylation changes (DMWS: N = 813; DMCs: N = 248) to hypomethylation changes (DMWs: N = 1139; DMCs: N = 416) compared to the overall ratio of hypermethylated to hypomethylated probes. An OR < 1 indicates that more hypermethylation changes occurred than expected compared to the overall ratio, an OR of > 1 indicates that more hypomethylation changes occurred than expected.

Genomic Region	Differentially Methylated Windows			Differentially Methylated CpG Sites		
	Odds Ratio	Confidence Interval	P Value	Odds Ratio	Confidence Interval	P Value
3' UTR	0.45	0.15 – 1.28	0.10	1.44	0.47 – 5.29	0.62
5' UTR	0.36	0.01 – 6.86	0.57	0.00	0 – Inf	1.00
Exon	1.06	0.66 – 1.72	0.82	1.25	0.66 – 2.43	0.55
Intergenic	0.88	0.72 – 1.07	0.19	1.35	0.93 – 1.96	0.11
Intron	1.10	0.90 – 1.35	0.35	1.24	0.85 – 1.81	0.28
Non-coding	1.43	0.30 – 8.86	0.74	Inf	0.25 – Inf	0.30
Promoter	1.52	0.84 – 2.86	0.17	0.79	0.13 – 5.46	0.72
TTS	1.98	0.84 – 5.17	0.13	0.89	0.10 – 10.77	1.00
CpG Island	Inf	0.84 – Inf	0.04	1.79	0.14 – 94.52	1.00
LINE	0.90	0.68 – 1.20	0.48	0.60	0.37 – 1.00	0.05
SINE	0.89	0.60 – 1.33	0.56	1.01	0.43 – 2.52	1.00
LTR	1.06	0.76 – 1.51	0.74	0.53	0.33 – 0.85	0.01
Other	0.96	0.52 – 1.82	0.88	0.99	0.19 – 0.45	1.00

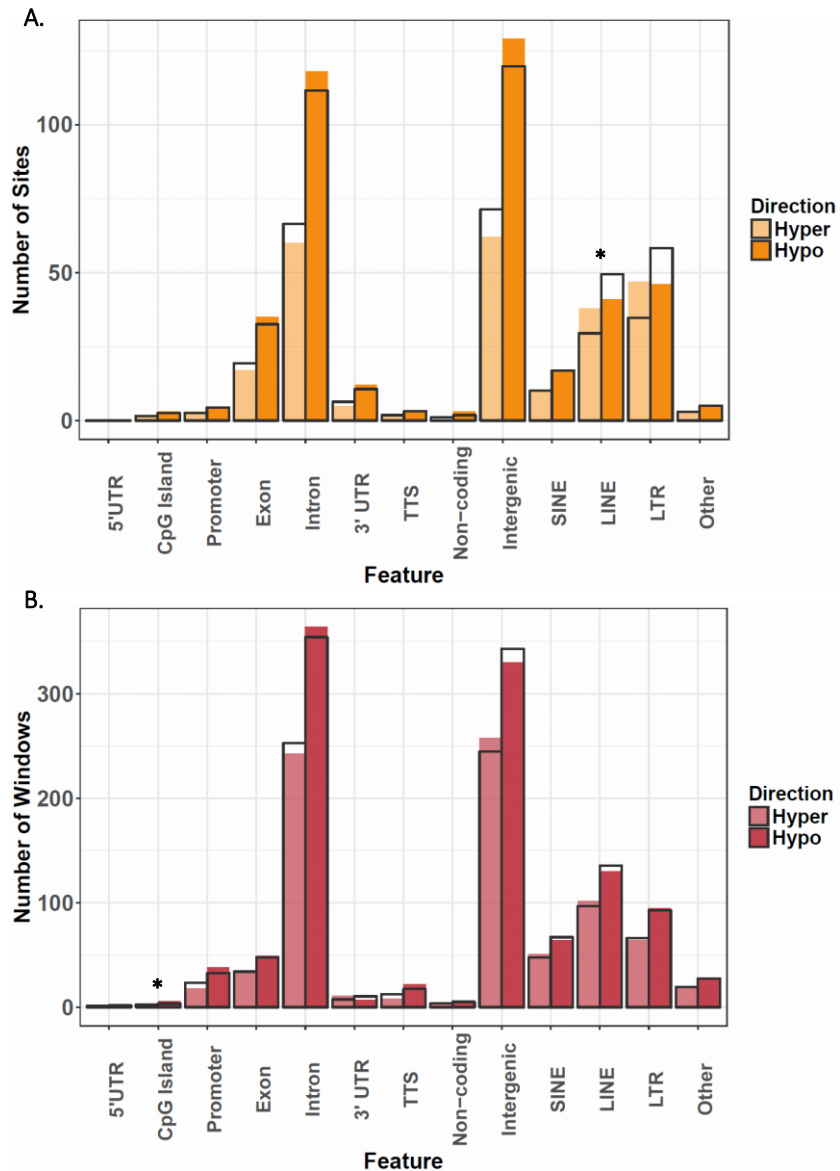


Figure 9.9 Comparison of the genomic distribution of hypermethylated (N = 248) and hypomethylated (N = 416) sites (A), and hypermethylated (N = 813) and hypomethylated (N = 1139) windows (B). The coloured bars show the number of significant probes, with the lighter and darker shades indicating hypermethylated and hypomethylated probes respectively. The grey bars indicate the expected distribution calculated based on the overall ratio of hypermethylated:hypomethylated results. In all cases, these results are for the treated vs untreated models. For all plots, *** indicates $p < 0.005$ following Fisher's Exact test.

9.2 Appendix 2 – Chapter 4 Supporting Tables and Figures

Table 9.7. Model results for the FDR significant ($p < 2.8 \times 10^{-5}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship between air PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status.

Probe ID	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-NL (N=132)		
	β Coefficient	95% Confidence Interval	P Value	β Coefficient	95% Confidence Interval	P Value	β Coefficient	95% Confidence Interval	P Value
cg00466488	0.353	0.278; 0.428	1.82E-20	0.069	-0.008; 0.146	0.08	0.467	0.049; 0.884	0.028
cg00695391	-0.064	-0.093; - 0.034	2.51E-05	-0.002	-0.04; 0.037	0.934	0.13	-0.076; 0.335	0.216
cg03317082	0.192	0.13; 0.255	1.92E-09	0.001	-0.072; 0.073	0.985	0.069	-0.34; 0.478	0.742
cg04117764	-0.119	-0.173; - 0.065	1.54E-05	-0.025	-0.103; 0.054	0.534	0.084	-0.28; 0.448	0.651
cg04273604	0.078	0.045; 0.111	2.64E-06	-0.023	-0.074; 0.027	0.364	0.165	-0.165; 0.494	0.327
cg06443644	-0.104	-0.15; -0.058	8.29E-06	0.084	0.02; 0.149	0.01	0.304	-0.064; 0.671	0.105
cg07826526	-0.118	-0.164; - 0.072	5.03E-07	-0.007	-0.069; 0.054	0.819	-0.186	-0.644; 0.272	0.426
cg12653146	-0.162	-0.21; -0.113	5.33E-11	0.029	-0.018; 0.077	0.222	0.113	-0.212; 0.438	0.495
cg16306900	-0.103	-0.147; - 0.058	5.23E-06	-0.086	-0.165; - 0.008	0.031	0.071	-0.247; 0.389	0.661
cg17900103	-0.086	-0.126; - 0.046	2.64E-05	0.026	-0.048; 0.1	0.485	0.419	-0.165; 1.002	0.159
cg18801028	-0.05	-0.074; - 0.027	2.80E-05	0.035	-0.005; 0.076	0.09	-0.184	-0.33; - 0.038	0.014
cg19925780	0.102	0.061; 0.142	9.52E-07	-0.028	-0.089; 0.033	0.368	-0.082	-0.383; 0.218	0.591

cg25876319	-0.049	-0.069; - 0.029	1.51E-06	0.037	0.002; 0.071	0.037	-0.223	-0.398; - 0.049	0.012
cg03219514	-0.144	-0.208; - 0.081	8.29E-06	-0.208	-0.332; - 0.084	0.001	-0.029	-0.424; 0.366	0.886
cg05703053	-0.184	-0.263; - 0.105	5.04E-06	0.106	-0.022; 0.234	0.104	-0.063	-0.612; 0.486	0.823
cg06856378	-0.299	-0.412; - 0.185	2.32E-07	0.313	0.157; 0.469	0	NA	NA; NA	NA
cg07480373	0.08	0.047; 0.113	1.98E-06	0.002	-0.052; 0.056	0.937	-0.062	-0.299; 0.174	0.605
cg08157672	0.141	0.08; 0.203	7.47E-06	-0.073	-0.184; 0.039	0.201	0.419	0.113; 0.726	0.007
cg09618309	-0.105	-0.152; - 0.058	1.09E-05	0.032	-0.03; 0.094	0.312	0.116	-0.393; 0.626	0.654
cg12448298	0.117	0.064; 0.171	1.76E-05	0.075	-0.004; 0.153	0.062	-0.104	-0.498; 0.29	0.604
cg13954213	0.063	0.034; 0.092	2.44E-05	-0.031	-0.079; 0.017	0.212	-0.087	-0.275; 0.102	0.369
cg14037665	-0.103	-0.147; -0.06	3.73E-06	0.011	-0.054; 0.076	0.748	0.061	-0.313; 0.435	0.749
cg14494451	-0.327	-0.438; - 0.217	6.06E-09	0.065	-0.014; 0.144	0.108	0.03	-0.532; 0.593	0.915
cg15845746	0.045	0.024; 0.066	2.31E-05	-0.023	-0.053; 0.007	0.128	0.019	-0.177; 0.216	0.848
cg19428444	-0.115	-0.166; - 0.064	8.96E-06	-0.085	-0.155; - 0.014	0.019	-0.383	-0.735; - 0.031	0.033
cg19485911	0.154	0.092; 0.217	1.33E-06	0.03	-0.035; 0.095	0.361	0.281	-0.102; 0.663	0.15
cg20356241	0.256	0.153; 0.359	1.08E-06	-0.052	-0.18; 0.077	0.43	0.063	-0.478; 0.603	0.821
cg22079149	-0.048	-0.069; - 0.026	1.63E-05	0.017	-0.02; 0.053	0.366	-0.114	-0.263; 0.036	0.135
cg22374586	-0.212	-0.286; - 0.138	1.92E-08	-0.06	-0.169; 0.049	0.28	-0.004	-0.509; 0.501	0.987
cg24935556	0.066	0.037; 0.094	6.95E-06	-0.061	-0.109; - 0.014	0.012	-0.238	-0.534; 0.057	0.114

cg06664959	-0.035	-0.051; - 0.019	2.11E-05	-0.011	-0.036; 0.013	0.36	-0.001	-0.125; 0.124	0.993
cg09435227	-0.041	-0.061; - 0.022	2.72E-05	0.016	-0.011; 0.044	0.245	-0.076	-0.214; 0.061	0.275
cg10178498	-0.078	-0.107; -0.05	6.88E-08	-0.019	-0.055; 0.018	0.312	0.153	0.011; 0.296	0.035
cg10991303	-0.195	-0.274; - 0.117	1.11E-06	-0.093	-0.232; 0.046	0.188	0.642	0.225; 1.059	0.003
cg12389423	-0.106	-0.148; - 0.063	1.13E-06	0.057	0.002; 0.112	0.043	0.246	-0.138; 0.63	0.21
cg18592273	0.372	0.248; 0.497	4.63E-09	0.165	-0.011; 0.341	0.065	NA	NA; NA	NA
cg19440233	0.11	0.062; 0.158	6.39E-06	0.005	-0.061; 0.072	0.872	-0.099	-0.457; 0.259	0.588
cg22471401	0.052	0.029; 0.074	7.68E-06	-0.007	-0.039; 0.026	0.679	-0.174	-0.418; 0.07	0.162
cg23202575	-0.127	-0.176; - 0.079	2.71E-07	0.057	-0.033; 0.148	0.215	-0.039	-0.522; 0.443	0.873
cg25679475	-0.12	-0.164; - 0.075	1.41E-07	-0.023	-0.098; 0.051	0.54	0.036	-0.204; 0.277	0.767
cg03931518	-0.077	-0.106; - 0.049	1.04E-07	-0.028	-0.062; 0.006	0.102	-0.048	-0.3; 0.203	0.706
cg06466757	0.15	0.089; 0.21	1.28E-06	-0.073	-0.172; 0.027	0.152	0.082	-0.245; 0.41	0.622
cg11060856	-0.068	-0.097; - 0.038	7.15E-06	-0.003	-0.047; 0.042	0.908	0.068	-0.114; 0.249	0.464
cg12857786	0.126	0.074; 0.178	1.87E-06	-0.109	-0.17; -0.049	0	0.255	-0.153; 0.663	0.22
cg13291296	-0.295	-0.399; - 0.191	2.64E-08	0.326	0.108; 0.545	0.003	0.391	-0.23; 1.012	0.217
cg13985817	-0.111	-0.161; -0.06	1.60E-05	-0.044	-0.134; 0.047	0.345	0.258	-0.054; 0.571	0.105
cg15407965	-0.144	-0.199; - 0.089	3.48E-07	0.049	-0.032; 0.13	0.237	0.231	-0.209; 0.67	0.304
cg17304168	0.071	0.039; 0.103	1.27E-05	0.063	0.014; 0.112	0.011	0.162	-0.212; 0.536	0.396

cg18299068	-0.094	-0.138; -0.05	2.76E-05	0.031	-0.041; 0.102	0.401	0.269	-0.101; 0.638	0.155
cg21304234	-0.165	-0.238; - 0.091	1.22E-05	0.016	-0.109; 0.14	0.806	-0.434	-0.807; - 0.061	0.023
cg23034496	-0.055	-0.08; -0.031	1.20E-05	0.016	-0.021; 0.054	0.395	-0.092	-0.306; 0.122	0.4
cg10629165	0.082	0.045; 0.119	1.20E-05	-0.079	-0.152; - 0.006	0.033	0.227	-0.074; 0.528	0.139
cg15043318	-0.078	-0.114; - 0.042	2.26E-05	-0.028	-0.07; 0.014	0.192	-0.01	-0.208; 0.188	0.919
cg18316500	-0.054	-0.077; - 0.032	2.96E-06	0.003	-0.035; 0.041	0.883	0.188	-0.054; 0.43	0.127
cg26496372	0.121	0.086; 0.156	1.15E-11	-0.01	-0.044; 0.024	0.553	-0.075	-0.263; 0.114	0.438
cg26650951	0.108	0.06; 0.156	9.49E-06	0.107	0.027; 0.187	0.008	0.108	-0.227; 0.443	0.527
cg08046604	-0.172	-0.251; - 0.093	2.13E-05	0.004	-0.109; 0.117	0.945	NA	NA; NA	NA
cg09620689	-0.075	-0.107; - 0.043	3.71E-06	-0.039	-0.09; 0.012	0.138	0.151	-0.19; 0.493	0.385
cg10179330	-0.168	-0.243; - 0.092	1.25E-05	-0.034	-0.111; 0.044	0.396	-0.108	-0.51; 0.295	0.6
cg11748650	0.104	0.059; 0.15	7.87E-06	0.023	-0.037; 0.084	0.451	-0.086	-0.474; 0.302	0.663
cg12802286	-0.1	-0.146; - 0.054	2.21E-05	0.062	0.004; 0.121	0.036	-0.001	-0.337; 0.334	0.994
cg13836183	-0.081	-0.119; - 0.043	2.56E-05	-0.025	-0.077; 0.026	0.336	-0.051	-0.444; 0.342	0.797
cg14279151	-0.063	-0.091; - 0.036	5.88E-06	0.098	0.058; 0.139	0	-0.154	-0.437; 0.129	0.286
cg15289190	0.129	0.079; 0.179	4.97E-07	0.013	-0.04; 0.066	0.641	0.082	-0.188; 0.352	0.553
cg16495982	0.235	0.154; 0.317	1.58E-08	0.028	-0.097; 0.152	0.662	NA	NA; NA	NA
cg16704889	-0.073	-0.106; - 0.039	2.51E-05	0.06	0.01; 0.111	0.019	0.015	-0.331; 0.361	0.931

cg24154161	0.188	0.1; 0.275	2.63E-05	0.107	-0.03; 0.243	0.125	NA	NA; NA	NA
cg24894158	0.195	0.116; 0.274	1.25E-06	0.077	-0.026; 0.179	0.143	-0.354	-0.922; 0.213	0.221
cg00460793	-0.074	-0.105; - 0.043	3.31E-06	0.02	-0.026; 0.066	0.392	0.247	-0.006; 0.501	0.056
cg00601953	-0.258	-0.359; - 0.158	4.54E-07	-0.171	-0.347; 0.004	0.056	NA	NA; NA	NA
cg00766497	-0.044	-0.063; - 0.025	5.28E-06	0.015	-0.016; 0.045	0.351	0.008	-0.184; 0.199	0.936
cg01509853	-0.07	-0.099; -0.04	3.69E-06	-0.025	-0.059; 0.01	0.157	0.034	-0.167; 0.236	0.74
cg01731811	-0.1	-0.144; - 0.056	9.55E-06	-0.064	-0.136; 0.007	0.079	-0.012	-0.33; 0.306	0.94
cg01827097	-0.077	-0.111; - 0.042	1.55E-05	0.007	-0.036; 0.05	0.741	-0.003	-0.23; 0.223	0.978
cg03521347	0.188	0.111; 0.264	1.52E-06	-0.001	-0.064; 0.062	0.971	-0.439	-0.955; 0.076	0.095
cg04106201	0.106	0.061; 0.15	2.94E-06	-0.042	-0.134; 0.049	0.363	0.072	-0.283; 0.427	0.69
cg04536807	-0.171	-0.248; - 0.095	1.21E-05	-0.012	-0.139; 0.114	0.852	0.533	0.078; 0.989	0.022
cg04678743	0.456	0.311; 0.602	7.92E-10	0.986	0.724; 1.248	0	NA	NA; NA	NA
cg06745145	0.093	0.054; 0.132	2.51E-06	0.015	-0.05; 0.08	0.651	-0.142	-0.434; 0.149	0.339
cg11405617	-0.095	-0.138; - 0.052	1.44E-05	0.026	-0.049; 0.102	0.489	-0.045	-0.486; 0.397	0.843
cg12020543	-0.069	-0.096; - 0.042	5.19E-07	0.013	-0.033; 0.06	0.574	-0.022	-0.281; 0.236	0.865
cg12921647	0.108	0.058; 0.157	2.35E-05	0.041	-0.045; 0.127	0.355	-0.044	-0.38; 0.293	0.799
cg26160086	-0.045	-0.066; - 0.025	1.13E-05	0.045	0.01; 0.08	0.012	-0.076	-0.275; 0.123	0.454
cg27658048	-0.079	-0.115; - 0.042	2.67E-05	-0.027	-0.078; 0.024	0.298	-0.069	-0.336; 0.198	0.613

cg01457883	0.126	0.068; 0.184	2.36E-05	-0.07	-0.169; 0.029	0.166	-0.112	-0.601; 0.377	0.653
cg05624932	-0.096	-0.141; - 0.052	2.01E-05	-0.013	-0.085; 0.058	0.717	-0.114	-0.419; 0.19	0.462
cg06009497	-0.052	-0.077; - 0.028	2.49E-05	0.005	-0.024; 0.033	0.753	-0.096	-0.217; 0.024	0.117
cg06440414	-0.134	-0.194; - 0.074	1.33E-05	0.057	-0.062; 0.176	0.347	0.132	-0.39; 0.653	0.621
cg08260790	0.102	0.058; 0.145	4.81E-06	0.151	0.081; 0.222	0	-0.152	-0.606; 0.302	0.511
cg09961689	-0.131	-0.184; - 0.079	9.35E-07	-0.022	-0.074; 0.03	0.409	0.153	-0.089; 0.395	0.214
cg12126859	0.109	0.063; 0.154	2.79E-06	-0.016	-0.094; 0.062	0.686	-0.233	-0.558; 0.091	0.159
cg14654239	0.122	0.065; 0.178	2.39E-05	0	-0.088; 0.088	0.996	-0.06	-0.398; 0.277	0.726
cg15948851	-0.055	-0.081; -0.03	2.10E-05	0.018	-0.022; 0.058	0.383	-0.064	-0.316; 0.188	0.618
cg16752592	-0.146	-0.214; - 0.078	2.28E-05	0.063	-0.037; 0.164	0.218	-0.045	-0.575; 0.486	0.869
cg22848598	-0.232	-0.324; - 0.141	6.96E-07	0.043	-0.081; 0.166	0.497	0.549	-0.051; 1.149	0.073
cg23303108	-0.181	-0.254; - 0.107	1.40E-06	0.069	-0.064; 0.201	0.312	0.173	-0.318; 0.663	0.491
cg24507266	0.063	0.036; 0.09	4.72E-06	-0.037	-0.081; 0.008	0.107	-0.245	-0.536; 0.046	0.098
cg26394257	-0.092	-0.132; - 0.051	1.03E-05	-0.061	-0.114; - 0.007	0.026	0.267	-0.039; 0.573	0.087
cg26530341	-0.086	-0.123; -0.05	3.87E-06	0.059	-0.014; 0.131	0.115	0.111	-0.181; 0.402	0.457
cg14286514	-0.129	-0.185; - 0.073	6.06E-06	-0.073	-0.158; 0.011	0.09	0.045	-0.351; 0.44	0.825
cg04714939	-0.094	-0.136; - 0.052	9.79E-06	-0.062	-0.125; 0.002	0.057	-0.293	-0.557; - 0.029	0.03
cg06539091	-0.284	-0.39; -0.177	1.77E-07	-0.005	-0.141; 0.13	0.938	-0.027	-0.42; 0.366	0.894

cg11518184	-0.06	-0.084; - 0.036	1.31E-06	-0.013	-0.053; 0.026	0.507	-0.168	-0.385; 0.048	0.128
cg14677612	0.154	0.095; 0.214	3.54E-07	-0.017	-0.122; 0.088	0.745	-0.538	-1.167; 0.091	0.094
cg14677909	-0.29	-0.366; - 0.214	8.55E-14	0.071	-0.147; 0.289	0.523	-0.358	-0.806; 0.091	0.118
cg15275103	-0.371	-0.473; -0.27	6.38E-13	0.069	-0.059; 0.197	0.292	-0.202	-0.549; 0.145	0.253
cg16696043	-0.05	-0.073; - 0.027	1.68E-05	-0.051	-0.085; - 0.017	0.003	-0.171	-0.36; 0.018	0.076
cg18129282	0.048	0.027; 0.07	1.36E-05	-0.022	-0.056; 0.013	0.214	-0.161	-0.353; 0.03	0.099
cg18308755	-0.143	-0.192; - 0.093	1.42E-08	0.058	-0.009; 0.126	0.092	-0.144	-0.544; 0.255	0.479
cg23881299	-0.104	-0.137; -0.07	2.23E-09	0.017	-0.015; 0.05	0.294	-0.03	-0.207; 0.148	0.744
cg25421941	0.133	0.076; 0.19	4.76E-06	0.095	0.012; 0.177	0.024	0.111	-0.442; 0.665	0.693
cg25928881	-0.165	-0.236; - 0.095	3.77E-06	-0.022	-0.151; 0.107	0.738	NA	NA; NA	NA
cg27190138	0.093	0.056; 0.13	8.74E-07	-0.013	-0.066; 0.04	0.635	-0.013	-0.29; 0.265	0.928
cg27286337	0.343	0.202; 0.485	1.98E-06	0.298	0.089; 0.506	0.005	NA	NA; NA	NA
cg00293245	0.12	0.073; 0.168	7.66E-07	-0.016	-0.07; 0.037	0.546	0.055	-0.259; 0.369	0.732
cg00338749	-0.121	-0.169; - 0.073	6.84E-07	0.059	-0.017; 0.135	0.128	-0.001	-0.291; 0.289	0.994
cg00719410	0.077	0.045; 0.109	2.22E-06	-0.015	-0.052; 0.022	0.433	-0.097	-0.269; 0.076	0.273
cg00877321	-0.124	-0.183; - 0.066	2.76E-05	-0.112	-0.204; -0.02	0.017	0.121	-0.344; 0.586	0.611
cg01981334	-0.192	-0.242; - 0.141	1.13E-13	0.012	-0.024; 0.049	0.515	0.112	-0.092; 0.317	0.282
cg04293602	-0.076	-0.11; -0.042	1.21E-05	-0.05	-0.106; 0.007	0.087	-0.074	-0.293; 0.145	0.507

cg06112087	-0.166	-0.231; - 0.101	5.66E-07	0.025	-0.064; 0.114	0.58	-0.338	-0.741; 0.064	0.1
cg10917952	-0.151	-0.21; -0.091	6.36E-07	-0.083	-0.133; - 0.033	0.001	-0.282	-0.551; - 0.013	0.04
cg15233880	-0.224	-0.321; - 0.127	6.53E-06	0.103	-0.032; 0.239	0.135	NA	NA; NA	NA
cg16082644	0.096	0.052; 0.14	2.15E-05	-0.035	-0.102; 0.031	0.295	0.122	-0.151; 0.395	0.38
cg20325517	-0.26	-0.379; - 0.141	1.92E-05	0.027	-0.171; 0.225	0.79	NA	NA; NA	NA
cg21773220	0.119	0.069; 0.169	3.11E-06	0.047	-0.023; 0.117	0.185	0.246	-0.063; 0.555	0.119
cg22049858	0.212	0.137; 0.287	3.42E-08	-0.041	-0.132; 0.051	0.383	-0.579	-0.985; - 0.174	0.005
cg26079428	0.085	0.046; 0.125	2.26E-05	-0.014	-0.08; 0.051	0.666	-0.237	-0.559; 0.084	0.148
cg27261733	-0.268	-0.354; - 0.182	9.45E-10	-0.005	-0.088; 0.077	0.897	0.044	-0.419; 0.507	0.851
cg00331237	0.087	0.053; 0.121	4.11E-07	-0.002	-0.045; 0.041	0.933	-0.161	-0.417; 0.094	0.215
cg00596508	-0.262	-0.365; -0.16	5.42E-07	-0.126	-0.32; 0.068	0.203	NA	NA; NA	NA
cg02574894	0.179	0.117; 0.242	1.45E-08	0.018	-0.048; 0.084	0.585	0.239	-0.119; 0.597	0.191
cg06400595	0.201	0.124; 0.278	3.46E-07	0.001	-0.059; 0.06	0.981	0.049	-0.289; 0.387	0.777
cg08693490	0.054	0.029; 0.079	2.28E-05	0.007	-0.037; 0.05	0.759	-0.247	-0.492; - 0.002	0.048
cg10730291	-0.274	-0.38; -0.168	4.17E-07	0.258	0.131; 0.386	0	NA	NA; NA	NA
cg11661631	-0.04	-0.059; - 0.021	2.65E-05	-0.014	-0.046; 0.018	0.404	-0.048	-0.193; 0.096	0.514
cg14603867	0.14	0.077; 0.202	1.23E-05	0.006	-0.07; 0.081	0.884	0.397	-0.114; 0.908	0.128
cg20781967	0.113	0.061; 0.165	1.82E-05	-0.069	-0.154; 0.016	0.113	-0.314	-0.691; 0.063	0.103

cg22660542	-0.129	-0.181; - 0.077	1.26E-06	-0.068	-0.162; 0.026	0.158	-0.346	-0.689; - 0.003	0.048
cg25367249	0.168	0.099; 0.237	1.95E-06	0.143	0.034; 0.252	0.01	-0.32	-0.758; 0.117	0.151
cg11315081	-0.204	-0.293; - 0.114	8.44E-06	-0.064	-0.192; 0.063	0.322	NA	NA; NA	NA
cg12275871	-0.133	-0.196; - 0.071	2.70E-05	-0.227	-0.334; -0.12	0	0.399	-0.262; 1.06	0.236
cg16564828	-0.254	-0.353; - 0.155	4.74E-07	-0.287	-0.414; - 0.161	0	NA	NA; NA	NA
cg19133973	0.046	0.026; 0.066	6.11E-06	-0.002	-0.034; 0.029	0.877	-0.121	-0.272; 0.031	0.12
cg02583546	0.223	0.137; 0.308	3.40E-07	-0.057	-0.14; 0.027	0.183	0.152	-0.177; 0.482	0.365
cg07285995	-0.199	-0.275; - 0.123	2.46E-07	-0.055	-0.163; 0.052	0.313	0.078	-0.409; 0.566	0.753
cg18837429	0.059	0.032; 0.086	1.80E-05	-0.025	-0.063; 0.014	0.205	0.204	-0.008; 0.417	0.06
cg02478956	0.153	0.083; 0.223	1.93E-05	-0.017	-0.143; 0.11	0.794	-0.016	-0.413; 0.382	0.939
cg03146503	-0.131	-0.185; - 0.078	1.17E-06	-0.026	-0.079; 0.027	0.338	-0.162	-0.467; 0.142	0.296
cg14209037	-0.146	-0.2; -0.093	8.12E-08	0.006	-0.072; 0.084	0.879	0.091	-0.309; 0.491	0.655
cg23991388	0.082	0.045; 0.118	9.94E-06	-0.052	-0.117; 0.013	0.119	-0.171	-0.462; 0.121	0.252
cg01003448	0.213	0.122; 0.304	4.31E-06	-0.066	-0.199; 0.067	0.332	0.09	-0.448; 0.627	0.743
cg01031475	-0.054	-0.079; - 0.029	2.37E-05	0.071	0.03; 0.111	0.001	-0.102	-0.429; 0.225	0.54
cg06181567	-0.103	-0.146; -0.06	3.24E-06	-0.033	-0.087; 0.021	0.227	-0.2	-0.472; 0.072	0.15
cg06672545	0.127	0.069; 0.185	1.83E-05	0.077	-0.014; 0.168	0.099	0.133	-0.365; 0.631	0.601
cg07482202	0.397	0.261; 0.533	1.09E-08	-0.064	-0.263; 0.135	0.527	0.139	-0.739; 1.016	0.757

cg16619935	-0.054	-0.078; - 0.031	5.82E-06	0.055	0.013; 0.098	0.011	-0.163	-0.374; 0.048	0.13
cg24303478	-0.047	-0.068; - 0.026	1.06E-05	0.02	-0.009; 0.048	0.173	-0.004	-0.156; 0.149	0.962
cg27605307	-0.123	-0.181; - 0.066	2.72E-05	-0.082	-0.158; - 0.006	0.035	-0.245	-0.647; 0.157	0.232
cg00590152	-0.114	-0.167; - 0.061	2.65E-05	-0.03	-0.124; 0.063	0.524	0.231	-0.162; 0.624	0.249
cg02216727	-0.101	-0.144; - 0.057	6.19E-06	0.035	-0.011; 0.081	0.139	-0.215	-0.475; 0.045	0.105
cg03452174	0.076	0.044; 0.108	2.56E-06	-0.037	-0.094; 0.02	0.198	0.13	-0.136; 0.396	0.337
cg04206797	0.184	0.102; 0.267	1.27E-05	-0.013	-0.153; 0.128	0.856	NA	NA; NA	NA
cg11329058	0.096	0.055; 0.138	4.84E-06	-0.148	-0.221; - 0.075	0	0.229	-0.157; 0.614	0.245
cg12679910	0.078	0.043; 0.113	1.27E-05	-0.011	-0.058; 0.036	0.643	-0.261	-0.589; 0.067	0.119
cg13117272	-0.039	-0.057; - 0.021	2.62E-05	-0.001	-0.032; 0.03	0.945	-0.011	-0.157; 0.134	0.877
cg14811011	-0.267	-0.385; -0.15	8.04E-06	0.152	-0.05; 0.354	0.14	0.074	-0.581; 0.728	0.825
cg16904599	0.088	0.05; 0.125	5.25E-06	-0.067	-0.109; - 0.025	0.002	0.038	-0.229; 0.305	0.781
cg18576374	0.181	0.139; 0.224	3.44E-17	-0.093	-0.175; - 0.011	0.026	0.047	-0.292; 0.385	0.787
cg22277994	-0.06	-0.085; - 0.034	3.93E-06	-0.001	-0.046; 0.045	0.97	-0.115	-0.36; 0.129	0.355
cg24580146	-0.113	-0.161; - 0.065	3.97E-06	0.008	-0.067; 0.084	0.829	-0.335	-0.646; - 0.023	0.035
cg25170034	-0.122	-0.175; - 0.069	6.44E-06	0.013	-0.064; 0.09	0.743	-0.201	-0.52; 0.117	0.216
cg25496297	-0.087	-0.125; - 0.048	1.30E-05	-0.025	-0.085; 0.035	0.415	-0.37	-0.684; - 0.055	0.021
cg06490951	-0.086	-0.126; - 0.046	2.72E-05	0.047	-0.028; 0.122	0.218	0.256	-0.1; 0.612	0.159

cg00541350	-0.062	-0.089; -0.035	5.31E-06	0.024	-0.017; 0.064	0.247	0.018	-0.211; 0.246	0.88
cg01017773	-0.176	-0.246; -0.106	8.48E-07	-0.093	-0.196; 0.01	0.076	-0.007	-0.509; 0.495	0.977
cg02633409	-0.083	-0.116; -0.05	9.07E-07	0.015	-0.021; 0.05	0.427	-0.074	-0.325; 0.177	0.565
cg02796790	-0.051	-0.074; -0.028	1.67E-05	0.041	0.009; 0.074	0.013	0.03	-0.215; 0.275	0.81
cg07356415	-0.067	-0.095; -0.039	2.39E-06	0.053	0.012; 0.095	0.012	-0.092	-0.282; 0.098	0.34
cg12074025	0.205	0.116; 0.295	7.35E-06	-0.217	-0.387; -0.046	0.013	NA	NA; NA	NA
cg12088773	-0.184	-0.263; -0.106	3.70E-06	-0.088	-0.216; 0.04	0.178	-0.148	-0.674; 0.377	0.58
cg12258179	-0.184	-0.268; -0.1	1.76E-05	-0.119	-0.295; 0.058	0.188	0.252	-0.302; 0.806	0.373
cg12497870	-0.062	-0.088; -0.036	2.09E-06	0.02	-0.017; 0.058	0.29	0.078	-0.144; 0.3	0.493
cg13670756	-0.138	-0.194; -0.082	1.55E-06	-0.069	-0.151; 0.013	0.098	-0.301	-0.674; 0.072	0.114
cg14032725	-0.062	-0.09; -0.033	1.96E-05	-0.027	-0.077; 0.023	0.286	0.16	-0.104; 0.423	0.235
cg18031747	0.387	0.272; 0.502	4.47E-11	0.107	0.029; 0.185	0.007	NA	NA; NA	NA
cg18344466	-0.114	-0.166; -0.062	1.81E-05	0.018	-0.064; 0.1	0.672	0.083	-0.231; 0.397	0.604
cg23037932	0.073	0.039; 0.107	2.15E-05	0.095	0.043; 0.148	0	-0.061	-0.337; 0.215	0.665
cg24217159	-0.304	-0.42; -0.187	3.22E-07	0.17	0.023; 0.317	0.023	0.365	-0.162; 0.891	0.175
cg26623885	-0.088	-0.128; -0.048	1.41E-05	-0.111	-0.175; -0.046	0.001	0.086	-0.255; 0.427	0.62
cg26709695	-0.099	-0.144; -0.053	1.89E-05	-0.093	-0.161; -0.024	0.008	-0.707	-1.097; -0.317	0
cg26776924	-0.05	-0.073; -0.027	1.75E-05	0.065	0.03; 0.1	0	0.196	0.017; 0.374	0.032

cg09576415	-0.05	-0.072; - 0.028	8.53E-06	0.004	-0.036; 0.044	0.84	-0.058	-0.249; 0.132	0.548
cg09801828	-0.041	-0.059; - 0.023	9.41E-06	-0.016	-0.053; 0.021	0.399	0.018	-0.134; 0.171	0.816
cg10629004	-0.166	-0.233; - 0.099	1.33E-06	-0.175	-0.286; - 0.064	0.002	0.329	-0.063; 0.72	0.1
cg14083397	0.316	0.238; 0.395	3.30E-15	-0.016	-0.066; 0.033	0.516	-0.084	-0.434; 0.265	0.636
cg17206393	0.055	0.031; 0.079	8.11E-06	0.026	-0.012; 0.064	0.184	0.066	-0.169; 0.301	0.584
cg18150852	-0.104	-0.153; - 0.056	2.44E-05	0.097	0.03; 0.164	0.005	0.226	-0.022; 0.474	0.074
cg04187708	-0.107	-0.154; -0.06	8.27E-06	0.045	-0.03; 0.121	0.24	-0.137	-0.475; 0.202	0.429
cg12826791	-0.253	-0.341; - 0.165	1.89E-08	0.121	0.038; 0.203	0.004	-0.066	-0.432; 0.3	0.723
cg21505925	0.087	0.051; 0.124	3.04E-06	-0.018	-0.074; 0.039	0.542	0.019	-0.225; 0.262	0.881
cg27376287	-0.175	-0.256; - 0.095	1.97E-05	-0.055	-0.16; 0.05	0.303	NA	NA; NA	NA
cg00256932	-0.164	-0.235; - 0.093	6.51E-06	0.062	-0.051; 0.174	0.282	0.21	-0.257; 0.677	0.378
cg16740427	-0.097	-0.141; - 0.052	2.25E-05	0.066	-0.007; 0.138	0.077	0.491	0.152; 0.83	0.005

Table 9.8. Table of characteristics of probes found to be significantly associated with air PAH8 exposure at the FDR level ($p < 2.8 \times 10^{-5}$ in the training dataset).

Probe ID	Chromosome	Position	UCSC RefGene Name	Gene Location	Relation to CpG Island	Methylation Change Direction
cg00466488	1	118148927	<i>FAM46C</i>	5'UTR	Island	+
cg00695391	1	2525548	<i>MMEL1</i>	Body	North Shore	-
cg03317082	1	234748618			South Shore	+
cg04117764	1	10917451				-
cg04273604	1	234609003	<i>TARBP1</i>	Body		+
cg06443644	1	1846046	<i>CALML6</i>	TSS1500		-
cg07826526	1	194087368				-
cg12653146	1	25919290				-
cg16306900	1	182026435	<i>ZNF648</i>	Body	Island	-
cg17900103	1	20940981	<i>CDA</i>	Body		-
cg18801028	1	29646475	<i>PTPRU</i>	Body		-
cg19925780	1	101509557				+
cg25876319	1	158037893	<i>KIRREL</i>	Body		-
cg03219514	2	98350753	<i>ZAP70</i>	Body	North Shore	-
cg05703053	2	169769616				-
cg06856378	2	160759118	<i>LY75</i>	Body	North Shore	-
cg07480373	2	216874286	<i>MREG</i>	Body	North Shelf	+
cg08157672	2	68547141	<i>CNRIP1</i>	1stExon	South Shore	+
cg09618309	2	99908962	<i>LYG1</i>	Body		-
cg12448298	2	115822039	<i>DPP10</i>	Body		+

cg13954213	2	217971590				+
cg14037665	2	1748617	<i>PXDN</i>	TSS1500	Island	-
cg14494451	2	153575552	<i>ARL6IP6</i>	Body	Island	-
cg15845746	2	120438237	<i>TMEM177</i>	5'UTR	South Shore	+
cg19428444	2	21023690	<i>C2orf43</i>	TSS1500	South Shore	-
cg19485911	2	220380542	<i>ACCN4</i>	Body	South Shelf	+
cg20356241	2	10691911				+
cg22079149	2	241568656	<i>GPR35</i>	TSS200	North Shore	-
cg22374586	2	232220566				-
cg24935556	2	21291088				+
cg06664959	3	24871703			South Shore	-
cg09435227	3	55505073	<i>WNT5A</i>	Body	South Shore	-
cg10178498	3	124103021	<i>KALRN</i>	Body		-
cg10991303	3	142608051	<i>PCOLCE2</i>	TSS200	Island	-
cg12389423	3	118864836	<i>C3orf30; IGSF11</i>	TSS200; 1stExon		-
cg18592273	3	161089930	<i>C3orf57</i>	TSS200	Island	+
cg19440233	3	52009002	<i>ABHD14A; ABHD14B</i>	TSS200; TSS1500	Island	+
cg22471401	3	183824717	<i>HTR3E</i>	3'UTR		+
cg23202575	3	183888295	<i>DVL3</i>	Body	Island	-
cg25679475	3	118705126	<i>IGSF11</i>	Body		-
cg03931518	4	79106308	<i>FRAS1</i>	Body	South Shelf	-

cg06466757	4	1255808				+
cg11060856	4	5895410	<i>CRMP1</i>	TSS1500	South Shore	-
cg12857786	4	83934023	<i>LIN54</i>	1stExon	Island	+
cg13291296	4	22390126	<i>GPR125</i>	Body		-
cg13985817	4	56685845	<i>LOC644145</i>	TSS1500		-
cg15407965	4	128707242	<i>HSPA4L</i>	Body	South Shelf	-
cg17304168	4	15626104	<i>FBXL5</i>	Body		+
cg18299068	4	1305425	<i>MAEA</i>	Body	South Shore	-
cg21304234	4	839646			North Shelf	-
cg23034496	4	741516	<i>PCGF3</i>	Body	Island	-
cg10629165	5	66124563	<i>MAST4</i>	TSS200		+
cg15043318	5	170947897				-
cg18316500	5	177855654	<i>COL23A1</i>	Body		-
cg26496372	5	37379396	<i>WDR70</i>	TSS200	Island	+
cg26650951	5	50686147	<i>ISL1</i>	Body	Island	+
cg08046604	6	135238772	<i>ALDH8A1</i>	3'UTR		-
cg09620689	6	169351108				-
cg10179330	6	163755233				-
cg11748650	6	32279016	<i>C6orf10</i>	Body		+
cg12802286	6	129513973	<i>LAMA2</i>	Body		-
cg13836183	6	32131254	<i>EGFL8; PPT2</i>	TSS1500; 3'UTR	North Shelf	-
cg14279151	6	52369502	<i>TRAM2</i>	Body		-

cg15289190	6	28831544			North Shore	+
cg16495982	6	30641015	<i>DHX16</i>	TSS200	South Shore	+
cg16704889	6	31696729	<i>DDAH2</i>	Body	Island	-
cg24154161	6	32820421	<i>TAP1</i>	Body	North Shore	+
cg24894158	6	145069599	<i>UTRN</i>	Body		+
cg00460793	7	36157021				-
cg00601953	7	155856104			Island	-
cg00766497	7	2148973	<i>MAD1L1</i>	Body		-
cg01509853	7	4754502	<i>FOXK1</i>	Body	North Shelf	-
cg01731811	7	157890171	<i>PTPRN2</i>	Body	North Shore	-
cg01827097	7	1937682	<i>MAD1L1</i>	Body	North Shore	-
cg03521347	7	44529975	<i>NUDCD3</i>	Body	Island	+
cg04106201	7	39332739	<i>POU6F2</i>	Body		+
cg04536807	7	32997190	<i>FKBP9</i>	1stExon	Island	-
cg04678743	7	130353515	<i>TSGA13; COPG2</i>	3'UTR; 5'UTR	Island	+
cg06745145	7	90664816	<i>CDK14</i>	Body		+
cg11405617	7	6541332	<i>GRID2IP</i>	Body	North Shore	-
cg12020543	7	6691949			Island	-
cg12921647	7	1336696			North Shore	+
cg26160086	7	157205545	<i>DNAJB6</i>	Body	North Shore	-
cg27658048	7	4201594	<i>SDK1</i>	Body		-
cg01457883	8	335353				+

cg05624932	8	75897310	<i>CRISPLD1</i>	5'UTR	South Shore	-
cg06009497	8	37695050	<i>GPR124</i>	Body	North Shelf	-
cg06440414	8	142149205	<i>DENND3</i>	Body		-
cg08260790	8	98294370			South Shelf	+
cg09961689	8	144590068	<i>ZC3H3</i>	Body		-
cg12126859	8	335281				+
cg14654239	8	335341				+
cg15948851	8	59059119	<i>FAM110B</i>	Body	Island	-
cg16752592	8	145537709	<i>HSF1</i>	Body		-
cg22848598	8	38965026	<i>ADAM32</i>	TSS200	Island	-
cg23303108	8	23083578	<i>TNFRSF10A</i>	TSS1500	South Shore	-
cg24507266	8	145027948	<i>PLEC1</i>	1stExon	Island	+
cg26394257	8	38964993	<i>ADAM32</i>	TSS200	North Shore	-
cg26530341	8	23083353	<i>TNFRSF10A</i>	TSS1500	South Shore	-
cg14286514	9	32525315	<i>DDX58</i>	Body	North Shore	-
cg04714939	10	80917620	<i>ZMIZ1</i>	5'UTR		-
cg06539091	10	116390942	<i>ABLIM1</i>	Body	North Shore	-
cg11518184	10	133849797			Island	-
cg14677612	10	131263962	<i>MGMT</i>	TSS1500	North Shore	+
cg14677909	10	48807341	<i>PTPN20B;</i> <i>PTPN20A</i>	Body; 5'UTR		-
cg15275103	10	124893024			Island	-

cg16696043	10	49658928	ARHGAP22	Body	Island	-
cg18129282	10	89987284				+
cg18308755	10	134065956	STK32C	Body		-
cg23881299	10	121116990	GRK5	Body		-
cg25421941	10	11506312	USP6NL	Body	South Shore	+
cg25928881	10	32632685	EPC1	Body	North Shelf	-
cg27190138	10	98479757	PIK3AP1	Body	Island	+
cg27286337	10	134555280	INPP5A	Body	Island	+
cg00293245	11	94884652			Island	+
cg00338749	11	1036677	MUC6	1stExon	Island	-
cg00719410	11	94884479			Island	+
cg00877321	11	627612	SCT	TSS1500	Island	-
cg01981334	11	64877237	C11orf2	Body	North Shore	-
cg04293602	11	65553660	OVOL1	TSS1500	North Shore	-
cg06112087	11	18740452	IGSF22	Body	North Shelf	-
cg10917952	11	5989222	OR56A5	1stExon		-
cg15233880	11	69454727	CCND1	TSS1500	Island	-
cg16082644	11	59324522			North Shelf	+
cg20325517	11	96123557	JRKL; CCDC82	5'UTR; 1stExon	Island	-
cg21773220	11	94883919			Island	+
cg22049858	11	94884121			Island	+
cg26079428	11	106513618				+

cg27261733	11	1891872	<i>LSP1</i>	5'UTR	North Shore	-
cg00331237	12	77273319	<i>CSRP2</i>	TSS1500	South Shore	+
cg00596508	12	41086800	<i>CNTN1</i>	5'UTR	Island	-
cg02574894	12	53693825	<i>C12orf10</i>	Body	Island	+
cg06400595	12	56551642	<i>MYL6B; MYL6</i>	3'UTR; TSS1500		+
cg08693490	12	116757896				+
cg10730291	12	2908075	<i>FKBP4</i>	Body	South Shelf	-
cg11661631	12	57586714	<i>LRP1</i>	Body		-
cg14603867	12	53693485	<i>C12orf10</i>	1stExon	Island	+
cg20781967	12	772688	<i>NINJ2</i>	1stExon		+
cg22660542	12	54402431	<i>HOXC8</i>	TSS1500	North Shore	-
cg25367249	12	96127072	<i>NTN4</i>	Body		+
cg11315081	13	22651243				-
cg12275871	13	114096262	<i>ADPRHL1</i>	Body	Island	-
cg16564828	13	113660298	<i>MCF2L</i>	Body	North Shelf	-
cg19133973	13	33467980				+
cg02583546	14	77494451	<i>C14orf4</i>	5'UTR	Island	+
cg07285995	14	92790364	<i>SLC24A4</i>	Body	Island	-
cg18837429	14	75050528	<i>LTBP2</i>	Body	North Shelf	+
cg02478956	15	22833681	<i>TUBGCP5</i>	Body	Island	+
cg03146503	15	67127653				-

cg14209037	15	41228521	<i>DLL4</i>	Body	Island	-
cg23991388	15	100654760	<i>ADAMTS17</i>	Body		+
cg01003448	16	745685	<i>FBXL16</i>	Body	Island	+
cg01031475	16	30004376	<i>HIRIP3</i>	3'UTR	North Shelf	-
cg06181567	16	50731115	<i>NOD2</i>	1stExon		-
cg06672545	16	19869965				+
cg07482202	16	745687	<i>FBXL16</i>	Body	Island	+
cg16619935	16	2037439	<i>GFER</i>	3'UTR	North Shelf	-
cg24303478	16	89143845			Island	-
cg27605307	16	20357506	<i>UMOD</i>	Body	North Shelf	-
cg00590152	17	74003831	<i>EVPL</i>	Body	Island	-
cg02216727	17	38520653	<i>GJD3</i>	1stExon	South Shore	-
cg03452174	17	27045113	<i>RAB34</i>	Body	North Shore	+
cg04206797	17	36689639	<i>SRCIN1</i>	Body		+
cg11329058	17	78549358	<i>RPTOR</i>	Body		+
cg12679910	17	73900713	<i>MRPL38</i>	Body	Island	+
cg13117272	17	79681052	<i>SLC25A10</i>	Body	North Shore	-
cg14811011	17	73055792	<i>KCTD2</i>	Body		-
cg16904599	17	57287367	<i>C17orf71</i>	TSS200	Island	+
cg18576374	17	78549371	<i>RPTOR</i>	Body		+
cg22277994	17	71161157	<i>SSTR2</i>	TSS200	Island	-

cg24580146	17	7644241	<i>DNAH2</i>	Body	Island	-
cg25170034	17	33288066	<i>ZNF830; CCT6B</i>	TSS1500; Body	North Shore	-
cg25496297	17	27411642	<i>MYO18A</i>	Body	South Shelf	-
cg06490951	18	580435	<i>CETN1</i>	1stExon	Island	-
cg00541350	19	13869979	<i>CCDC130</i>	Body	Island	-
cg01017773	19	57988497	<i>ZNF772</i>	Body		-
cg02633409	19	5051110	<i>KDM4B</i>	Body	South Shelf	-
cg02796790	19	10823761	<i>QTRT1</i>	Body	Island	-
cg07356415	19	19655352	<i>CILP2</i>	Body	Island	-
cg12074025	19	58238850	<i>ZNF671</i>	1stExon	Island	+
cg12088773	19	44128330	<i>CADM4</i>	Body		-
cg12258179	19	49232187	<i>RASIP1</i>	Body	Island	-
cg12497870	19	36210913	<i>MLL4</i>	Body	Island	-
cg13670756	19	3662042	<i>PIP5K1C</i>	Body	South Shelf	-
cg14032725	19	46289742	<i>DMWD</i>	Body	Island	-
cg18031747	19	9929709	<i>FBXL12</i>	1stExon	Island	+
cg18344466	19	57702324	<i>ZNF264</i>	TSS1500	North Shore	-
cg23037932	19	55042806	<i>KIR3DX1</i>	TSS1500		+
cg24217159	19	47844652	<i>GPR77</i>	Body		-
cg26623885	19	38709496	<i>DPF1</i>	Body	North Shelf	-

cg26709695	19	3587573	<i>GIPC3</i>	Body	South Shore	-
cg26776924	19	1969666	<i>CSNK1G2</i>	5'UTR	Island	-
cg09576415	20	62059559	<i>KCNQ2</i>	Body	Island	-
cg09801828	20	33462656	<i>ACSS2</i>	TSS200	North Shore	-
cg10629004	20	21696467	<i>PAX1</i>	3'UTR	South Shore	-
cg14083397	20	388473	<i>RBCK1</i>	TSS1500	Island	+
cg17206393	20	33681223	<i>TRPC4AP</i>	TSS1500	South Shore	+
cg18150852	20	4807932			South Shelf	-
cg04187708	21	36693060				-
cg12826791	21	45926719	<i>C21orf29</i>	Body	Island	-
cg21505925	21	38807423	<i>DYRK1A</i>	Body		+
cg27376287	21	19192138	<i>C21orf91</i>	TSS1500	South Shore	-
cg00256932	22	51041732	<i>MAPK8IP2</i>	1stExon	North Shore	-
cg16740427	22	42469976	<i>FAM109B</i>	TSS1500	North Shore	-

Table 9.9. Table comparing results published by Tryndyak et al. (2018)³²⁵ and the air PAH8 exposure EWAS results

<u>Gene Name</u>	Tryndyak et al. 2018 ³²⁵					Combined PAH8 EWAS Results		
	<u>Chromosome</u>	<u>Start</u>	<u>End</u>	<u>Genomic Location</u>	<u>Direction of Methylation Change</u>	<u>Probe ID</u>	<u>Genomic Location</u>	<u>Direction of Methylation Change</u>
<i>DLL4</i>	chr15	41230332	41230369	3' UTR	-	cg14209037	-	Body
	chr15	41230466	41230643					
<i>SSTR2</i>	chr17	71166008	71166009	Exon	+	cg22277994	-	TSS200
<i>SDK1</i>	chr7	4308985	4309123	TTS	+	cg27658048	-	Body
<i>TNFRSF10A</i>	chr8	23054712	23054713	Exon	+	cg23303108	-	TSS1500
						cg26530341		

Table 9.10. Table showing overlaps between results of the air PAH8 exposure EWAS, and results from published smoking EWAS. Overlaps were identified by looking for exact CpG probes and by looking for probes with the same genes.

Gene	Published Smoking EWAS Results				Air PAH8 EWAS Results	
	Study	CpG Probe	Direction	Tissue	CpG Probe	Direction
<i>ABLIM1</i>	Joubert, 2016 ⁴¹⁶	cg09649347	-	Blood	cg06539091	-
		cg13587915				
<i>ARHGAP22</i>	Dogan, 2014 ⁴²¹	cg24385334	+	Blood	cg16696043	-
<i>C11orf2</i>	Joubert, 2016 ⁴¹⁶	cg13626866	+	Blood	cg01981334	-
<i>C14orf4</i>	Besingi, 2014 ³⁸⁰	cg02583546	+	Whole blood	cg02583546	+
<i>CCND1</i>	Lee, 2016 ⁴⁵⁰	cg09520904	-	Blood	cg15233880	-
<i>CILP2</i>	Joubert, 2016 ⁴¹⁶	cg07942040	+	Blood	cg07356415	-
<i>COL23A1</i>	Dogan, 2014 ⁴²¹	cg17731547	-	Blood	cg18316500	-
	Joubert, 2016 ⁴¹⁶	cg12194832	-			
		cg21871330	+			
<i>CSNK1G2</i>	Joehanes, 2016 ⁴¹⁵	cg17884674	+	Blood	cg26776924	-
<i>DDAH2</i>	Ivorra, 2015 ⁴²⁷	cg15264752	+	Cord blood, Blood	cg16704889	-
	Joubert, 2016 ⁴¹⁶	cg26111283		Blood		
<i>DENND3</i>	Allione, 2015 ⁴¹⁷	cg11538410	-	Whole blood	cg06440414	-
	Joubert, 2016 ⁴¹⁶	cg03401656	+	Blood		
		cg06623899				
<i>DHX16</i>	Joubert, 2016 ⁴¹⁶	cg20117675	+	Blood	cg16495982	+
<i>DPF1</i>	Joehanes, 2016 ⁴¹⁵	cg26950531	+	Blood	cg26623885	-
<i>DPP10</i>	Chhabra, 2014 ⁴²⁰	cg22670147	-	Lung	cg12448298	+
<i>EGFL8</i>	Allione, 2015 ⁴¹⁷	cg10502563	-	Whole blood	cg13836183	-

FBXL12	Dogan, 2014 ⁴²¹	cg03621406 cg21112148	-	Blood	cg18031747	+
FBXL16	Joehanes, 2016 ⁴¹⁵	cg05542681	+	Blood	cg01003448 cg07482202	+
	Joubert, 2016 ⁴¹⁶	cg02713960 cg02958327 cg26804595	-			
	Markunas, 2014 ⁴³³	cg26804595	-			
FBXL5	Joubert, 2016 ⁴¹⁶	cg02630888 cg15175162	+	Blood	cg17304168	+
GJD3	Joubert, 2016 ⁴¹⁶	cg05568941 cg05930207 cg06949812 cg11758793	+	Blood	cg02216727	-
	Markunas, 2014 ⁴³³	cg05568941 cg06949812				
GPR124	Joehanes, 2016 ⁴¹⁵	cg20272648 cg01226742		Blood	cg06009497	-
	Joubert, 2016 ⁴¹⁶	cg05552035 cg12424646 cg12869334	-			
GPR77	Joehanes, 2016 ⁴¹⁵	cg16734795	+	Blood	cg24217159	-
	Joubert, 2016 ⁴¹⁶	cg24217159				
GRK5	Joehanes, 2016 ⁴¹⁵	cg24539517 cg00522048	- -	Blood	cg23881299	-
	Joubert, 2016 ⁴¹⁶	cg09085932 cg14351425	- +			
		cg23537932	-			

<i>INPP5A</i>	Freeman, 2016 ⁴²⁴	cg02250553	+	Lung adenocarcinoma, Lung squamous cell	cg27286337	+
	Joubert, 2016 ⁴¹⁶	cg09893465 cg12673559 cg19087643 cg26193427	-	Blood		
<i>KALRN</i>	Allione, 2015 ⁴¹⁷	cg05766129	-	Whole blood	cg10178498	-
<i>KCNQ2</i>	Joubert, 2016 ⁴¹⁶	cg13379325	-	Blood	cg09576415	-
<i>LRP1</i>	Dogan, 2014 ⁴²¹	cg09749862	+	Blood	cg11661631	-
<i>LSP1</i>	Joehanes, 2016 ⁴¹⁵	cg04085571	-	Blood	cg27261733	-
	Joubert, 2016 ⁴¹⁶	cg04085571 cg09989681 cg15079934 cg20331155 cg21529477 cg22043296 cg24024833 cg24552015 cg26868156 cg26897904				
	Markunas, 2014 ⁴³³	cg22043296				
<i>LTBP2</i>	Joubert, 2016 ⁴¹⁶	cg27317046	+	Blood	cg18837429	+
<i>MAD1L1</i>	Allione, 2015 ⁴¹⁷	cg01843768	-	Whole blood	cg00766497 cg01827097	- -
		cg17018896 cg23393892				

	Guida, 2015 ⁴²⁵	cg17551891	-	Blood		
		cg06857018	+			
	Joehanes, 2016 ⁴¹⁵	cg08712631	-	Blood		
		cg17551891	-			
		cg01119831	-			
		cg03694580	-			
		cg07065756	-			
		cg09367467	-			
	Joubert, 2016 ⁴¹⁶	cg12492273	-	Blood		
		cg15804231	-			
		cg17551891	-			
		cg21852842	+			
		cg25573368	-			
		cg26580761	-			
	Dogan, 2014 ⁴²¹	cg11168432	+			
		cg06696815				
		cg10133462				
MAEA	Joubert, 2016 ⁴¹⁶	cg18299068	-	Blood	cg18299068	-
		cg18515868				
		cg24973755				
		cg03883572				
MAST4	Joubert, 2016 ⁴¹⁶	cg17278401	-	Blood	cg10629165	+
		cg25399309				
	Allione, 2015 ⁴¹⁷	cg04157979	-	Whole blood		
		cg05197164				
		cg01435643				
MCF2L	Joehanes, 2016 ⁴¹⁵	cg01899620	+	Blood	cg16564828	-
		cg06885459				
		cg14765414				

		cg16822035				
		cg25745937				
	Zeilinger, 2013 ⁴⁴⁸	cg06885459	+	Blood		
MGMT	Allione, 2015 ⁴¹⁷	cg14312783		Whole blood	cg14677612	+
		cg27483317	-			
	Joubert, 2012 ⁴²⁸	cg09993459		Cord blood		
MUC6	Joubert, 2016 ⁴¹⁶	cg01279538	-	Blood	cg00338749	-
		cg07729916				
MYO18A	Dogan, 2014 ⁴²¹	cg11253957	+	Blood	cg25496297	-
NOD2	Joubert, 2016 ⁴¹⁶	cg01020263	+	Blood	cg06181567	-
	Monick, 2012 ⁴³⁴	cg02486161	-	Lymphoblast		
NTN4	Dogan, 2014 ⁴²¹	cg10964388	-	Blood	cg25367249	+
OVOL1	Allione, 2015 ⁴¹⁷	cg10604040	-	Whole blood	cg04293602	-
PCGF3	Joehanes, 2016 ⁴¹⁵	cg10843276	+	Blood	cg23034496	-
	Joubert, 2016 ⁴¹⁶	cg13917589	-			
PIP5K1C	Dogan, 2014 ⁴²¹	cg02322048	+	Blood	cg13670756	-
	Joehanes, 2016 ⁴¹⁵	cg13561409	+			
	Joubert, 2016 ⁴¹⁶	cg21865657	-			
PLEC1	Besingi, 2014 ³⁸⁰	cg09550697	-	Whole blood	cg24507266	+
	Joehanes, 2016 ⁴¹⁵	cg03958308		Blood		
		cg11147309				
		cg12621745	-			
		cg13389508				
	Joubert, 2016 ⁴¹⁶	cg25325005				
		cg07898713	+	Blood		
Monick, 2012 ⁴³⁴	cg14913216	-				
	cg25325005	-	Lymphoblast			

	Zhu, 2016 ⁴⁴⁹	cg13389508 cg25325005	-	Leukocytes			
PPT2	Joubert, 2016 ⁴¹⁶	cg00086577	-	Blood	cg13836183	-	
		cg03995156	-				
		cg05133205	+				
cg06832687		-					
cg08110052		+					
cg14130039		+					
	Kupers, 2015 ⁴²⁹	cg12629909	-	Cord blood			
PTPRN2	Allione, 2015 ⁴¹⁷	cg07305000	-	Whole blood	cg01731811	-	
	Besingi, 2014 ³⁸⁰	cg15340709	+	Whole blood			
	Dogan, 2014 ⁴²¹	cg14743683	-	Blood			
	Joehanes, 2016 ⁴¹⁵	cg00566158	+	Blood			
		cg02223801	-				
		cg05433557	+				
	Joubert, 2012 ⁴²⁸	cg23385492	+	Cord blood			
	Joubert, 2016 ⁴¹⁶	cg02356647	-				
		cg02637474	-				Blood
		cg02660277					
cg14338779							
cg17748769							
		cg18064706					
		cg19350216					
PXDN	Joehanes, 2016 ⁴¹⁵	cg01328473	-	Blood	cg14037665	-	
RAB34	Joehanes, 2016 ⁴¹⁵	cg05668853	-	Blood	cg03452174	+	
RPTOR	Allione, 2015 ⁴¹⁷	cg21289763	-	Whole blood	cg11329058	+	
		cg26469982					
	Dogan, 2014 ⁴²¹	cg02933375	-	Blood	cg18576374	+	

		cg17872658				
	Joehanes, 2016 ⁴¹⁵	cg01498832	+	Blood		
		cg01561259	+			
		cg15228441	-			
		cg18780100	+			
	Joubert, 2016 ⁴¹⁶	cg01498832	+	Blood		
		cg03794617	+			
		cg07126783	+			
		cg08939850	+			
		cg15230985	+			
		cg16541275	-			
		cg16638092	+			
		cg18780100	+			
		cg26360197	-			
		cg27511181	+			
SDK1	Joehanes, 2016 ⁴¹⁵	cg05642264	+	Blood	cg27658048	-
	Joubert, 2012 ⁴²⁸	cg21005410	-	Cord blood		
	Joubert, 2016 ⁴¹⁶	cg16639880	-	Blood		
		cg26180191				
SLC24A4	Joehanes, 2016 ⁴¹⁵	cg01499816	-	Blood	cg07285995	-
TAP1	Joehanes, 2016 ⁴¹⁵	cg02181920	-	Blood	cg24154161	+
		cg10666909				
TMEM177	Philibert, 2013 ⁴³⁷	cg12108912	-	Lymphocyte	cg15845746	+
TRPC4AP	Joehanes, 2016 ⁴¹⁵	cg16151538	+	Blood	cg17206393	+
	Joubert, 2016 ⁴¹⁶					
ZAP70	Joehanes, 2016 ⁴¹⁵	cg09006159	+	Blood	cg03219514	-
		cg12332902				
	Joubert, 2016 ⁴¹⁶	cg15933451	-	Blood		

ZC3H3	Allione, 2015 ⁴¹⁷	cg26361535	-	Whole blood	cg09961689	-
	Guida, 2015 ⁴²⁵	cg26361535	-	Blood		
	Joehanes, 2016 ⁴¹⁵	cg21404980	+	Blood		
		cg26361535	-			
	Joubert, 2016 ⁴¹⁶	cg12688965	-	Blood		
		cg14740860	+			
cg26361535		-				
Zeilinger, 2013 ⁴⁴⁸	cg26361535	-	Blood			
ZMIZ1	Besingi, 2014 ³⁸⁰	cg03450842	-	Whole blood	cg04714939	-
		cg17065712				
	Guida, 2015 ⁴²⁵	cg02743070	-	Blood		
		cg03450842				
		cg18295744				
	Joehanes, 2016 ⁴¹⁵	cg02145310	-	Blood		
		cg03450842	-			
		cg11961495	-			
		cg14371731	+			
		cg14841514	-			
		cg17065712	-			
Joubert, 2016 ⁴¹⁶	cg17823346	-	Blood			
	cg03450842	-				
	cg23865980	+				
ZNF264	Joubert, 2016 ⁴¹⁶	cg16636110	+	Blood	cg18344466	-
		cg18344466				
		cg27176357				

9.3 Appendix 3 – Chapter 5 Supporting Tables and Figures

Table 9.11. Model results for the FDR significant ($p < 2.4 \times 10^{-5}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship between dietary PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status.

Probe ID	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-NL (N=132)		
	β Coefficient	95% Confidence Interval	P Value	β Coefficient	95% Confidence Interval	P Value	β Coefficient	95% Confidence Interval	P Value
cg00575674	0.00038	0.00021; 0.00055	1.33E-05	-0.00025	-0.00052; 1e- 05	1.33E-05	0.00034	-0.00038; 0.00106	0.357
cg01833436	2.00E-04	0.00011; 0.00029	9.55E-06	-2.00E-05	-0.00014; 9e- 05	9.55E-06	0.00025	-0.00016; 0.00066	0.232
cg01931994	-3.00E-04	-0.00043; - 0.00016	1.57E-05	0.00016	-3e-05; 0.00035	1.57E-05	-0.00031	-8e-04; 0.00018	0.217
cg03256904	0.00035	0.00021; 5e- 04	2.64E-06	-0.00012	-3e-04; 7e-05	2.64E-06	6.00E-05	-5e-04; 0.00062	0.833
cg03586847	-0.00035	-0.00051; -2e- 04	9.85E-06	-0.00037	-0.00062; - 0.00012	9.85E-06	-0.00063	-0.00117; - 9e-05	0.023
cg05262877	-2.00E-04	-3e-04; - 0.00011	1.11E-05	-4.00E-05	-0.00019; 0.00011	1.11E-05	-6.00E-05	-0.00032; 2e- 04	0.663
cg06182121	2.00E-04	0.00012; 0.00027	3.32E-07	-0.00011	-0.00023; 1e- 05	3.32E-07	-0.00012	-0.00057; 0.00033	0.609
cg06221222	0.00028	0.00016; 4e- 04	6.88E-06	0.00011	-6e-05; 0.00027	6.88E-06	9.00E-04	0.00035; 0.00144	0.001
cg07173049	0.00029	0.00016; 0.00042	1.81E-05	1.00E-05	-0.00017; 2e- 04	1.81E-05	0.00025	-0.00022; 0.00072	0.293
cg11673291	-0.00023	-0.00032; - 0.00014	2.45E-07	2.00E-05	-0.00011; 0.00014	2.45E-07	-0.00027	-0.00061; 6e- 05	0.106
cg12632832	0.00017	9e-05; 0.00025	1.67E-05	NA	NA; NA	1.67E-05	-0.00018	-6e-04; 0.00024	0.403
cg14287788	0.00061	0.00036; 0.00086	1.73E-06	0.00013	-0.00023; 0.00049	1.73E-06	-0.00026	-0.00108; 0.00056	0.539

cg16515477	0.00027	0.00015; 0.00039	1.58E-05	-0.00019	-0.00037; - 1e-05	1.58E-05	0.00031	-4e-04; 0.00102	0.389
cg16598810	-0.00031	-0.00045; - 0.00017	2.18E-05	0.00016	-0.00012; 0.00043	2.18E-05	0.00061	-3e-05; 0.00125	0.062
cg18936620	0.00019	0.00012; 0.00026	1.44E-07	4.00E-05	-6e-05; 0.00014	1.44E-07	0.00016	-0.00021; 0.00053	0.399
cg20253172	0.00063	0.00034; 0.00091	1.87E-05	0.00055	1e-05; 0.00109	1.87E-05	NA	NA; NA	NA
cg24634746	-0.00074	-0.001; - 0.00047	6.03E-08	0.00041	0.00019; 0.00063	6.03E-08	-8.00E-05	-7e-04; 0.00054	0.794
cg24937768	-0.00016	-0.00024; -9e- 05	2.24E-05	-3.00E-05	-0.00014; 9e- 05	2.24E-05	0.00033	-0.00021; 0.00087	0.229
cg26913155	0.00021	0.00012; 0.00031	7.29E-06	-2.00E-05	-0.00015; 0.00011	7.29E-06	-0.00047	-0.00086; - 7e-05	0.02
cg26916166	-0.00035	-0.00051; - 0.00019	2.13E-05	0.00013	-0.00012; 0.00038	2.13E-05	-0.00057	-0.00162; 0.00047	0.284
cg03135351	0.00047	0.00025; 0.00069	2.06E-05	-2.00E-05	-0.00038; 0.00034	2.06E-05	NA	NA; NA	NA
cg04528072	-0.00018	-0.00027; -1e- 04	2.13E-05	-3.00E-05	-0.00016; 9e- 05	2.13E-05	0.00021	-0.00025; 0.00067	0.362
cg07468585	-0.00071	-0.001; - 0.00043	1.14E-06	-0.00018	-0.00062; 0.00026	1.14E-06	0.00054	-9e-05; 0.00116	0.091
cg08455099	-0.00025	-0.00037; - 0.00014	1.08E-05	-4.00E-05	-0.00022; 0.00014	1.08E-05	0.00032	-2e-04; 0.00083	0.225
cg10288578	-0.00081	-0.00116; - 0.00047	4.51E-06	0	-0.00024; 0.00023	4.51E-06	6.00E-04	9e-05; 0.00112	0.022
cg11419304	0.00015	8e-05; 0.00021	6.00E-06	-0.00013	-0.00022; - 5e-05	6.00E-06	8.00E-05	-0.00028; 0.00045	0.659
cg14356440	0.00027	0.00015; 0.00039	1.42E-05	-0.00015	-0.00033; 3e- 05	1.42E-05	-0.00049	-0.00099; 1e- 05	0.054
cg15517113	0.00055	3e-04; 8e-04	1.44E-05	-3.00E-05	-0.00046; 4e- 04	1.44E-05	NA	NA; NA	NA
cg19067791	0.00082	0.00047; 0.00117	4.20E-06	-0.00038	-0.00087; 0.00011	4.20E-06	0.00017	-0.00012; 0.00046	0.256
cg19697911	-0.00026	-0.00037; - 0.00015	2.04E-06	3.00E-05	-0.00014; 2e- 04	2.04E-06	0.00048	2e-05; 0.00095	0.043

cg21679970	-0.00027	-4e-04; -0.00015	1.71E-05	-6.00E-05	-0.00019; 6e-05	1.71E-05	0.00029	-9e-05; 0.00067	0.135
cg23759710	-0.00012	-0.00017; -7e-05	1.07E-05	-8.00E-05	-0.00016; 1e-05	1.07E-05	9.00E-05	-1e-04; 0.00027	0.367
cg00712146	3.00E-04	0.00016; 0.00044	2.04E-05	-9.00E-05	-3e-04; 0.00012	2.04E-05	0.00015	-5e-04; 8e-04	0.648
cg04016621	0.00025	0.00014; 0.00037	1.64E-05	-3.00E-05	-0.00025; 2e-04	1.64E-05	-0.00052	-0.00093; -0.00012	0.011
cg14595003	-0.00094	-0.00133; -0.00055	2.67E-06	0.00042	-1e-04; 0.00095	2.67E-06	NA	NA; NA	NA
cg21548131	6.00E-04	0.00034; 0.00086	7.67E-06	5.00E-05	-0.00039; 0.00048	7.67E-06	NA	NA; NA	NA
cg00598021	0.00046	0.00025; 0.00066	1.43E-05	-0.00031	-0.00067; 5e-05	1.43E-05	3.00E-04	-0.00077; 0.00137	0.578
cg01142579	0.00022	0.00012; 0.00032	1.72E-05	1.00E-04	-5e-05; 0.00025	1.72E-05	-4.00E-05	-0.00057; 0.00049	0.889
cg05229229	-0.00029	-0.00041; -0.00016	1.21E-05	-2.00E-04	-4e-04; 0	1.21E-05	0.00012	-0.00036; 6e-04	0.617
cg08415391	0.00024	0.00013; 0.00035	1.28E-05	0.00011	-1e-04; 0.00031	1.28E-05	-0.00012	-0.00073; 0.00049	0.705
cg13534095	3.00E-04	0.00017; 0.00044	1.68E-05	0.00016	-7e-05; 0.00039	1.68E-05	-0.00014	-0.00065; 0.00038	0.608
cg16870595	0.00037	0.00022; 0.00052	8.34E-07	0.00024	0; 0.00048	8.34E-07	0.00057	0; 0.00114	0.049
cg27526346	-0.00039	-0.00058; -0.00021	2.14E-05	1.00E-04	-0.00013; 0.00034	2.14E-05	-9.00E-04	-0.00141; -4e-04	0
cg01256674	-0.00027	-0.00039; -0.00014	2.29E-05	-1.00E-04	-0.00029; 1e-04	2.29E-05	0.00042	-6e-05; 0.00091	0.089
cg13652314	-0.00025	-0.00036; -0.00015	2.41E-06	0.00012	-2e-05; 0.00027	2.41E-06	-0.00014	-0.00054; 0.00025	0.467
cg22101141	-0.00032	-0.00046; -0.00019	4.29E-06	0.00016	-3e-05; 0.00035	4.29E-06	0.00019	-0.00059; 0.00096	0.638
cg26650655	-0.00041	-0.00059; -0.00022	2.01E-05	5.00E-05	-0.00025; 0.00035	2.01E-05	-0.00027	-0.00116; 0.00063	0.557
cg00686197	-0.00018	-0.00026; -1e-04	8.39E-06	-0.00016	-3e-04; -3e-05	8.39E-06	3.00E-04	-6e-05; 0.00067	0.102

cg06686436	0.00037	0.00022; 0.00053	3.21E-06	0.00016	-6e-05; 0.00038	3.21E-06	0.00015	-0.00071; 0.00102	0.726
cg06991565	-0.00023	-0.00033; - 0.00013	7.81E-06	-0.00012	-0.00026; 2e- 05	7.81E-06	0.00024	-7e-05; 0.00056	0.131
cg07057617	0.00024	0.00013; 0.00034	7.04E-06	3.00E-05	-0.00013; 2e- 04	7.04E-06	2.00E-04	-0.00022; 0.00063	0.348
cg10552964	-0.00013	-0.00019; -7e- 05	9.77E-06	0.00011	3e-05; 0.00019	9.77E-06	0.00014	-0.00012; 0.00039	0.299
cg13369999	0.00016	9e-05; 0.00023	1.56E-05	1.00E-05	-1e-04; 0.00012	1.56E-05	5.00E-05	-0.00033; 0.00043	0.803
cg15209921	0.00037	2e-04; 0.00054	1.95E-05	-0.00035	-0.00066; - 4e-05	1.95E-05	0.00036	-0.00044; 0.00116	0.379
cg15245581	0.00044	0.00024; 0.00064	2.00E-05	-0.00044	-0.00082; - 7e-05	2.00E-05	0.0011	0.00014; 0.00205	0.024
cg16370701	0.00019	0.00011; 0.00027	5.43E-06	1.00E-04	-2e-05; 0.00021	5.43E-06	-2.00E-04	-0.00063; 0.00023	0.362
cg19954341	-3.00E-04	-0.00044; - 0.00016	1.70E-05	-5.00E-05	-0.00027; 0.00017	1.70E-05	-1.00E-05	-0.00048; 0.00046	0.981
cg22322679	0.00016	9e-05; 0.00023	5.77E-06	9.00E-05	-1e-05; 2e-04	5.77E-06	-2.00E-05	-0.00046; 0.00043	0.945
cg22615992	-0.00049	-7e-04; - 0.00027	6.70E-06	0.00051	0.00021; 8e- 04	6.70E-06	0.00043	-0.00047; 0.00132	0.352
cg24873872	-0.00049	-0.00068; - 0.00031	1.77E-07	-9.00E-05	-0.00038; 0.00019	1.77E-07	2.00E-04	-0.00051; 0.00091	0.579
cg00622763	-0.00043	-0.00063; - 0.00023	2.28E-05	0.00016	-1e-04; 0.00041	2.28E-05	0.00062	-0.00025; 0.00149	0.165
cg03308706	-6.00E-04	-0.00088; - 0.00033	1.27E-05	-0.00039	-0.00079; 1e- 05	1.27E-05	-0.00057	-0.00131; 0.00016	0.127
cg04960665	0.00017	9e-05; 0.00024	1.18E-05	-0.00016	-0.00028; - 5e-05	1.18E-05	4.00E-05	-0.00029; 0.00037	0.808
cg09214099	-0.00021	-3e-04; - 0.00011	1.61E-05	0.00015	0; 3e-04	1.61E-05	0.00044	8e-05; 0.00081	0.018
cg09327911	0.00067	0.00042; 0.00092	2.05E-07	0.00014	-0.00019; 0.00047	2.05E-07	-0.00057	-0.00125; 0.00011	0.102
cg11610702	-0.00019	-0.00028; - 0.00011	5.37E-06	-9.00E-05	-0.00023; 5e- 05	5.37E-06	-2.00E-05	-0.00034; 0.00029	0.887

cg13886135	0.00033	0.00018; 0.00048	1.79E-05	-0.00023	-0.00046; 0	1.79E-05	0.00113	0.00041; 0.00185	0.002
cg21886541	0.00026	0.00014; 0.00037	1.34E-05	1.00E-05	-0.00016; 0.00018	1.34E-05	-0.00018	-0.00053; 0.00018	0.329
cg22322818	0.00054	0.00033; 0.00076	8.09E-07	-1.00E-04	-0.00037; 0.00017	8.09E-07	0.00028	-0.00015; 7e- 04	0.202
cg26022064	-0.00018	-0.00026; - 0.00011	2.18E-06	-0.00015	-0.00028; - 2e-05	2.18E-06	-0.00024	-5e-04; 2e-05	0.065
cg00642970	0.00029	0.00015; 0.00042	2.15E-05	3.00E-05	-0.00018; 0.00023	2.15E-05	0.00034	-0.00016; 0.00085	0.177
cg04689061	0.00057	0.00032; 0.00081	5.69E-06	6.00E-05	-0.00028; 0.00039	5.69E-06	0.00145	0.00052; 0.00237	0.002
cg06102330	-0.00027	-0.00039; - 0.00016	2.17E-06	3.00E-05	-0.00012; 0.00019	2.17E-06	0.00012	-0.00064; 0.00088	0.755
cg12672713	0.00029	0.00016; 0.00042	1.22E-05	-4.00E-05	-0.00021; 0.00013	1.22E-05	-1.00E-05	-0.00051; 0.00049	0.981
cg14213394	-0.00029	-0.00041; - 0.00016	6.32E-06	4.00E-04	0.00018; 0.00062	6.32E-06	0.00059	0.00012; 0.00105	0.014
cg15706250	-0.00024	-0.00035; - 0.00014	7.21E-06	0.00015	1e-05; 0.00029	7.21E-06	-8.00E-05	-0.00049; 0.00032	0.694
cg20585869	-0.00059	-0.00087; - 0.00032	2.29E-05	1.00E-04	-0.00032; 0.00052	2.29E-05	NA	NA; NA	NA
cg02538891	-0.00013	-0.00019; -7e- 05	9.48E-06	2.00E-05	-7e-05; 0.00012	9.48E-06	3.00E-05	-0.00014; 0.00019	0.731
cg08022012	-3.00E-04	-0.00042; - 0.00019	1.52E-07	-0.00017	-0.00031; - 4e-05	1.52E-07	4.00E-05	-0.00028; 0.00036	0.815
cg12044689	0.00024	0.00013; 0.00035	1.45E-05	-2.00E-05	-0.00017; 0.00012	1.45E-05	0.00057	-0.00012; 0.00126	0.104
cg13683194	0.00017	9e-05; 0.00024	1.02E-05	2.00E-05	-9e-05; 0.00013	1.02E-05	8.00E-05	-0.00028; 0.00044	0.662
cg14027524	-0.00022	-3e-04; - 0.00014	1.52E-07	-4.00E-05	-0.00017; 1e- 04	1.52E-07	0.00027	-2e-05; 0.00056	0.07
cg14050363	0.00039	0.00022; 0.00055	5.60E-06	-0.00018	-0.00043; 7e- 05	5.60E-06	0.00027	-0.00056; 0.00109	0.525
cg14276133	0.00019	1e-04; 0.00027	2.35E-05	-0.00015	-3e-04; -1e- 05	2.35E-05	0.00067	0.00024; 0.0011	0.002

cg21207730	0.00064	4e-04; 0.00087	7.89E-08	-4.00E-04	-0.00073; - 7e-05	7.89E-08	-0.00039	-0.00106; 0.00027	0.247
cg03246584	0.00108	0.00067; 0.0015	2.79E-07	-1.00E-04	-0.00055; 0.00036	2.79E-07	NA	NA; NA	NA
cg04203742	0.00035	0.00019; 0.00051	1.55E-05	3.00E-04	1e-04; 0.00051	1.55E-05	0.00015	-0.00037; 0.00068	0.571
cg04366815	0.00021	0.00012; 3e- 04	8.52E-06	-6.00E-05	-0.00021; 1e- 04	8.52E-06	-6.00E-05	-0.00046; 0.00035	0.791
cg14494090	-2.00E-04	-0.00029; - 0.00011	1.98E-05	-7.00E-05	-0.00021; 7e- 05	1.98E-05	0.00013	-2e-04; 0.00046	0.429
cg19154600	-0.00016	-0.00023; -9e- 05	8.19E-06	-2.00E-05	-0.00012; 9e- 05	8.19E-06	-8.00E-05	-0.00034; 0.00018	0.549
cg23083424	-0.00016	-0.00023; -9e- 05	4.84E-06	-1.00E-05	-0.00011; 1e- 04	4.84E-06	-0.00024	-5e-04; 3e-05	0.077
cg03423942	-0.00013	-2e-04; -7e-05	1.40E-05	2.00E-05	-7e-05; 0.00012	1.40E-05	6.00E-05	-0.00016; 0.00028	0.615
cg05269359	0.00062	0.00035; 0.00089	5.17E-06	-0.00034	-0.00071; 4e- 05	5.17E-06	0.00138	0.00038; 0.00238	0.007
cg07832061	0.00063	0.00039; 0.00088	5.70E-07	-0.00018	-0.00055; 0.00019	5.70E-07	NA	NA; NA	NA
cg11310820	-0.00021	-0.00031; - 0.00012	1.44E-05	-9.00E-05	-0.00025; 7e- 05	1.44E-05	-3.00E-05	-4e-04; 0.00033	0.854
cg12438576	-0.00026	-0.00037; - 0.00015	5.40E-06	-8.00E-05	-0.00026; 1e- 04	5.40E-06	-5.00E-05	-6e-04; 0.00049	0.847
cg15659420	3.00E-04	0.00016; 0.00044	1.48E-05	-6.00E-05	-0.00024; 0.00013	1.48E-05	0.00021	-0.00046; 0.00088	0.543
cg22996681	0.00024	0.00013; 0.00036	1.85E-05	-8.00E-05	-0.00025; 9e- 05	1.85E-05	5.00E-04	-2e-05; 0.00102	0.059
cg24413662	-0.00031	-0.00046; - 0.00017	1.47E-05	0.00011	-1e-04; 0.00032	1.47E-05	0.00042	-0.00019; 0.00104	0.178
cg26327442	0.00024	0.00013; 0.00035	1.74E-05	2.00E-05	-0.00013; 0.00016	1.74E-05	5.00E-05	-0.00045; 0.00055	0.847
cg01414572	0.00025	0.00014; 0.00036	1.45E-05	-4.00E-05	-0.00021; 0.00012	1.45E-05	-9.00E-05	-0.00068; 5e- 04	0.767
cg05419385	0.00022	0.00014; 3e- 04	1.20E-07	4.00E-05	-7e-05; 0.00015	1.20E-07	4.00E-05	-0.00034; 0.00042	0.829

cg09111484	0.00026	0.00014; 0.00038	1.47E-05	-3.00E-05	-2e-04; 0.00015	1.47E-05	-0.00023	-0.00077; 0.00031	0.401
cg09260514	0.00034	0.00019; 0.00049	7.32E-06	-8.00E-05	-0.00028; 0.00013	7.32E-06	0.00031	-3e-04; 0.00092	0.322
cg10832076	0.00031	0.00018; 0.00044	3.78E-06	0.00025	6e-05; 0.00043	3.78E-06	-0.00041	-0.00096; 0.00014	0.147
cg11294269	-2.00E-04	-3e-04; - 0.00011	1.77E-05	3.00E-05	-0.00019; 0.00024	1.77E-05	-5.00E-05	-2e-04; 1e-04	0.545
cg14733535	-3.00E-04	-0.00042; - 0.00017	6.21E-06	-7.00E-05	-0.00031; 0.00017	6.21E-06	-0.00012	-0.00069; 0.00046	0.687
cg18202562	0.00028	0.00015; 0.00041	1.38E-05	-0.00021	-0.00042; 0	1.38E-05	3.00E-04	-0.00017; 0.00077	0.21
cg18827332	-0.00031	-0.00043; - 0.00018	1.57E-06	-0.00014	-0.00032; 4e- 05	1.57E-06	2.00E-05	-0.00038; 0.00042	0.917
cg20901167	-0.00034	-5e-04; - 0.00019	1.68E-05	-2.00E-05	-0.00023; 0.00019	1.68E-05	-0.00051	-0.00142; 4e- 04	0.269
cg21858255	-0.00052	-0.00074; -3e- 04	5.13E-06	2.00E-04	-0.00014; 0.00054	5.13E-06	-0.00065	-0.00143; 0.00013	0.102
cg23320862	-0.00021	-0.00031; - 0.00012	1.61E-05	-2.00E-05	-0.00013; 1e- 04	1.61E-05	0.00016	-6e-05; 0.00037	0.161
cg01226614	-0.00022	-0.00031; - 0.00012	5.45E-06	-9.00E-05	-0.00024; 5e- 05	5.45E-06	0.00023	-9e-05; 0.00054	0.154
cg13147013	0.00018	0.00011; 0.00026	3.08E-06	-5.00E-05	-0.00018; 8e- 05	3.08E-06	2.00E-04	-7e-05; 0.00048	0.152
cg13487183	0.00033	0.00018; 0.00048	1.70E-05	-0.00019	-0.00041; 3e- 05	1.70E-05	-0.00018	-0.00106; 7e- 04	0.694
cg18481241	2.00E-04	0.00011; 0.00029	1.75E-05	3.00E-05	-8e-05; 0.00014	1.75E-05	0.00017	-2e-04; 0.00054	0.364
cg26658125	0.00015	8e-05; 0.00021	2.36E-05	-9.00E-05	-0.00019; 0	2.36E-05	0.00014	-0.00027; 0.00055	0.501
cg03523785	-0.00019	-0.00028; - 0.00011	4.69E-06	NA	NA; NA	4.69E-06	0.00011	-0.00037; 0.00058	0.66
cg05239310	-0.00025	-0.00036; - 0.00014	9.50E-06	-1.00E-04	-0.00026; 7e- 05	9.50E-06	0.00052	-8e-05; 0.00112	0.09
cg05881436	-3.00E-04	-0.00042; - 0.00017	4.87E-06	7.00E-05	-0.00013; 0.00026	4.87E-06	0.00056	0.00016; 0.00096	0.006

cg10001646	0.00061	4e-04; 0.00083	2.45E-08	0.00017	3e-05; 0.00031	2.45E-08	0.00029	-0.00012; 7e-04	0.169
cg12940991	0.00032	0.00021; 0.00044	7.91E-08	2.00E-04	5e-05; 0.00035	7.91E-08	-0.00023	-0.00082; 0.00037	0.454
cg27158340	-2.00E-04	-0.00028; - 0.00011	5.34E-06	8.00E-05	-6e-05; 0.00021	5.34E-06	0.00034	1e-05; 0.00066	0.042
cg01273232	-4.00E-04	-0.00058; - 0.00021	2.30E-05	0.00028	-1e-05; 0.00057	2.30E-05	4.00E-04	-0.00024; 0.00105	0.219
cg01321816	0.00036	2e-04; 0.00052	1.30E-05	0.00015	-9e-05; 0.00039	1.30E-05	0.00045	-0.00033; 0.00124	0.257
cg03829137	-0.00047	-0.00067; - 0.00026	8.06E-06	0.00012	-0.00023; 0.00047	8.06E-06	0.00039	-0.00024; 0.00102	0.23
cg04624362	0.00016	9e-05; 0.00024	1.93E-05	2.00E-05	-1e-04; 0.00013	1.93E-05	2.00E-04	-0.00013; 0.00052	0.235
cg02233835	-0.00021	-3e-04; - 0.00012	1.13E-05	5.00E-05	-1e-04; 0.00021	1.13E-05	0.00048	8e-05; 0.00088	0.019
cg02302035	0.00036	2e-04; 0.00052	1.22E-05	-3.00E-05	-0.00024; 0.00018	1.22E-05	-0.00056	-0.00125; 0.00013	0.113
cg05570739	0.00054	0.00029; 0.00078	1.81E-05	-0.00072	-0.00111; - 0.00033	1.81E-05	NA	NA; NA	NA
cg06420305	-0.00032	-0.00046; - 0.00017	1.72E-05	-0.00019	-0.00045; 6e-05	1.72E-05	2.00E-04	-0.00045; 0.00085	0.546
cg08394248	0.00035	0.00019; 0.00051	2.30E-05	7.00E-05	-0.00014; 0.00028	2.30E-05	0.00073	-0.00032; 0.00178	0.173
cg26780022	-4.00E-04	-0.00057; - 0.00024	1.30E-06	5.00E-05	-2e-04; 0.00029	1.30E-06	-0.00024	-0.00069; 0.00022	0.312
cg08870588	0.00019	0.00011; 0.00028	1.10E-05	NA	NA; NA	1.10E-05	-2.00E-04	-0.00069; 0.00029	0.421
cg10999136	0.00037	0.00021; 0.00054	1.10E-05	-9.00E-05	-0.00035; 0.00016	1.10E-05	0.00082	0.00011; 0.00154	0.024
cg12187586	-0.00045	-0.00065; - 0.00025	8.28E-06	-8.00E-05	-0.00041; 0.00024	8.28E-06	-0.00029	-0.00078; 2e-04	0.242
cg12550399	0.00023	0.00015; 0.00031	6.10E-09	-0.00015	-0.00026; - 3e-05	6.10E-09	0.00032	0; 0.00065	0.052
cg14949292	-0.00071	-0.00103; - 0.00039	1.18E-05	0.00012	-0.00037; 0.00061	1.18E-05	-0.00012	-0.00092; 0.00067	0.764

cg15159588	-0.00026	-0.00037; -0.00014	2.26E-05	5.00E-05	-8e-05; 0.00018	2.26E-05	0.00012	-0.00015; 0.00039	0.382
cg16548154	-0.00015	-0.00021; -9e-05	1.85E-06	-2.00E-05	-0.00011; 7e-05	1.85E-06	3.00E-05	-0.00017; 0.00023	0.781
cg17301311	-0.00019	-0.00028; -0.00011	5.66E-06	-3.00E-05	-0.00016; 1e-04	5.66E-06	2.00E-05	-0.00024; 0.00029	0.875
cg19935128	-0.00016	-0.00023; -9e-05	1.52E-05	-5.00E-05	-0.00015; 5e-05	1.52E-05	1.00E-04	-0.00019; 4e-04	0.501
cg25930644	0.00021	0.00012; 0.00029	1.52E-06	-0.00015	-0.00027; -3e-05	1.52E-06	0.00046	7e-05; 0.00085	0.022
cg14307471	0.00044	0.00025; 0.00063	4.35E-06	8.00E-05	-2e-04; 0.00036	4.35E-06	NA	NA; NA	NA
cg00086493	-0.00023	-0.00033; -0.00013	2.57E-06	-5.00E-05	-0.00019; 8e-05	2.57E-06	-0.00021	-0.00055; 0.00014	0.243
cg00910067	-3.00E-04	-0.00042; -0.00017	3.59E-06	8.00E-05	-8e-05; 0.00024	3.59E-06	0.00012	-0.00037; 6e-04	0.639
cg02644494	-0.00019	-0.00027; -1e-04	1.42E-05	-2.00E-05	-0.00015; 0.00011	1.42E-05	-6.00E-05	-0.00034; 0.00022	0.659
cg03013172	-0.00025	-0.00036; -0.00014	1.30E-05	8.00E-05	-9e-05; 0.00025	1.30E-05	-0.00031	-0.00078; 0.00016	0.2
cg04253011	-0.00016	-0.00023; -9e-05	1.42E-05	-3.00E-05	-0.00014; 9e-05	1.42E-05	-0.00019	-0.00046; 9e-05	0.182
cg04351156	-0.00017	-0.00025; -1e-04	4.71E-06	-3.00E-05	-0.00014; 9e-05	4.71E-06	-0.00011	-0.00028; 5e-05	0.183
cg04556210	-0.00013	-0.00019; -7e-05	1.11E-05	2.00E-05	-6e-05; 1e-04	1.11E-05	0.00043	7e-05; 8e-04	0.019
cg08065565	-0.00017	-0.00025; -1e-04	8.29E-06	-4.00E-05	-0.00016; 8e-05	8.29E-06	0.00018	-6e-05; 0.00042	0.134
cg12568707	-0.00049	-0.00071; -0.00027	1.07E-05	0.00038	1e-05; 0.00075	1.07E-05	0.00047	5e-05; 0.00088	0.029
cg12610917	-0.00035	-0.00051; -0.00019	1.59E-05	-0.00016	-4e-04; 8e-05	1.59E-05	1.00E-04	-0.00042; 0.00061	0.709
cg14930737	0.00059	0.00033; 0.00084	6.13E-06	0.00046	0.00011; 0.00082	6.13E-06	-0.00107	-0.00202; -0.00012	0.027
cg17583504	-0.00031	-0.00045; -0.00017	1.20E-05	0.00044	0.00019; 0.00069	1.20E-05	-3.00E-05	-0.00051; 0.00045	0.903

cg19007908	-0.00036	-0.00051; - 0.00021	2.83E-06	0.00022	3e-05; 0.00041	2.83E-06	-0.00034	-0.00078; 1e- 04	0.134
cg19864007	0.00073	0.00043; 0.00103	2.04E-06	0.00017	-0.00036; 7e- 04	2.04E-06	NA	NA; NA	NA
cg07727233	0.00102	0.00063; 0.0014	2.16E-07	-0.00053	-0.0011; 4e- 05	2.16E-07	0.00026	-0.00031; 0.00084	0.372
cg08553950	0.00018	0.00011; 0.00026	3.77E-06	7.00E-05	-5e-05; 2e-04	3.77E-06	1.00E-04	-3e-04; 0.00051	0.622
cg17512522	0.00025	0.00015; 0.00036	1.57E-06	9.00E-05	-3e-05; 0.00021	1.57E-06	0.00016	-4e-04; 0.00073	0.567
cg08073527	0.00031	0.00019; 0.00043	7.94E-07	-0.00056	-0.00081; - 0.00032	7.94E-07	0.00012	-0.00018; 0.00041	0.43
cg13588826	-0.00036	-5e-04; - 0.00023	1.36E-07	-4.00E-05	-0.00021; 0.00012	1.36E-07	6.00E-05	-0.00055; 0.00066	0.854
cg13882606	0.00061	0.00034; 0.00089	1.16E-05	0.00044	3e-05; 0.00085	1.16E-05	NA	NA; NA	NA
cg19312314	0.00125	0.00067; 0.00182	2.03E-05	-0.00014	-0.00106; 0.00077	2.03E-05	0.00147	0.00018; 0.00276	0.025
cg19902195	0.00022	0.00012; 0.00031	2.03E-05	-7.00E-05	-0.00024; 9e- 05	2.03E-05	3.00E-05	-0.00044; 0.00049	0.913
cg10898989	-0.00019	-0.00028; - 0.00011	4.34E-06	-1.00E-04	-0.00022; 1e- 05	4.34E-06	1.00E-04	-0.00022; 0.00043	0.531
cg12130797	-0.00019	-0.00026; - 0.00011	1.44E-06	2.00E-05	-9e-05; 0.00013	1.44E-06	6.00E-05	-0.00023; 0.00035	0.686
cg14043774	0.00017	9e-05; 0.00025	2.08E-05	-0.00014	-0.00022; - 5e-05	2.08E-05	0.00036	0; 0.00071	0.047
cg21811450	-0.00034	-0.00049; - 0.00019	8.78E-06	0.00014	-8e-05; 0.00037	8.78E-06	-0.00014	-0.00084; 0.00056	0.692
cg25795369	0.00033	0.00019; 0.00047	1.87E-06	-0.00012	-0.00031; 7e- 05	1.87E-06	0.00033	-0.00034; 0.00099	0.335

Table 9.12. Table of characteristics of probes found to be significantly associated with dietary PAH8 exposure at the FDR level ($p < 2.4 \times 10^{-5}$) in the training dataset.

Probe ID	Chromosome	Position	UCSC RefGene Name	Gene Location	Relation to CpG Island	Methylation Change Direction
cg00575674	1	61314297				+
cg01833436	1	243653490	<i>SDCCAG8; AKT3</i>	Body; 3'UTR	South Shore	+
cg01931994	1	177225850	<i>FAM5B</i>	Body		-
cg03256904	1	3696735				+
cg03586847	1	40726012	<i>ZMPSTE24</i>	Body	South Shelf	-
cg05262877	1	42631835	<i>GUCA2A</i>	TSS1500		-
cg06182121	1	3080723	<i>PRDM16</i>	Body	North Shore	+
cg06221222	1	94147831	<i>BCAR3</i>	TSS1500	South Shore	+
cg07173049	1	7289937	<i>CAMTA1</i>	Body		+
cg11673291	1	36787145			Island	-
cg12632832	1	157013346	<i>ARHGEF11</i>	Body	North Shore	+
cg14287788	1	6284844	<i>ICMT</i>	3'UTR		+
cg16515477	1	49511332	<i>AGBL4</i>	Body		+
cg16598810	1	68962176	<i>DEPDC1</i>	Body	North Shore	-
cg18936620	1	43811019	<i>MPL</i>	Body	North Shelf	+
cg20253172	1	3107290	<i>PRDM16</i>	Body	South Shelf	+
cg24634746	1	7538723	<i>CAMTA1</i>	Body		-
cg24937768	1	2092853	<i>PRKCZ</i>	Body		-

cg26913155	1	3128175	<i>PRDM16</i>	Body		+
cg26916166	1	183387420	<i>NMNAT2</i>	5'UTR	Island	-
cg03135351	2	29338258	<i>CLIP4</i>	TSS200	Island	+
cg04528072	2	27371642	<i>TCF23</i>	TSS1500	North Shore	-
cg07468585	2	56192635				-
cg08455099	2	95663959			Island	-
cg10288578	2	16816994	<i>FAM49A</i>	5'UTR		-
cg11419304	2	69248344	<i>ANTXR1</i>	Body		+
cg14356440	2	135050894	<i>MGAT5</i>	Body		+
cg15517113	2	620265			North Shore	+
cg19067791	2	166809971	<i>TTC21B</i>	Body	Island	+
cg19697911	2	241080057	<i>OTOS</i>	5'UTR		-
cg21679970	2	10581770	<i>ODC1</i>	Body		-
cg23759710	2	42990957	<i>OXER1</i>	1stExon		-
cg00712146	3	65340538	<i>MAGI1</i>	3'UTR	North Shore	+
cg04016621	3	141495947	<i>GRK7</i>	TSS1500	North Shore	+
cg14595003	3	129694156	<i>TRH</i>	5'UTR	Island	-
cg21548131	3	173639566	<i>NLGN1</i>	Body		+
cg00598021	4	10113794	<i>WDR1</i>	Body	North Shelf	+
cg01142579	4	183769187				+
cg05229229	4	3644703			South Shore	-
cg08415391	4	20574555	<i>SLIT2</i>	Body		+

cg13534095	4	12926002				+
cg16870595	4	175839423	<i>ADAM29</i>	TSS200		+
cg27526346	4	21699534	<i>KCNIP4</i>	5'UTR;TSS1500		-
cg01256674	5	72716074			South Shore	-
cg13652314	5	131893144	<i>RAD50</i>	5'UTR	South Shore	-
cg22101141	5	39425191	<i>DAB2</i>	1stExon	Island	-
cg26650655	5	88178400	<i>MEF2C</i>	5'UTR	North Shore	-
cg00686197	6	31733619	<i>C6orf27</i>	Body		-
cg06686436	6	32138008	<i>AGPAT1</i>	Body	South Shelf	+
cg06991565	6	31733799	<i>C6orf27</i>	Body		-
cg07057617	6	170405951			South Shelf	+
cg10552964	6	35991802	<i>SLC26A8</i>	5'UTR	North Shelf	-
cg13369999	6	29711465	<i>LOC285830</i>	Body		+
cg15209921	6	29430506	<i>OR2H1</i>	3'UTR		+
cg15245581	6	96651792	<i>FUT9</i>	Body		+
cg16370701	6	43029051	<i>KLC4</i>	5'UTR	South Shore	+
cg19954341	6	166583523	<i>T</i>	TSS1500	South Shore	-
cg22322679	6	33244178	<i>B3GALT4 ;RPS18</i>	TSS1500; Body	North Shore	+
cg22615992	6	164093099			Island	-
cg24873872	6	36391494	<i>PXT1</i>	Body	South Shore	-
cg00622763	7	128556770			South Shore	-
cg03308706	7	91763433	<i>CYP51A1</i>	5'UTR	Island	-

cg04960665	7	885594	<i>UNC84A</i>	Body	North Shore	+
cg09214099	7	72791740			Island	-
cg09327911	7	6617264	<i>ZDHHC4</i>	5'UTR	Island	+
cg11610702	7	39773227	<i>LOC349114</i>	Body	Island	-
cg13886135	7	31126479	<i>ADCYAP1R1</i>	Body		+
cg21886541	7	155616422				+
cg22322818	7	55497587	<i>LANCL2</i>	Body		+
cg26022064	7	98739782	<i>SMURF1</i>	Body	North Shore	-
cg00642970	8	125953928	<i>LOC157381</i>	TSS1500		+
cg04689061	8	79427993	<i>PKIA</i>	TSS1500	North Shore	+
cg06102330	8	72756932	<i>MSC</i>	TSS1500	South Shore	-
cg12672713	8	30889709	<i>PURG; WRN</i>	1stExon; TSS1500		+
cg14213394	8	38508585			Island	-
cg15706250	8	41583321	<i>ANK1</i>	Body	Island	-
cg20585869	8	24772333	<i>NEFM</i>	TSS200	Island	-
cg02538891	9	139549426			North Shore	-
cg08022012	9	138678461	<i>KCNT1</i>	Body	Island	-
cg12044689	9	97203357	<i>HIATL1</i>	Body		+
cg13683194	9	104237697	<i>C9orf125</i>	3'UTR		+
cg14027524	9	140120587	<i>C9orf169</i>	3'UTR	South Shelf	-
cg14050363	9	89958644				+
cg14276133	9	89061100				+

cg21207730	9	86821905				+
cg03246584	10	134663467				+
cg04203742	10	2371007				+
cg04366815	10	101690668	<i>NCRNA00093; DNMBP</i>	Body; Body		+
cg14494090	10	134972969	<i>KNDC1</i>	TSS1500	North Shore	-
cg19154600	10	75415868	<i>SYNPO2L</i>	TSS200		-
cg23083424	10	75415875	<i>SYNPO2L</i>	TSS200		-
cg03423942	11	77908087	<i>USP35</i>	Body	Island	-
cg05269359	11	118004193	<i>SCN4B</i>	3'UTR		+
cg07832061	11	67236265	<i>TMEM134</i>	Body	Island	+
cg11310820	11	62648102	<i>SLC3A2</i>	Body	North Shore	-
cg12438576	11	89232216	<i>NOX4</i>	5'UTR		-
cg15659420	11	20034979	<i>NAV2</i>	Body		+
cg22996681	11	15847036				+
cg24413662	11	122311293				-
cg26327442	11	82447767			South Shelf	+
cg01414572	12	5248588			North Shelf	+
cg05419385	12	27352945				+
cg09111484	12	80749651				+
cg09260514	12	50225376	<i>LOC100286844</i>	Body	South Shelf	+
cg10832076	12	21418929	<i>SLCO1A2</i>	3'UTR		+
cg11294269	12	132685428	<i>GALNT9</i>	Body	North Shelf	-
cg14733535	12	99287793	<i>ANKS1B</i>	Body	North Shore	-

cg18202562	12	100375604	<i>ANKS1B</i>	Body	North Shelf	+
cg18827332	12	103344506			Island	-
cg20901167	12	132946255				-
cg21858255	12	104609609	<i>TXNRD1</i>	1stExon	Island	-
cg23320862	12	114843932	<i>TBX5</i>	TSS200	North Shore	-
cg01226614	13	44947593	<i>SERP2</i>	TSS1500	Island	-
cg13147013	13	99852409	<i>UBAC2</i>	TSS1500	Island	+
cg13487183	13	78428203				+
cg18481241	13	48893727	<i>RB1</i>	Body	Island	+
cg26658125	13	112885464				+
cg03523785	14	29234981	<i>FOXG1</i>	TSS1500	Island	-
cg05239310	14	95651984	<i>CLMN</i>	3'UTR		-
cg05881436	14	62331619			Island	-
cg10001646	14	24683737	<i>MDP1; CHMP4A</i>	Body	South Shore	+
cg12940991	14	77525744				+
cg27158340	14	105603389			Island	-
cg01273232	15	93652756			Island	-
cg01321816	15	91358514	<i>BLM</i>	3'UTR	North Shelf	+
cg03829137	15	93653073			South Shore	-
cg04624362	15	90730573	<i>SEMA4B</i>	5'UTR	South Shelf	+
cg02233835	16	89156772			North Shelf	-
cg02302035	16	54155477			North Shore	+
cg05570739	16	89757371	<i>CDK10</i>	Body	North Shelf	+

cg06420305	16	78133211	<i>WWOX</i>	TSS1500	Island	-
cg08394248	16	83848109				+
cg26780022	16	1336537				-
cg08870588	17	71232538	<i>C17orf80</i>	Body	South Shelf	+
cg10999136	17	76988558	<i>CANT1</i>	3'UTR	North Shore	+
cg12187586	17	2627661			Island	-
cg12550399	17	19482275	<i>SLC47A1</i>	3'UTR	North Shore	+
cg14949292	17	78079608	<i>GAA</i>	Body	Island	-
cg15159588	17	26672798	<i>TNFAIP1</i>	3'UTR		-
cg16548154	17	74565757	<i>ST6GALNAC2</i>	Body		-
cg17301311	17	48641896	<i>CACNA1G</i>	Body	South Shelf	-
cg19935128	17	79507243	<i>C17orf70</i>	3'UTR	South Shelf	-
cg25930644	17	8531915	<i>MYH10</i>	5'UTR	North Shore	+
cg14307471	18	31432117	<i>NOL4</i>	3'UTR		+
cg00086493	19	51535348	<i>KLK12</i>	Body	Island	-
cg00910067	19	33717545	<i>SLC7A10</i>	TSS1500	Island	-
cg02644494	19	6412686			North Shelf	-
cg03013172	19	5688456	<i>HSD11B1L</i>	3'UTR	North Shore	-
cg04253011	19	39906496	<i>PLEKHG2</i>	Body	South Shelf	-
cg04351156	19	10562415	<i>PDE4A</i>	Body		-
cg04556210	19	47840110	<i>GPR77</i>	TSS1500		-

cg08065565	19	8008274	<i>TIMM44</i>	Body	Island	-
cg12568707	19	19042904	<i>HOMER3</i>	Body	Island	-
cg12610917	19	46387992	<i>IRF2BP1</i>	1stExon	Island	-
cg14930737	19	58109761	<i>ZNF530</i>	TSS1500	North Shore	+
cg17583504	19	49669542	<i>TRPM4</i>	Body	Island	-
cg19007908	19	49686020	<i>TRPM4</i>	Body	Island	-
cg19864007	19	52408259	<i>ZNF649</i>	TSS200		+
cg07727233	20	33543679	<i>GSS</i>	TSS200	Island	+
cg08553950	20	36012016	<i>SRC</i>	5'UTR	North Shore	+
cg17512522	20	6195391			South Shore	+
cg08073527	21	43256581	<i>PRDM15</i>	Body	South Shore	+
cg13588826	21	47533197	<i>COL6A2</i>	Body	South Shore	-
cg13882606	21	17101011	<i>USP25</i>	TSS1500	North Shore	+
cg19312314	21	44473962	<i>CBS</i>	3'UTR	Island	+
cg19902195	21	46357106	<i>C21orf67</i>	Body	North Shelf	+
cg10898989	22	45060369			North Shelf	-
cg12130797	22	20143041			Island	-
cg14043774	22	36013398	<i>MB</i>	TSS200		+
cg21811450	22	47022471	<i>GRAMD4</i>	TSS200	Island	-
cg25795369	22	50094962			North Shelf	+

Table 9.13. Table comparing results published by Tryndyak et al. (2018)³²⁵ and the combined PAH8 exposure EWAS results

<u>Gene Name</u>	Tryndyak <i>et al.</i> 2018 ³²⁵					Combined PAH8 EWAS Results		
	<u>Chromosome</u>	<u>Start</u>	<u>End</u>	<u>Genomic Location</u>	<u>Direction of Methylation Change</u>	<u>Probe ID</u>	<u>Genomic Location</u>	<u>Direction of Methylation Change</u>
<i>CAMTA1</i>	chr1	7728779	7728879	Promoter	+	cg07173049	Body	+
						cg24634746	Body	-
<i>NAV2</i>	chr11	19955509	19955640	Exon	-	cg15659420	Body	+

Table 9.14. Table showing overlaps between results of the dietary PAH8 exposure EWAS, and results from published smoking EWAS. Overlaps were identified by looking for exact CpG probes and by looking for probes with the same genes.

Gene	Study	CpG	Direction	Tissue	CpG	Direction	
<i>ADCYAP1R1</i>	Lee, 2016 ⁴⁵⁰	cg20165074	-	Blood	cg13886135	+	
<i>AGBL4</i>	Joubert, 2016 ⁴¹⁶	cg12127196	+	Blood	cg16515477	+	
		cg16260421					
<i>AKT3</i>	Guida, 2015 ⁴²⁵	cg11314684	-	Blood	cg01833436	+	
	Harlid, 2014 ⁴²⁶	cg11314684	-				
	Joehanes, 2016 ⁴¹⁵	cg04221461	+				
		cg11314684	-				
		cg11496569	-				
Sun, 2013 ⁴⁴²	cg11314684	-					
<i>ANK1</i>	Joehanes, 2016 ⁴¹⁵	cg12634208	+	Blood	cg15706250	-	
	Joubert, 2016 ⁴¹⁶	cg01453458					
<i>C6orf27</i>	Guida, 2015 ⁴²⁵	cg19868593	-	Blood	cg00686197 cg06991565	-	
	Joehanes, 2016 ⁴¹⁵	cg08409562	+				
	Joubert, 2016 ⁴¹⁶	cg24065328	-				
<i>CACNA1G</i>	Joubert, 2016 ⁴¹⁶	cg20271361	-	Blood	cg17301311	-	
<i>CAMTA1</i>	Joehanes, 2016 ⁴¹⁵	cg23972860	+	Blood	cg07173049 cg24634746	+	
		cg00452133	+				
		cg06077003	+				
		Joubert, 2016 ⁴¹⁶	cg11755201				+
		cg12097989	-				
		cg20800117	+				
cg24999973	-						
<i>CYP51A1</i>	Allione, 2015 ⁴¹⁷	cg10655371	-	Whole blood	cg03308706	-	

DAB2	Monick, 2012 ⁴³⁴	cg17576603	+	Alveolar macrophage	cg22101141	-
DEPDC1	Joubert, 2016 ⁴¹⁶	cg14609721	-	Blood	cg16598810	-
FAM49A	Dogan, 2014 ⁴²¹	cg21646084	-	Blood	cg10288578	-
	Joubert, 2016 ⁴¹⁶	cg07091529	+			
		cg07712663	-			
		cg10106284	+			
		cg10502303	+			
GALNT9	Allione, 2015 ⁴¹⁷	cg17320856		Whole blood	cg11294269	-
	Dogan, 2014 ⁴²¹	cg19834585	-	Blood		
	Joubert, 2016 ⁴¹⁶	cg07782603		Blood		
GPR77	Joehanes, 2016 ⁴¹⁵	cg16734795		Blood	cg04556210	-
	Joubert, 2016 ⁴¹⁶	cg24217159	+			
GSS	Joubert, 2016 ⁴¹⁶	cg00352780	+	Blood	cg07727233	+
	Sun, 2013 ⁴⁴²	cg08743392	-			
HOMER3	Joubert, 2016 ⁴¹⁶	cg11601336	+	Blood	cg12568707	-
IRF2BP1	Joehanes, 2016 ⁴¹⁵	cg08097614	-	Blood	cg12610917	-
KNDC1	Joubert, 2016 ⁴¹⁶	cg01258050	-	Blood	cg14494090	-
LOC157381	Joubert, 2016 ⁴¹⁶	cg01209566	+	Blood	cg00642970	+
LOC285830	Joehanes, 2016 ⁴¹⁵	cg23606396	+	Blood	cg13369999	+
MEF2C	Dogan, 2014 ⁴²¹	cg16105594		Blood	cg26650655	-
	Joubert, 2016 ⁴¹⁶	cg06835212	+			
MYH10	Dogan, 2014 ⁴²¹	cg06557376		Blood	cg25930644	+
	Joehanes, 2016 ⁴¹⁵	cg09975715	+			
NAV2	Guida, 2015 ⁴²⁵	cg04039799	-	Blood	cg15659420	+
	Ivorra, 2015 ⁴²⁷	cg01249134		Cord blood, Blood		
		cg03529555	+			

	Joehanes, 2016 ⁴¹⁵	cg12535090	+	Blood		
		cg03220447				
	Joubert, 2016 ⁴¹⁶	cg04039799	-	Blood		
		cg12711760				
	Zeilinger, 2013 ⁴⁴⁸	cg04039799	-	Blood		
ODC1	Joehanes, 2016 ⁴¹⁵	cg26236235	-	Blood	cg21679970	-
PRDM15	Joehanes, 2016 ⁴¹⁵	cg18151030	-	Blood	cg08073527	+
	Allione, 2015 ⁴¹⁷	cg00109293				
		cg25372239	-	Whole blood		
	Dogan, 2014 ⁴²¹	cg00068377	+	Blood		
		cg15386853	-			
		cg00806481	-			
		cg03126058	-			
		cg04134748	+			
	Joehanes, 2016 ⁴¹⁵	cg10493186	+	Blood		
		cg12297125	-			
		cg22510139	-			
PRDM16		cg25618424	-		cg06182121	
		cg01261194	-		cg20253172	+
		cg01418153	-		cg26913155	
		cg01431482	-			
		cg03254465	-			
	Joubert, 2016 ⁴¹⁶	cg04134748	-	Blood		
		cg05804170	-			
		cg08262220	-			
		cg11138362	-			
		cg11731671	-			
		cg12133962	-			

	cg12408250	-				
	cg12436196	-				
	cg12441214	+				
	cg13388191	-				
	cg13393782	-				
	cg15090440	-				
	cg17001566	-				
	cg17445936	-				
	cg17940849	-				
	cg18369939	-				
	cg18509466	-				
	cg19243842	-				
	cg19904265	-				
	cg21848084	-				
	cg22122862	-				
	cg22510139	-				
	cg22726349	-				
	cg22729726	-				
	cg24939838	+				
	cg25618424	-				
	cg26425711	-				
	Kupers, 2015 ⁴²⁹	cg252153667	-	Cord blood		
	Allione, 2015 ⁴¹⁷	cg16059943	-	Whole blood		
	Dogan, 2014 ⁴²¹	cg09180820	-	Blood		
		cg23629792	-			
PRKCZ	Freeman, 2016 ⁴²⁴	cg11345323	-	Lung adenocarcinoma,	cg24937768	-
		cg22865720	-	Lung squamous cell		
	Joehanes, 2016 ⁴¹⁵	cg24842354	-	Blood		
	Joubert, 2016 ⁴¹⁶	cg02393699	+	Blood		

		cg09225489	-			
		cg27264462	+			
PXT1	Joubert, 2016 ⁴¹⁶	cg23678210	+	Blood	cg24873872	-
	Dogan, 2014 ⁴²¹	cg27182159				
RPS18	Joehanes, 2016 ⁴¹⁵	cg12583553	-	Blood	cg22322679	+
		cg27182159				
	Joubert, 2016 ⁴¹⁶	cg27182159				
SEMA4B	Joehanes, 2016 ⁴¹⁵	cg24924577	+	Blood	cg04624362	+
	Joubert, 2016 ⁴¹⁶	cg25913761	-			
SLC26A8	Joubert, 2016 ⁴¹⁶	cg23807646	+	Blood	cg10552964	-
SLCO1A2	Dogan, 2014 ⁴²¹	cg20529334	+	Blood	cg10832076	+
SRC	Joubert, 2016 ⁴¹⁶	cg01141721	+	Blood	cg08553950	+
ST6GALNAC2	Shenker, 2013 ⁴⁴⁰	cg14385325	-	Blood	cg16548154	-
SYNPO2L	Joubert, 2016 ⁴¹⁶	cg27550918	-	Blood	cg19154600	-
					cg23083424	
TRPM4	Dogan, 2014 ⁴²¹	cg19017254	-	Blood	cg17583504	-
	Monick, 2012 ⁴³⁴	cg10951975		Lymphoblast	cg19007908	
TXNRD1	Joehanes, 2016 ⁴¹⁵	cg25684105	-	Blood	cg21858255	-
	Joubert, 2016 ⁴¹⁶	cg19722698	+			
WDR1	Joubert, 2016 ⁴¹⁶	cg22821355	-	Blood	cg00598021	+
WWOX	Joehanes, 2016 ⁴¹⁵	cg10001715	+	Blood	cg06420305	-
	Joubert, 2016 ⁴¹⁶	cg08549497				

9.4 Appendix 4 - Chapter 6 Supporting Tables and Figures

Table 9.15 Model results for the FDR significant ($p < 3.8 \times 10^{-5}$) EWAS probes in the three datasets: training, testing and EPIC-NL. All results are from beta regression models assessing the relationship between combined air and dietary PAH8 exposure and the methylation beta values for each probe. The training model adjusted for chip, position on chip, WBC proportions, age, sex, smoking status, cancer case status, and subject centre. The testing model included all covariates with the exception of chip. The EPIC-NL model did not include chip, sex, and cancer case status.

Probe ID	EPIC-Italy – Training (N=493)			EPIC-Italy – Testing (N=208)			EPIC-NL (N=132)		
	β Coefficient	95% Confidence Interval	P Value	β Coefficient	95% Confidence Interval	P Value	β Coefficient	95% Confidence Interval	P Value
cg00030047	-0.023	-0.034; -0.012	2.60E-05	0.031	0.011; 0.05	0.002	-0.012	-0.048; 0.024	0.523
cg00466488	0.099	0.071; 0.126	2.48E-12	0.011	-0.016; 0.037	0.422	0.081	0.022; 0.141	0.008
cg02155655	-0.023	-0.033; -0.014	1.68E-06	0	-0.01; 0.011	0.957	0.009	-0.003; 0.021	0.152
cg02767788	-0.034	-0.049; -0.019	8.54E-06	-0.003	-0.026; 0.02	0.799	0	-0.044; 0.043	0.99
cg03317082	0.062	0.04; 0.084	2.24E-08	-0.001	-0.026; 0.025	0.954	-0.017	-0.073; 0.039	0.553
cg04117764	-0.041	-0.059; -0.022	1.84E-05	0.013	-0.014; 0.04	0.356	-0.007	-0.057; 0.043	0.782
cg04226892	-0.028	-0.039; -0.016	1.37E-06	0.019	-0.01; 0.047	0.203	-0.02	-0.065; 0.025	0.382
cg04662939	-0.017	-0.025; -0.009	2.95E-05	0.013	-0.002; 0.028	0.096	0.005	-0.018; 0.028	0.678
cg04830546	0.018	0.01; 0.026	9.12E-06	-0.017	-0.029; -0.004	0.011	-0.016	-0.047; 0.014	0.29
cg05262877	-0.022	-0.032; -0.013	2.59E-06	0.011	-0.005; 0.027	0.163	0.014	-0.008; 0.036	0.205
cg06182121	0.02	0.012; 0.028	4.26E-07	-0.012	-0.025; 0.001	0.081	0.005	-0.033; 0.043	0.804
cg09009380	-0.023	-0.033; -0.012	2.41E-05	0.017	0.001; 0.032	0.038	-0.019	-0.04; 0.002	0.078

cg09353985	0.022	0.013; 0.031	2.82E-06	-0.015	-0.041; 0.012	0.277	0.017	-0.039; 0.074	0.553
cg11388802	0.018	0.01; 0.027	3.61E-05	0.006	-0.009; 0.02	0.456	0.013	-0.019; 0.045	0.421
cg12653146	-0.038	-0.055; - 0.02	1.95E-05	0.016	0; 0.032	0.047	0.018	-0.026; 0.062	0.42
cg13355424	-0.039	-0.055; - 0.023	1.15E-06	0.028	0.009; 0.047	0.005	-0.006	-0.046; 0.035	0.786
cg16164356	0.04	0.022; 0.059	1.76E-05	-0.007	-0.036; 0.022	0.624	-0.008	-0.065; 0.049	0.774
cg17515347	0.037	0.02; 0.055	3.73E-05	0	-0.024; 0.024	0.989	0.029	-0.045; 0.102	0.44
cg18936620	0.016	0.009; 0.023	1.94E-05	-0.006	-0.016; 0.005	0.267	0.005	-0.026; 0.036	0.731
cg19695266	0.021	0.012; 0.03	8.23E-06	-0.017	-0.033; -0.001	0.041	-0.019	-0.048; 0.011	0.208
cg20962500	-0.098	-0.136; - 0.061	3.34E-07	-0.014	-0.047; 0.019	0.414	-0.022	-0.064; 0.021	0.32
cg21862529	-0.02	-0.029; - 0.011	1.55E-05	0.01	-0.004; 0.024	0.166	0.008	-0.019; 0.036	0.553
cg21908208	-0.057	-0.083; - 0.031	1.48E-05	-0.034	-0.079; 0.011	0.143	NA	NA; NA	NA
cg22025064	-0.022	-0.032; - 0.011	2.72E-05	0.006	-0.01; 0.022	0.463	0.014	-0.015; 0.044	0.348
cg24843511	-0.016	-0.023; - 0.009	1.78E-05	0.007	-0.006; 0.019	0.294	0.005	-0.021; 0.031	0.716
cg24937768	-0.016	-0.024; - 0.008	3.67E-05	0.003	-0.01; 0.015	0.693	0.017	-0.029; 0.063	0.472
cg26272105	-0.019	-0.028; - 0.01	2.87E-05	-0.013	-0.026; 0.001	0.062	-0.011	-0.04; 0.019	0.477
cg26913155	0.021	0.012; 0.031	1.31E-05	-0.018	-0.032; -0.004	0.013	-0.04	-0.073; -0.006	0.019
cg00771084	-0.012	-0.017; - 0.006	2.80E-05	-0.004	-0.014; 0.006	0.434	0	-0.015; 0.016	0.954
cg00901401	-0.049	-0.071; - 0.027	1.35E-05	-0.016	-0.049; 0.016	0.33	0.033	-0.022; 0.087	0.244
cg03025473	0.02	0.011; 0.028	1.59E-05	-0.001	-0.016; 0.014	0.874	0.01	-0.02; 0.04	0.513

cg04042861	-0.031	-0.044; -0.018	3.75E-06	-0.016	-0.039; 0.008	0.187	-0.034	-0.079; 0.012	0.143
cg04850055	0.042	0.023; 0.061	1.09E-05	-0.033	-0.063; -0.004	0.028	0.06	-0.013; 0.134	0.109
cg05511924	0.018	0.01; 0.027	1.88E-05	-0.007	-0.02; 0.006	0.317	-0.032	-0.069; 0.006	0.097
cg05703053	-0.062	-0.088; -0.035	5.38E-06	-0.007	-0.051; 0.038	0.766	0.056	-0.02; 0.131	0.149
cg06856378	-0.087	-0.125; -0.048	1.06E-05	0.056	0.006; 0.106	0.027	NA	NA; NA	NA
cg07480373	0.026	0.015; 0.037	6.51E-06	0.006	-0.013; 0.024	0.553	-0.012	-0.045; 0.021	0.477
cg10459425	0.017	0.009; 0.025	3.60E-05	-0.002	-0.015; 0.01	0.708	-0.009	-0.033; 0.014	0.435
cg12448298	0.049	0.031; 0.068	1.05E-07	-0.008	-0.034; 0.019	0.561	0.009	-0.045; 0.064	0.736
cg14950134	-0.025	-0.037; -0.013	3.21E-05	-0.002	-0.02; 0.016	0.843	0.05	0.001; 0.1	0.046
cg15742848	-0.055	-0.08; -0.03	1.63E-05	0.004	-0.034; 0.043	0.824	0.005	-0.064; 0.074	0.887
cg16723488	-0.046	-0.067; -0.025	2.01E-05	0.03	-0.006; 0.066	0.098	0	-0.042; 0.042	0.988
cg17866732	-0.021	-0.03; -0.012	7.29E-06	0	-0.015; 0.014	0.963	0.011	-0.016; 0.038	0.429
cg19428444	-0.039	-0.057; -0.022	9.15E-06	-0.009	-0.033; 0.016	0.495	-0.064	-0.112; -0.015	0.011
cg19485911	0.052	0.031; 0.073	1.56E-06	0.011	-0.011; 0.034	0.324	0.046	-0.006; 0.099	0.085
cg19697911	-0.023	-0.034; -0.012	3.47E-05	-0.016	-0.034; 0.001	0.07	0.017	-0.022; 0.056	0.386
cg22374586	-0.085	-0.109; -0.06	3.38E-11	0.012	-0.026; 0.05	0.549	0.079	0.006; 0.152	0.035
cg22591002	-0.02	-0.029; -0.011	8.28E-06	0.003	-0.011; 0.018	0.635	0.009	-0.021; 0.038	0.57
cg22939193	-0.036	-0.054; -0.019	2.61E-05	0.011	-0.012; 0.034	0.341	0.055	-0.023; 0.132	0.166

cg23665824	-0.053	-0.077; - 0.03	9.65E-06	0.042	-0.001; 0.085	0.055	0.015	-0.049; 0.079	0.644
cg23759710	-0.013	-0.018; - 0.007	4.10E-06	-0.006	-0.015; 0.003	0.213	0.007	-0.009; 0.022	0.406
cg24935556	0.023	0.014; 0.033	2.21E-06	-0.02	-0.036; -0.003	0.021	-0.033	-0.075; 0.008	0.113
cg00121562	-0.021	-0.031; - 0.011	3.03E-05	-0.003	-0.022; 0.016	0.738	-0.009	-0.031; 0.014	0.448
cg11162839	0.021	0.011; 0.031	2.19E-05	0.011	-0.002; 0.025	0.104	0.025	-0.02; 0.07	0.284
cg12361223	0.052	0.029; 0.075	1.09E-05	-0.038	-0.074; -0.003	0.036	0.106	0.037; 0.176	0.003
cg12389423	-0.039	-0.054; - 0.024	1.39E-07	0.01	-0.009; 0.029	0.284	-0.042	-0.093; 0.008	0.098
cg14875327	-0.042	-0.062; - 0.022	3.03E-05	-0.012	-0.048; 0.025	0.523	0.043	-0.019; 0.105	0.177
cg18334977	0.02	0.011; 0.028	1.69E-05	-0.011	-0.027; 0.004	0.153	0.014	-0.032; 0.061	0.55
cg18592273	0.117	0.072; 0.161	3.18E-07	0.094	0.03; 0.157	0.004	NA	NA; NA	NA
cg19516404	-0.017	-0.025; - 0.009	3.18E-05	-0.001	-0.012; 0.011	0.919	-0.022	-0.05; 0.005	0.116
cg20912272	-0.018	-0.026; - 0.009	3.49E-05	-0.004	-0.019; 0.01	0.57	0.029	0.002; 0.056	0.034
cg21548131	0.063	0.036; 0.09	3.89E-06	0.032	-0.015; 0.079	0.18	NA	NA; NA	NA
cg24620761	0.032	0.018; 0.046	1.29E-05	0.003	-0.016; 0.022	0.767	0.004	-0.041; 0.049	0.862
cg25315362	-0.028	-0.041; - 0.015	3.32E-05	-0.016	-0.038; 0.006	0.145	-0.039	-0.078; 0	0.052
cg25679475	-0.033	-0.049; - 0.018	2.95E-05	-0.001	-0.027; 0.024	0.915	-0.015	-0.048; 0.018	0.372
cg06466757	0.053	0.032; 0.073	4.25E-07	-0.027	-0.06; 0.005	0.1	0.009	-0.036; 0.055	0.686
cg09084892	0.026	0.014; 0.038	2.52E-05	-0.009	-0.024; 0.006	0.253	-0.02	-0.068; 0.029	0.421

cg11060856	-0.022	-0.032; - 0.012	1.58E-05	0.004	-0.011; 0.02	0.592	-0.002	-0.027; 0.023	0.861
cg13842222	0.034	0.018; 0.05	2.81E-05	-0.005	-0.03; 0.019	0.67	-0.029	-0.105; 0.047	0.455
cg15407965	-0.043	-0.063; - 0.024	1.07E-05	0.015	-0.013; 0.043	0.285	-0.017	-0.076; 0.041	0.568
cg17304168	0.029	0.019; 0.04	7.31E-08	0.007	-0.009; 0.024	0.389	-0.003	-0.054; 0.049	0.92
cg17910931	-0.019	-0.028; - 0.011	1.10E-05	0.007	-0.006; 0.02	0.296	-0.003	-0.029; 0.024	0.847
cg20536207	-0.018	-0.025; - 0.01	7.19E-06	0.009	-0.004; 0.021	0.196	-0.006	-0.027; 0.014	0.538
cg25144207	-0.034	-0.05; - 0.019	9.34E-06	-0.012	-0.034; 0.009	0.262	0.002	-0.032; 0.037	0.892
cg00489401	-0.027	-0.037; - 0.016	1.06E-06	0.009	-0.006; 0.024	0.252	0.036	-0.005; 0.078	0.086
cg00618323	-0.016	-0.023; - 0.008	2.88E-05	-0.007	-0.018; 0.005	0.26	-0.007	-0.032; 0.018	0.582
cg05184550	-0.041	-0.056; - 0.025	1.72E-07	0.031	0.005; 0.057	0.018	-0.001	-0.056; 0.053	0.968
cg07287793	-0.018	-0.026; - 0.01	7.01E-06	0.005	-0.007; 0.018	0.412	-0.003	-0.02; 0.014	0.736
cg09458384	-0.032	-0.044; - 0.019	1.04E-06	0.006	-0.013; 0.025	0.53	0.016	-0.027; 0.058	0.476
cg11081752	0.025	0.013; 0.037	3.36E-05	0.006	-0.013; 0.026	0.526	-0.011	-0.06; 0.039	0.669
cg11190434	0.023	0.012; 0.034	2.70E-05	-0.018	-0.034; -0.002	0.032	0.03	-0.015; 0.074	0.191
cg19423543	-0.042	-0.062; - 0.022	3.03E-05	0.017	-0.015; 0.05	0.289	0.029	-0.013; 0.071	0.178
cg25110832	0.038	0.022; 0.054	4.25E-06	-0.023	-0.047; 0.002	0.069	-0.028	-0.08; 0.024	0.289
cg26496372	0.031	0.019; 0.044	5.69E-07	-0.008	-0.02; 0.004	0.17	0.004	-0.023; 0.031	0.769
cg00686197	-0.017	-0.025; - 0.009	3.35E-05	-0.015	-0.029; 0	0.049	0.015	-0.016; 0.045	0.349

cg01359933	0.036	0.021; 0.051	4.18E-06	0	-0.021; 0.022	0.965	-0.021	-0.093; 0.051	0.572
cg02595760	-0.029	-0.041; -0.016	6.83E-06	0.003	-0.017; 0.024	0.763	0.008	-0.04; 0.057	0.731
cg03349397	-0.026	-0.036; -0.017	1.12E-07	0.021	0.005; 0.037	0.009	0.006	-0.03; 0.041	0.758
cg05629964	0.04	0.023; 0.057	7.13E-06	-0.027	-0.054; -0.001	0.042	0.012	-0.065; 0.09	0.756
cg05927817	-0.089	-0.126; -0.053	1.57E-06	-0.024	-0.077; 0.029	0.375	-0.01	-0.091; 0.071	0.809
cg08014661	0.026	0.015; 0.036	3.53E-06	-0.003	-0.017; 0.012	0.733	-0.007	-0.047; 0.032	0.709
cg08097157	0.023	0.012; 0.033	3.29E-05	-0.016	-0.029; -0.004	0.012	-0.009	-0.045; 0.026	0.596
cg12256206	-0.022	-0.032; -0.012	2.47E-05	-0.016	-0.033; 0	0.053	0.006	-0.025; 0.038	0.691
cg13312976	0.057	0.031; 0.082	1.08E-05	0.025	-0.024; 0.075	0.314	-0.02	-0.102; 0.062	0.631
cg15289190	0.038	0.021; 0.056	2.17E-05	-0.004	-0.023; 0.015	0.671	0.024	-0.014; 0.061	0.218
cg17427198	-0.029	-0.043; -0.016	2.75E-05	-0.037	-0.063; -0.011	0.005	0.054	0; 0.107	0.049
cg21286967	-0.02	-0.029; -0.011	1.03E-05	0.002	-0.011; 0.016	0.72	0.027	-0.001; 0.055	0.058
cg23973371	-0.014	-0.021; -0.008	1.08E-05	0.014	0.003; 0.024	0.013	-0.014	-0.029; 0.001	0.065
cg24225668	-0.02	-0.03; -0.011	3.20E-05	-0.002	-0.017; 0.012	0.745	-0.024	-0.053; 0.005	0.102
cg25748868	0.029	0.017; 0.041	2.08E-06	0.009	-0.008; 0.026	0.316	-0.023	-0.061; 0.016	0.255
cg26590603	-0.035	-0.052; -0.019	2.38E-05	0.011	-0.017; 0.04	0.438	-0.001	-0.042; 0.04	0.966
cg00870778	0.038	0.02; 0.055	2.06E-05	-0.007	-0.031; 0.017	0.575	-0.031	-0.095; 0.032	0.331
cg00984540	-0.053	-0.078; -0.028	2.36E-05	0.042	0.001; 0.082	0.043	-0.037	-0.095; 0.021	0.214

cg01353448	-0.042	-0.062; -0.023	2.27E-05	0.009	-0.028; 0.046	0.628	0.02	-0.031; 0.071	0.444
cg01731811	-0.034	-0.05; -0.019	8.33E-06	-0.004	-0.029; 0.021	0.78	0.041	-0.004; 0.085	0.073
cg01740202	-0.013	-0.018; -0.007	3.25E-06	0.002	-0.006; 0.009	0.648	0.013	-0.001; 0.028	0.072
cg02883229	0.028	0.015; 0.041	2.29E-05	-0.005	-0.024; 0.014	0.611	-0.024	-0.056; 0.007	0.125
cg03308706	-0.057	-0.084; -0.03	3.68E-05	-0.026	-0.068; 0.016	0.226	-0.06	-0.126; 0.006	0.076
cg04493169	-0.048	-0.068; -0.028	2.32E-06	-0.024	-0.054; 0.007	0.124	0.022	-0.038; 0.082	0.473
cg04678743	0.107	0.056; 0.157	3.18E-05	0.337	0.244; 0.431	0	NA	NA; NA	NA
cg06745145	0.033	0.02; 0.046	7.11E-07	-0.025	-0.047; -0.004	0.021	0.017	-0.025; 0.06	0.426
cg09214099	-0.026	-0.036; -0.016	1.01E-07	0.005	-0.011; 0.022	0.534	0.009	-0.022; 0.04	0.552
cg09424595	-0.019	-0.028; -0.011	5.25E-06	0.007	-0.007; 0.021	0.325	-0.064	-0.093; -0.036	0
cg09505513	0.02	0.011; 0.03	2.08E-05	0.007	-0.007; 0.021	0.324	0.021	-0.011; 0.052	0.2
cg12844895	0.017	0.009; 0.025	2.50E-05	0.003	-0.009; 0.015	0.633	-0.018	-0.052; 0.017	0.315
cg15035350	-0.07	-0.102; -0.039	1.30E-05	0.061	0.008; 0.114	0.023	NA	NA; NA	NA
cg17604655	-0.051	-0.074; -0.029	8.90E-06	-0.037	-0.072; -0.001	0.041	0.018	-0.044; 0.08	0.569
cg26914299	-0.022	-0.032; -0.013	4.51E-06	0.012	-0.003; 0.027	0.105	0.005	-0.022; 0.033	0.707
cg06009497	-0.022	-0.031; -0.014	8.80E-08	0.006	-0.004; 0.016	0.228	-0.016	-0.033; 0.001	0.059
cg08772302	-0.017	-0.024; -0.009	1.36E-05	-0.002	-0.012; 0.009	0.758	0.002	-0.029; 0.033	0.913
cg09961689	-0.04	-0.059; -0.022	1.17E-05	-0.014	-0.032; 0.004	0.118	0.009	-0.024; 0.042	0.599

cg10140583	-0.041	-0.059; -0.023	6.27E-06	-0.015	-0.057; 0.027	0.481	0.012	-0.017; 0.041	0.429
cg12126859	0.033	0.018; 0.049	2.72E-05	-0.017	-0.044; 0.01	0.206	-0.017	-0.062; 0.029	0.471
cg15706250	-0.026	-0.037; -0.015	2.35E-06	0.004	-0.011; 0.019	0.59	-0.037	-0.07; -0.003	0.031
cg17683573	-0.05	-0.072; -0.029	3.03E-06	0.039	0.02; 0.058	0	0.005	-0.038; 0.049	0.806
cg20585869	-0.06	-0.088; -0.032	2.81E-05	0.008	-0.036; 0.052	0.724	NA	NA; NA	NA
cg21045828	0.04	0.023; 0.057	3.25E-06	0.028	0.007; 0.05	0.01	-0.041	-0.115; 0.033	0.278
cg22848598	-0.077	-0.108; -0.046	1.02E-06	0.031	-0.011; 0.073	0.147	0.034	-0.053; 0.121	0.438
cg22932649	0.06	0.031; 0.088	3.74E-05	-0.031	-0.067; 0.004	0.085	-0.065	-0.139; 0.008	0.083
cg24495007	-0.015	-0.022; -0.008	2.11E-05	0.003	-0.009; 0.015	0.651	0	-0.016; 0.016	0.989
cg13845049	0.024	0.013; 0.035	1.61E-05	-0.009	-0.025; 0.007	0.29	0.017	-0.031; 0.065	0.481
cg14027524	-0.019	-0.027; -0.01	1.40E-05	-0.011	-0.026; 0.003	0.129	0.016	-0.009; 0.041	0.199
cg14286514	-0.052	-0.071; -0.033	6.83E-08	-0.03	-0.059; -0.001	0.041	0.002	-0.052; 0.057	0.934
cg14410137	-0.012	-0.017; -0.006	2.07E-05	0	-0.008; 0.008	0.958	0.004	-0.014; 0.021	0.681
cg03246584	0.11	0.066; 0.154	9.78E-07	0.021	-0.031; 0.073	0.432	NA	NA; NA	NA
cg14494090	-0.02	-0.029; -0.01	3.17E-05	-0.02	-0.035; -0.005	0.007	-0.001	-0.028; 0.027	0.962
cg14677612	0.044	0.024; 0.064	1.37E-05	0.02	-0.015; 0.056	0.266	-0.16	-0.243; -0.077	0
cg14677909	-0.074	-0.102; -0.047	8.82E-08	0.024	-0.047; 0.096	0.505	-0.017	-0.082; 0.048	0.603
cg15275103	-0.114	-0.148; -0.079	1.21E-10	0.057	0.014; 0.1	0.01	-0.012	-0.061; 0.037	0.633

cg16692757	-0.022	-0.033; - 0.012	2.42E-05	0.003	-0.014; 0.019	0.754	-0.011	-0.047; 0.025	0.552
cg18308755	-0.049	-0.066; - 0.032	1.39E-08	0.017	-0.006; 0.04	0.146	0.037	-0.02; 0.094	0.2
cg18979589	-0.041	-0.058; - 0.025	1.62E-06	0.053	0.022; 0.083	0.001	0.004	-0.045; 0.052	0.881
cg19295034	0.05	0.028; - 0.073	1.27E-05	0.014	-0.022; 0.049	0.449	NA	NA; NA	NA
cg27190138	0.027	0.015; 0.04	2.81E-05	0.001	-0.018; 0.019	0.928	-0.001	-0.041; 0.039	0.963
cg01981334	-0.051	-0.069; - 0.033	3.44E-08	0.011	-0.001; 0.024	0.081	0.032	0.005; 0.06	0.023
cg02508204	0.037	0.02; 0.054	2.11E-05	-0.039	-0.07; -0.008	0.013	-0.055	-0.104; -0.006	0.028
cg03728580	-0.029	-0.042; - 0.016	2.00E-05	-0.011	-0.037; 0.015	0.409	0.009	-0.018; 0.036	0.511
cg04174538	-0.019	-0.028; - 0.01	2.70E-05	-0.009	-0.025; 0.007	0.26	0.008	-0.017; 0.033	0.531
cg04293602	-0.028	-0.039; - 0.016	1.99E-06	0.004	-0.015; 0.024	0.667	0.004	-0.026; 0.035	0.786
cg06193239	-0.033	-0.048; - 0.018	1.78E-05	-0.027	-0.05; -0.004	0.02	-0.02	-0.062; 0.021	0.338
cg10337956	0.029	0.016; - 0.042	1.46E-05	-0.018	-0.04; 0.004	0.106	0.019	-0.024; 0.063	0.378
cg10854423	0.023	0.012; - 0.033	2.46E-05	-0.019	-0.037; -0.001	0.04	0.002	-0.043; 0.047	0.921
cg12472022	-0.028	-0.04; - 0.016	7.82E-06	0.005	-0.015; 0.025	0.629	0.009	-0.047; 0.066	0.744
cg12575434	0.038	0.021; - 0.056	1.81E-05	-0.025	-0.052; 0.002	0.067	0.052	-0.011; 0.115	0.106
cg12690575	-0.017	-0.025; - 0.009	2.55E-05	0.003	-0.009; 0.015	0.64	0.001	-0.029; 0.031	0.947
cg12781794	-0.041	-0.059; - 0.022	1.27E-05	0.026	-0.002; 0.054	0.07	-0.017	-0.056; 0.022	0.398
cg12782933	-0.026	-0.038; - 0.015	1.44E-05	0.009	-0.01; 0.028	0.356	0.023	-0.021; 0.068	0.304

cg15233880	-0.072	-0.105; - 0.04	1.26E-05	0.064	0.015; 0.112	0.01	NA	NA; NA	NA
cg15659420	0.029	0.015; 0.043	3.44E-05	-0.001	-0.021; 0.018	0.89	-0.008	-0.064; 0.048	0.77
cg19204693	-0.029	-0.041; - 0.016	4.54E-06	0.007	-0.006; 0.02	0.287	-0.014	-0.043; 0.015	0.337
cg19375210	0.02	0.011; 0.03	3.36E-05	-0.013	-0.028; 0.001	0.076	-0.001	-0.031; 0.03	0.973
cg20051949	0.017	0.009; 0.025	2.99E-05	-0.009	-0.022; 0.004	0.179	-0.001	-0.039; 0.038	0.975
cg20122043	0.025	0.015; 0.035	1.29E-06	-0.022	-0.037; -0.007	0.005	-0.012	-0.057; 0.033	0.599
cg22049858	0.068	0.041; 0.095	8.72E-07	-0.044	-0.073; -0.015	0.003	-0.062	-0.116; -0.009	0.023
cg22458194	-0.022	-0.032; - 0.012	8.10E-06	0.003	-0.015; 0.021	0.75	-0.025	-0.055; 0.005	0.102
cg22747802	-0.023	-0.034; - 0.013	1.30E-05	-0.022	-0.037; -0.006	0.006	-0.011	-0.033; 0.012	0.346
cg23313445	0.02	0.011; 0.03	2.91E-05	0	-0.014; 0.014	0.988	0.005	-0.029; 0.039	0.772
cg24413662	-0.033	-0.047; - 0.019	7.04E-06	0.037	0.016; 0.059	0.001	0.041	-0.01; 0.092	0.112
cg02574894	0.057	0.034; 0.079	6.04E-07	-0.012	-0.035; 0.011	0.319	0.021	-0.031; 0.073	0.421
cg05419385	0.018	0.01; 0.027	2.05E-05	0	-0.012; 0.012	0.963	0.009	-0.023; 0.04	0.593
cg09576209	-0.032	-0.047; - 0.017	3.59E-05	0.01	-0.013; 0.032	0.413	-0.008	-0.041; 0.025	0.644
cg14990076	0.033	0.018; 0.048	2.39E-05	0.019	-0.006; 0.043	0.143	-0.079	-0.132; -0.026	0.003
cg15545035	-0.026	-0.036; - 0.015	1.15E-06	0.001	-0.016; 0.018	0.869	-0.005	-0.038; 0.029	0.785
cg16708880	-0.042	-0.062; - 0.022	3.63E-05	0.024	-0.009; 0.057	0.159	0.02	-0.03; 0.069	0.436
cg17844553	0.07	0.04; 0.101	6.05E-06	0.021	-0.021; 0.064	0.324	NA	NA; NA	NA
cg18827332	-0.032	-0.045; - 0.018	2.26E-06	-0.015	-0.034; 0.004	0.119	0.004	-0.029; 0.038	0.797

cg23432930	-0.016	-0.024; -0.009	2.05E-05	0.005	-0.007; 0.017	0.391	-0.02	-0.043; 0.004	0.099
cg23897083	0.024	0.013; 0.034	1.77E-05	-0.008	-0.027; 0.01	0.382	0.008	-0.028; 0.043	0.673
cg24445165	-0.031	-0.045; -0.016	3.12E-05	-0.019	-0.041; 0.002	0.074	0.003	-0.027; 0.032	0.853
cg10512779	0.022	0.012; 0.032	2.53E-05	-0.025	-0.041; -0.009	0.002	0.031	-0.007; 0.069	0.107
cg11315081	-0.072	-0.103; -0.041	4.32E-06	-0.045	-0.089; -0.001	0.047	NA	NA; NA	NA
cg23137039	-0.02	-0.03; -0.011	3.15E-05	-0.02	-0.037; -0.004	0.013	-0.01	-0.037; 0.017	0.462
cg23440004	-0.018	-0.026; -0.009	3.33E-05	0.003	-0.011; 0.017	0.697	0.003	-0.026; 0.033	0.83
cg01636662	0.026	0.014; 0.037	1.53E-05	0.021	0.004; 0.038	0.017	0.016	-0.023; 0.056	0.423
cg02583546	0.067	0.038; 0.097	6.90E-06	-0.043	-0.071; -0.015	0.003	0.047	0.001; 0.094	0.046
cg05881436	-0.031	-0.044; -0.018	4.21E-06	0.016	-0.005; 0.036	0.138	0.022	-0.012; 0.055	0.204
cg18225991	0.077	0.048; 0.106	1.63E-07	0.007	-0.04; 0.054	0.76	NA	NA; NA	NA
cg27158340	-0.024	-0.033; -0.015	4.57E-08	0	-0.014; 0.015	0.97	0.014	-0.013; 0.042	0.301
cg01908020	0.022	0.012; 0.032	2.62E-05	-0.016	-0.028; -0.003	0.014	0.003	-0.036; 0.041	0.887
cg05239311	-0.05	-0.074; -0.027	3.41E-05	0.022	-0.019; 0.063	0.291	-0.035	-0.108; 0.038	0.348
cg08815652	-0.02	-0.03; -0.011	1.87E-05	0.002	-0.013; 0.016	0.837	0.027	0.007; 0.047	0.008
cg14209037	-0.048	-0.067; -0.029	7.39E-07	-0.008	-0.035; 0.018	0.535	0.013	-0.042; 0.067	0.644
cg22475358	-0.035	-0.051; -0.02	8.78E-06	-0.005	-0.032; 0.022	0.725	0.009	-0.028; 0.046	0.643
cg23625341	0.025	0.015; 0.036	3.37E-06	0.004	-0.013; 0.021	0.63	-0.048	-0.093; -0.003	0.035

cg23735339	-0.037	-0.055; -0.02	3.38E-05	-0.036	-0.063; -0.01	0.007	0.075	0.01; 0.14	0.025
cg26481829	-0.038	-0.056; -0.02	3.60E-05	0.017	-0.009; 0.044	0.204	0.024	-0.037; 0.084	0.448
cg00876678	-0.021	-0.03; -0.012	8.68E-06	-0.003	-0.016; 0.01	0.651	-0.004	-0.025; 0.017	0.707
cg01003448	0.07	0.038; 0.101	1.28E-05	-0.009	-0.055; 0.038	0.715	0.01	-0.066; 0.085	0.8
cg02412803	-0.018	-0.026; -0.01	9.62E-06	-0.012	-0.024; 0	0.059	0.004	-0.024; 0.032	0.784
cg05131483	-0.03	-0.042; -0.017	4.16E-06	0.001	-0.019; 0.021	0.947	0.008	-0.031; 0.048	0.679
cg05406475	-0.021	-0.028; -0.013	1.82E-07	0.007	-0.006; 0.019	0.287	0.008	-0.01; 0.026	0.39
cg07482202	0.121	0.073; 0.168	7.36E-07	0.023	-0.049; 0.094	0.532	-0.002	-0.13; 0.127	0.98
cg08550881	-0.02	-0.028; -0.011	1.26E-05	-0.005	-0.017; 0.008	0.454	-0.046	-0.074; -0.018	0.001
cg09074450	-0.036	-0.051; -0.021	4.30E-06	0.004	-0.019; 0.027	0.757	0.013	-0.016; 0.042	0.38
cg09942293	-0.017	-0.025; -0.009	9.52E-06	0.013	0; 0.027	0.056	0.005	-0.013; 0.022	0.582
cg16619935	-0.02	-0.028; -0.012	1.18E-06	0.012	-0.003; 0.027	0.121	-0.022	-0.052; 0.007	0.142
cg24303478	-0.015	-0.022; -0.008	2.85E-05	0.01	0; 0.019	0.056	-0.017	-0.038; 0.004	0.113
cg26586719	0.028	0.015; 0.041	2.66E-05	-0.01	-0.027; 0.008	0.291	-0.006	-0.046; 0.033	0.756
cg26780022	-0.041	-0.057; -0.024	1.83E-06	0.006	-0.02; 0.031	0.66	-0.013	-0.051; 0.025	0.508
cg26916410	0.025	0.014; 0.037	1.63E-05	-0.001	-0.017; 0.015	0.892	-0.028	-0.083; 0.027	0.313
cg26929163	0.04	0.023; 0.057	2.32E-06	0.004	-0.027; 0.034	0.819	0.007	-0.036; 0.05	0.756
cg27605307	-0.042	-0.062; -0.022	3.24E-05	-0.005	-0.032; 0.022	0.738	0.03	-0.028; 0.088	0.311

cg00830755	-0.024	-0.035; - 0.013	2.36E-05	-0.001	-0.016; 0.014	0.907	-0.001	-0.037; 0.036	0.974
cg02216727	-0.035	-0.05; - 0.02	3.96E-06	0.005	-0.011; 0.022	0.509	-0.008	-0.045; 0.029	0.681
cg02958960	0.025	0.014; 0.036	4.29E-06	0	-0.014; 0.015	0.965	-0.018	-0.071; 0.035	0.508
cg10060065	-0.029	-0.043; - 0.016	2.15E-05	0.005	-0.015; 0.024	0.63	-0.016	-0.05; 0.019	0.371
cg11653266	0.031	0.017; 0.046	1.56E-05	-0.027	-0.05; -0.003	0.027	-0.016	-0.083; 0.052	0.647
cg12187586	-0.047	-0.067; - 0.027	3.63E-06	0.02	-0.015; 0.054	0.259	0.023	-0.017; 0.064	0.262
cg12550399	0.02	0.012; 0.028	1.36E-06	NA	NA; NA	NA	0	-0.028; 0.027	0.985
cg12964144	-0.022	-0.031; - 0.013	3.30E-06	0.014	-0.003; 0.031	0.101	0.028	-0.002; 0.059	0.072
cg15209885	0.011	0.006; 0.017	3.08E-05	0	-0.009; 0.008	0.909	0.005	-0.017; 0.027	0.643
cg16548154	-0.014	-0.021; - 0.008	6.30E-06	-0.001	-0.01; 0.009	0.909	-0.004	-0.021; 0.013	0.644
cg17325958	0.02	0.011; 0.03	3.61E-05	-0.009	-0.025; 0.006	0.245	0.018	-0.028; 0.063	0.444
cg17996892	-0.014	-0.021; - 0.008	2.72E-05	0.006	-0.004; 0.017	0.242	0.009	-0.009; 0.027	0.314
cg18576374	0.044	0.028; 0.059	2.79E-08	-0.056	-0.083; -0.029	0	0.027	-0.019; 0.073	0.255
cg21507719	-0.02	-0.029; - 0.011	2.02E-05	0.011	-0.002; 0.024	0.095	-0.014	-0.041; 0.014	0.324
cg25170034	-0.038	-0.056; - 0.02	3.34E-05	0.015	-0.011; 0.042	0.265	-0.008	-0.052; 0.037	0.733
cg25930644	0.021	0.013; 0.03	1.10E-06	-0.014	-0.028; -0.001	0.031	0.011	-0.023; 0.045	0.52
cg03012785	-0.093	-0.137; - 0.049	3.75E-05	-0.043	-0.117; 0.03	0.249	0.117	0.005; 0.229	0.04
cg14307471	0.043	0.024; 0.062	1.02E-05	0.031	0.001; 0.061	0.041	NA	NA; NA	NA

cg15704408	-0.02	-0.029; -0.011	1.37E-05	0.001	-0.013; 0.014	0.926	-0.028	-0.053; -0.003	0.029
cg00086493	-0.023	-0.033; -0.013	5.82E-06	-0.011	-0.025; 0.004	0.146	0.006	-0.023; 0.035	0.701
cg00910067	-0.028	-0.041; -0.015	2.42E-05	0.003	-0.014; 0.02	0.745	0.025	-0.016; 0.065	0.23
cg02281038	-0.039	-0.057; -0.021	3.57E-05	-0.019	-0.051; 0.012	0.231	0.044	-0.012; 0.099	0.122
cg03781262	-0.035	-0.051; -0.02	1.14E-05	0.015	-0.006; 0.036	0.153	-0.066	-0.119; -0.012	0.016
cg04351156	-0.016	-0.024; -0.009	2.67E-05	-0.004	-0.016; 0.008	0.481	-0.009	-0.023; 0.005	0.229
cg07356415	-0.021	-0.031; -0.012	1.14E-05	0.017	0.003; 0.031	0.02	-0.002	-0.029; 0.024	0.856
cg12088773	-0.061	-0.088; -0.034	1.09E-05	-0.006	-0.051; 0.038	0.779	0.028	-0.046; 0.103	0.451
cg12497870	-0.023	-0.032; -0.014	2.52E-07	0.002	-0.011; 0.015	0.792	0.002	-0.029; 0.032	0.914
cg12610917	-0.039	-0.055; -0.023	2.67E-06	0.014	-0.011; 0.039	0.27	-0.02	-0.063; 0.023	0.361
cg16574191	-0.035	-0.05; -0.02	3.24E-06	0.004	-0.021; 0.03	0.735	-0.04	-0.084; 0.004	0.074
cg17583504	-0.031	-0.046; -0.017	1.60E-05	0.028	0.001; 0.055	0.042	-0.015	-0.056; 0.025	0.46
cg19299952	-0.13	-0.187; -0.074	6.38E-06	0.151	0.043; 0.258	0.006	-0.126	-0.258; 0.006	0.062
cg19985870	-0.022	-0.032; -0.012	1.25E-05	-0.011	-0.026; 0.005	0.175	0.009	-0.017; 0.035	0.507
cg25378939	-0.023	-0.033; -0.013	8.65E-06	-0.005	-0.021; 0.012	0.559	0.006	-0.028; 0.039	0.739
cg00199007	0.035	0.019; 0.051	2.30E-05	-0.031	-0.059; -0.003	0.028	-0.008	-0.06; 0.043	0.75
cg06457011	0.043	0.023; 0.064	2.86E-05	-0.004	-0.033; 0.024	0.771	-0.013	-0.099; 0.072	0.76
cg07436991	-0.046	-0.067; -0.025	1.92E-05	-0.015	-0.043; 0.013	0.302	0.009	-0.064; 0.082	0.816

cg09576415	-0.016	-0.023; -0.008	3.53E-05	0.006	-0.008; 0.02	0.374	-0.006	-0.032; 0.021	0.67
cg10187707	-0.036	-0.052; -0.02	7.42E-06	0.017	-0.007; 0.04	0.157	0.021	-0.016; 0.059	0.267
cg10629004	-0.048	-0.071; -0.026	2.75E-05	-0.025	-0.063; 0.013	0.19	0.071	0.014; 0.129	0.014
cg11004890	-0.036	-0.05; -0.021	2.18E-06	-0.007	-0.027; 0.014	0.518	0.007	-0.028; 0.042	0.696
cg12978433	-0.044	-0.065; -0.023	3.05E-05	0.053	0.023; 0.083	0.001	-0.01	-0.068; 0.049	0.742
cg14083397	0.084	0.055; 0.112	1.32E-08	-0.011	-0.029; 0.006	0.195	0.017	-0.033; 0.067	0.507
cg17181362	-0.031	-0.044; -0.017	7.27E-06	0.006	-0.015; 0.027	0.581	0.032	-0.003; 0.067	0.071
cg24147187	-0.046	-0.066; -0.026	8.10E-06	-0.01	-0.046; 0.025	0.575	0.022	-0.031; 0.075	0.418
cg24849555	0.057	0.031; 0.083	1.71E-05	0.025	-0.02; 0.069	0.273	NA	NA; NA	NA
cg25188239	0.025	0.013; 0.037	2.61E-05	0.004	-0.018; 0.026	0.71	0.036	0.013; 0.06	0.003
cg27207308	-0.029	-0.041; -0.016	7.84E-06	0.02	-0.001; 0.04	0.059	0.015	-0.029; 0.058	0.506
cg08073527	0.031	0.018; 0.043	1.85E-06	-0.023	-0.051; 0.004	0.095	-0.005	-0.03; 0.019	0.68
cg09443102	-0.022	-0.032; -0.012	2.80E-05	-0.006	-0.022; 0.009	0.43	0.013	-0.019; 0.044	0.424
cg12826791	-0.078	-0.109; -0.046	1.07E-06	-0.003	-0.032; 0.026	0.849	0.023	-0.028; 0.073	0.375
cg19312314	0.125	0.066; 0.183	2.88E-05	-0.003	-0.098; 0.092	0.944	0.023	-0.082; 0.128	0.671
cg24936695	0.022	0.012; 0.032	1.85E-05	-0.002	-0.017; 0.012	0.774	-0.024	-0.059; 0.011	0.173
cg00077838	-0.022	-0.031; -0.013	1.87E-06	-0.002	-0.018; 0.013	0.758	-0.013	-0.05; 0.025	0.5
cg00256932	-0.054	-0.079; -0.03	1.54E-05	0.027	-0.011; 0.065	0.17	-0.007	-0.07; 0.057	0.837

cg04195684	0.016	0.009; 0.024	3.74E-05	-0.007	-0.019; 0.004	0.198	-0.009	-0.04; 0.021	0.551
cg08215954	-0.019	-0.027; - 0.011	7.31E-06	0.004	-0.01; 0.019	0.532	0.008	-0.022; 0.038	0.592
cg10047755	-0.032	-0.046; - 0.017	2.82E-05	0.019	-0.006; 0.044	0.13	-0.003	-0.046; 0.041	0.904
cg11985360	-0.041	-0.06; - 0.022	2.00E-05	-0.031	-0.065; 0.002	0.068	0.102	0.043; 0.161	0.001
cg17145402	-0.023	-0.033; - 0.014	9.38E-07	0.009	-0.005; 0.022	0.218	0.012	-0.031; 0.055	0.597
cg17998530	0.02	0.011; - 0.029	3.50E-05	0	-0.012; 0.012	0.997	0.008	-0.022; 0.039	0.598
cg23018242	0.046	0.026; - 0.065	6.88E-06	0.016	-0.018; 0.051	0.347	0.034	-0.008; 0.077	0.108
cg27665648	-0.026	-0.038; - 0.014	1.87E-05	0.002	-0.016; 0.02	0.834	-0.002	-0.035; 0.031	0.914

Table 9.16. Table of characteristics of probes found to be significantly associated with combined air and dietary PAH8 exposure at the FDR level ($p < 3.8 \times 10^{-5}$) in the training dataset.

Probe ID	Chromosome	Position	UCSC RefGene Name	Location in Gene	Relation to CpG Island	Direction of Methylation Change
cg19695266	1	1241672	<i>ACAP3</i>	Body	North Shore	+
cg24937768	1	2092853	<i>PRKCZ</i>	Body		-
cg06182121	1	3080723	<i>PRDM16</i>	Body	North Shore	+
cg02767788	1	3102750	<i>PRDM16</i>	Body	Island	-
cg26913155	1	3128175	<i>PRDM16</i>	Body		+
cg00030047	1	6268790	<i>RNF207</i>	Body	North Shore	-
cg04117764	1	10917451				-
cg21908208	1	17865737	<i>ARHGEF10L</i>	TSS1500	North Shore	-
cg22025064	1	22141400	<i>LDLRAD2</i>	Body	Island	-
cg12653146	1	25919290				-
cg21862529	1	41948212	<i>EDN2</i>	Body		-
cg05262877	1	42631835	<i>GUCA2A</i>	TSS1500		-
cg18936620	1	43811019	<i>MPL</i>	Body	North Shelf	+
cg26272105	1	47644977				-
cg02155655	1	53566481	<i>SLC1A7</i>	Body		-
cg04830546	1	95969622				+
cg16164356	1	111683801	<i>DRAM2</i>	TSS1500	SouthShore	+
cg00466488	1	118148927	<i>FAM46C</i>	5'UTR	Island	+

cg20962500	1	149174971				-
cg24843511	1	153579799	<i>S100A16</i>	3'UTR		-
cg13355424	1	157165134			Island	-
cg17515347	1	159047163	<i>AIM2</i>	TSS1500		+
cg11388802	1	181577847	<i>CACNA1E</i>	Body		+
cg09009380	1	201252974	<i>PKP1</i>	1stExon	Island	-
cg04226892	1	201637173	<i>NAV1</i>	Body		-
cg04662939	1	204380572	<i>PPP1R15B</i>	5'UTR	Island	-
cg09353985	1	207062674				+
cg03317082	1	234748618			SouthShore	+
cg14950134	2	3261155	<i>TSSC1</i>	Body		-
cg22939193	2	20190182	<i>WDR35</i>	TSS1500	Island	-
cg19428444	2	21023690	<i>C2orf43</i>	TSS1500	SouthShore	-
cg16723488	2	21266947	<i>APOB</i>	TSS200	Island	-
cg24935556	2	21291088				+
cg00901401	2	24272044	<i>FKBP1B</i>	TSS1500	North Shore	-
cg23759710	2	42990957	<i>OXER1</i>	1stExon		-
cg23665824	2	80530701	<i>CTNNA2</i>	Body	Island	-
cg22591002	2	97530695	<i>SEMA4C</i>	Body	Island	-
cg17866732	2	110372875	<i>Sep-10</i>	TSS1500	Island	-
cg12448298	2	115822039	<i>DPP10</i>	Body		+
cg06856378	2	160759118	<i>LY75</i>	Body	North Shore	-

cg15742848	2	169769501				-
cg05703053	2	169769616				-
cg04850055	2	170682941	<i>UBR3</i>	TSS1500	North Shore	+
cg03025473	2	205415509	<i>PARD3B</i>	Body		+
cg07480373	2	216874286	<i>MREG</i>	Body	North Shelf	+
cg05511924	2	218418852	<i>DIRC3</i>	Body		+
cg19485911	2	220380542	<i>ACCN4</i>	Body	SouthShelf	+
cg04042861	2	231989824	<i>HTR2B</i>	TSS200		-
cg22374586	2	232220566				-
cg10459425	2	240405980				+
cg19697911	2	241080057	<i>OTOS</i>	5'UTR		-
cg00771084	2	242598843	<i>ATG4B</i>	Body		-
cg25315362	3	116164576	<i>LSAMP</i>	TSS200		-
cg25679475	3	118705126	<i>IGSF11</i>	Body		-
cg12389423	3	118864836	<i>C3orf30</i>	TSS200		-
cg18334977	3	128272847			North Shore	+
cg20912272	3	128765374			Island	-
cg11162839	3	135688971	<i>PPP2R3A</i>	5'UTR	SouthShelf	+
cg12361223	3	147089362				+
cg24620761	3	147123199	<i>ZIC4</i>	5'UTR	North Shelf	+
cg19516404	3	147123475	<i>ZIC4</i>	5'UTR	North Shelf	-
cg18592273	3	161089930	<i>C3orf57</i>	TSS200	Island	+

cg21548131	3	173639566	<i>NLGN1</i>	Body		+
cg00121562	3	196014011	<i>PCYT1A</i>	5'UTR	North Shore	-
cg14875327	3	197081654				-
cg20536207	4	883948	<i>GAK</i>	Body	North Shelf	-
cg06466757	4	1255808				+
cg17910931	4	1534933			Island	-
cg25144207	4	4864302	<i>MSX1</i>	Body	North Shore	-
cg11060856	4	5895410	<i>CRMP1</i>	TSS1500	SouthShore	-
cg17304168	4	15626104	<i>FBXL5</i>	Body		+
cg15407965	4	128707242	<i>HSPA4L</i>	Body	SouthShelf	-
cg13842222	4	185972881				+
cg09084892	4	187557837	<i>FAT1</i>	Body		+
cg09458384	5	3030971				-
cg07287793	5	6447238	<i>UBE2QL1</i>	TSS1500	North Shore	-
cg26496372	5	37379396	<i>WDR70</i>	TSS200	Island	+
cg25110832	5	71360847				+
cg11081752	5	117380176				+
cg11190434	5	172106450	<i>NEURL1B</i>	Body	North Shelf	+
cg19423543	5	172655948			Island	-
cg00618323	5	176515533	<i>FGFR4</i>	5'UTR	SouthShore	-
cg05184550	5	178632517	<i>ADAMTS2</i>	Body		-
cg00489401	5	180075875	<i>FLT4</i>	Body	Island	-
cg03349397	6	6588693	<i>LY86</i>	TSS1500		-

cg05927817	6	18020357				-
cg08014661	6	22334619				+
cg15289190	6	28831544			North Shore	+
cg08097157	6	31080048	<i>C6orf15</i>	Body		+
cg21286967	6	31696710	<i>DDAH2</i>	Body	Island	-
cg00686197	6	31733619	<i>C6orf27</i>	Body		-
cg23973371	6	32132649	<i>EGFL8</i>	5'UTR	North Shore	-
cg05629964	6	41085111	<i>LOC221442</i>	Body		+
cg25748868	6	41131213	<i>TREM2</i>	TSS1500		+
cg13312976	6	43303332				+
cg26590603	6	43478530	<i>C6orf154</i>	TSS200	Island	-
cg17427198	6	52529191	<i>LOC730101</i>	TSS200	Island	-
cg02595760	6	100066673			Island	-
cg01359933	6	144612696	<i>UTRN</i>	TSS200		+
cg24225668	6	160679811	<i>SLC22A2</i>	1stExon	SouthShore	-
cg12256206	6	168719708	<i>DACT2</i>	Body	Island	-
cg01740202	7	550568	<i>PDGFA</i>	Body	Island	-
cg09424595	7	4147330	<i>SDK1</i>	Body		-
cg09505513	7	5432820	<i>TNRC18</i>	Body	SouthShelf	+
cg12844895	7	6774169	<i>PMS2CL</i>	TSS1500	SouthShelf	+
cg04493169	7	27912112	<i>JAZF1</i>	Body		-
cg01353448	7	31726912	<i>C7orf16</i>	5'UTR		-

cg09214099	7	72791740			Island	-
cg15035350	7	83824382	SEMA3A	TSS200		-
cg06745145	7	90664816	CDK14	Body		+
cg03308706	7	91763433	CYP51A1	5'UTR	Island	-
cg00870778	7	97598492	MGC72080	Body	North Shelf	+
cg26914299	7	122326807	CADPS2	Body		-
cg04678743	7	130353515	TSGA13	3'UTR	Island	+
cg00984540	7	155589293			Island	-
cg02883229	7	155616337				+
cg17604655	7	156735383			Island	-
cg01731811	7	157890171	PTPRN2	Body	North Shore	-
cg12126859	8	335281				+
cg22932649	8	20140018				+
cg20585869	8	24772333	NEFM	TSS200	Island	-
cg06009497	8	37695050	GPR124	Body	North Shelf	-
cg08772302	8	37826337			SouthShelf	-
cg22848598	8	38965026	ADAM32	TSS200	Island	-
cg15706250	8	41583321	ANK1	Body	Island	-
cg21045828	8	89310065	MMP16	Body		+
cg24495007	8	143867989	LY6D	1stExon		-
cg10140583	8	143868110	LY6D	TSS200		-
cg09961689	8	144590068	ZC3H3	Body		-

cg17683573	8	145584770	<i>GPR172A</i>	3'UTR	SouthShelf	-
cg14286514	9	32525315	<i>DDX58</i>	Body	North Shore	-
cg13845049	9	77701327	<i>C9orf95</i>	Body	North Shore	+
cg14410137	9	96269920	<i>FAM120A</i>	Body		-
cg14027524	9	140120587	<i>C9orf169</i>	3'UTR	SouthShelf	-
cg14677909	10	48807341	<i>PTPN20B</i>	Body		-
cg16692757	10	88024572	<i>MIR346</i>	TSS200	SouthShore	-
cg19295034	10	95721819	<i>PIPSL</i>	TSS200		+
cg27190138	10	98479757	<i>PIK3AP1</i>	Body	Island	+
cg15275103	10	124893024			Island	-
cg18979589	10	125034818				-
cg14677612	10	131263962	<i>MGMT</i>	TSS1500	North Shore	+
cg18308755	10	134065956	<i>STK32C</i>	Body		-
cg03246584	10	134663467				+
cg14494090	10	134972969	<i>KNDC1</i>	TSS1500	North Shore	-
cg23313445	11	4792272	<i>OR51F1</i>	TSS1500		+
cg12575434	11	14214472	<i>SPON1</i>	Body		+
cg15659420	11	20034979	<i>NAV2</i>	Body		+
cg19375210	11	20183057	<i>DBX1</i>	TSS1500	North Shore	+
cg03728580	11	34460856	<i>CAT</i>	Body	Island	-
cg02508204	11	39367436				+

cg12472022	11	61462803	DAGLA	5'UTR		-
cg12690575	11	61536955	C11orf9	Body;Body		-
cg01981334	11	64877237	C11orf2	Body	North Shore	-
cg04293602	11	65553660	OVOL1	TSS1500	North Shore	-
cg04174538	11	66673303	PC	5'UTR		-
cg19204693	11	68206027	LRP5	Body	Island	-
cg22747802	11	68417633				-
cg06193239	11	68417787				-
cg15233880	11	69454727	CCND1	TSS1500	Island	-
cg10337956	11	85436135	SYTL2	Body		+
cg22049858	11	94884121			Island	+
cg10854423	11	100275402				+
cg22458194	11	113345686	DRD2	5'UTR	Island	-
cg12782933	11	116451038			Island	-
cg12781794	11	118505740	PHLDB1	Body	Island	-
cg24413662	11	122311293				-
cg20051949	11	126619800	KIRREL3	Body		+
cg20122043	11	132912205	OPCML	Body		+
cg09576209	12	2339614	CACNA1C	Body	Island	-
cg17844553	12	11322611	PRR4	5'UTR	North Shore	+
cg14990076	12	23116550				+
cg05419385	12	27352945				+
cg15545035	12	28128288			Island	-

cg23897083	12	34755568				+
cg16708880	12	49257591	<i>RND1</i>	Body		-
cg02574894	12	53693825	<i>C12orf10</i>	Body	Island	+
cg18827332	12	103344506			Island	-
cg24445165	12	113549832	<i>RASAL1</i>	Body	North Shore	-
cg23432930	12	133464933	<i>CHFR</i>	TSS1500	SouthShore	-
cg11315081	13	22651243				-
cg10512779	13	30998440			SouthShelf	+
cg23137039	13	50069509	<i>PHF11</i>	TSS1500	North Shore	-
cg23440004	13	79175611	<i>POU4F1</i>	Body	Island	-
cg01636662	14	60046159			SouthShelf	+
cg05881436	14	62331619			Island	-
cg02583546	14	77494451	<i>C14orf4</i>	5'UTR	Island	+
cg18225991	14	96851949	<i>C14orf129</i>	Body		+
cg27158340	14	105603389			Island	-
cg22475358	15	23455599			Island	-
cg14209037	15	41228521	<i>DLL4</i>	Body	Island	-
cg01908020	15	45494784	<i>SHF</i>	TSS1500	SouthShelf	+
cg23735339	15	51350231	<i>TNFAIP8L3</i>	Body		-
cg26481829	15	69744762	<i>RPLP1</i>	TSS1500	North Shore	-
cg08815652	15	71055727	<i>UACA</i>	1stExon	Island	-
cg23625341	15	71520142	<i>THSD4</i>	Body		+
cg05239311	15	78913147	<i>CHRNA3</i>	5'UTR	Island	-

cg01003448	16	745685	<i>FBXL16</i>	Body	Island	+
cg07482202	16	745687	<i>FBXL16</i>	Body	Island	+
cg02412803	16	1099138			North Shore	-
cg09074450	16	1198900			North Shore	-
cg26780022	16	1336537				-
cg05406475	16	1837011	<i>NUBP2</i>	Body	Island	-
cg16619935	16	2037439	<i>GFER</i>	3'UTR	North Shelf	-
cg00876678	16	2319586	<i>RNPS1</i>	TSS1500	SouthShore	-
cg27605307	16	20357506	<i>UMOD</i>	Body	North Shelf	-
cg05131483	16	23706242	<i>ERN2</i>	Body		-
cg26929163	16	34598029	<i>LOC283914</i>	Body		+
cg26586719	16	54555550				+
cg09942293	16	66957496	<i>RRAD</i>	Body	North Shore	-
cg08550881	16	84213572	<i>TAF1C</i>	Body	North Shore	-
cg24303478	16	89143845			Island	-
cg26916410	16	89642016	<i>CPNE7</i>	TSS200	Island	+
cg12187586	17	2627661			Island	-
cg12964144	17	5974448	<i>WSCD1</i>	5'UTR	Island	-
cg21507719	17	7256673	<i>KCTD11</i>	1stExon	SouthShore	-
cg25930644	17	8531915	<i>MYH10</i>	5'UTR	North Shore	+
cg17325958	17	18628980	<i>TRIM16L</i>	5'UTR		+
cg12550399	17	19482275	<i>SLC47A1</i>	3'UTR	North Shore	+

cg17996892	17	26832584				-
cg10060065	17	32366901	<i>ACCN1</i>	Body	Island	-
cg25170034	17	33288066	<i>ZNF830</i>	TSS1500	North Shore	-
cg02216727	17	38520653	<i>GJD3</i>	1stExon	SouthShore	-
cg11653266	17	73901339	<i>MRPL38</i>	TSS200	Island	+
cg16548154	17	74565757	<i>ST6GALNAC2</i>	Body		-
cg02958960	17	76098276	<i>TNRC6C</i>	Body	North Shelf	+
cg15209885	17	77753199	<i>CBX2</i>	Body	SouthShore	+
cg18576374	17	78549371	<i>RPTOR</i>	Body		+
cg00830755	17	80819020	<i>TBCD</i>	Body	Island	-
cg14307471	18	31432117	<i>NOL4</i>	3'UTR		+
cg03012785	18	54788429			North Shore	-
cg15704408	18	77245548	<i>NFATC1</i>	Body	North Shore	-
cg19299952	19	2078176	<i>MOBKL2A</i>	Body	Island	-
cg04351156	19	10562415	<i>PDE4A</i>	Body		-
cg16574191	19	14063234	<i>PODNL1</i>	Body	Island	-
cg02281038	19	14139137	<i>RLN3</i>	1stExon	North Shelf	-
cg07356415	19	19655352	<i>CILP2</i>	Body	Island	-
cg00910067	19	33717545	<i>SLC7A10</i>	TSS1500	Island	-
cg19985870	19	34398004			Island	-
cg12497870	19	36210913	<i>MLL4</i>	Body	Island	-

cg25378939	19	36912702			Island	-
cg12088773	19	44128330	<i>CADM4</i>	Body		-
cg12610917	19	46387992	<i>IRF2BP1</i>	1stExon	Island	-
cg17583504	19	49669542	<i>TRPM4</i>	Body	Island	-
cg00086493	19	51535348	<i>KLK12</i>	Body	Island	-
cg03781262	19	58879871	<i>ZNF837</i>	Body	Island	-
cg14083397	20	388473	<i>RBCK1</i>	TSS1500	Island	+
cg11004890	20	3218500	<i>SLC4A11</i>	TSS200	North Shore	-
cg07436991	20	11871311	<i>BTBD3</i>	TSS200	North Shore	-
cg10629004	20	21696467	<i>PAX1</i>	3'UTR	SouthShore	-
cg24147187	20	36535798	<i>VSTM2L</i>	Body	SouthShelf	-
cg06457011	20	39767490	<i>PLCG1</i>	Body	SouthShore	+
cg27207308	20	47935683			Island	-
cg25188239	20	48532311	<i>SPATA2</i>	TSS1500	Island	+
cg10187707	20	49626842	<i>KCNG1</i>	Body	Island	-
cg24849555	20	52781187	<i>CYP24A1</i>	Body		+
cg12978433	20	52789956	<i>CYP24A1</i>	1stExon	Island	-
cg17181362	20	57090749	<i>APCDD1L</i>	TSS1500	SouthShore	-
cg00199007	20	61583910	<i>SLC17A9</i>	TSS200	Island	+
cg09576415	20	62059559	<i>KCNQ2</i>	Body	Island	-
cg24936695	21	31538995	<i>CLDN17</i>	TSS200		+

cg08073527	21	43256581	<i>PRDM15</i>	Body	SouthShore	+
cg19312314	21	44473962	<i>CBS</i>	3'UTR	Island	+
cg12826791	21	45926719	<i>C21orf29</i>	Body	Island	-
cg09443102	21	46824976	<i>COL18A1</i>	TSS200	Island	-
cg11985360	22	19138209	<i>GSC2</i>	TSS1500	Island	-
cg10047755	22	19751776	<i>TBX1</i>	Body	Island	-
cg27665648	22	30112403			North Shelf	-
cg08215954	22	30962134	<i>GAL3ST1</i>	TSS1500		-
cg23018242	22	31608245	<i>LIMK2</i>	TSS200	Island	+
cg17998530	22	35388865				+
cg00077838	22	47020733			North Shore	-
cg17145402	22	47189485	<i>TBC1D22A</i>	Body	Island	-
cg04195684	22	50515473	<i>MLC1</i>	Body	Island	+
cg00256932	22	51041732	<i>MAPK8IP2</i>	1stExon	North Shore	-

Table 9.17. Table comparing results published by Tryndyak *et al.* (2018)³²⁵ and the combined PAH8 exposure EWAS results

Gene Name	Tryndyak <i>et al.</i> 2018³²⁵					Combined PAH8 EWAS Results		
	<u>Chromosome</u>	<u>Start</u>	<u>End</u>	<u>Genomic Location</u>	<u>Direction of Methylation Change</u>	<u>Probe ID</u>	<u>Genomic Location</u>	<u>Direction of Methylation Change</u>
<i>BTBD3</i>	chr20	11870709	11870818	Promoter; TSS	+	cg07436991	TSS200	-
<i>DLL4</i>	chr15	41230332	41230369	3' UTR	-	cg14209037	Gene body	-
	chr15	41230466	41230643	3' UTR	-			
<i>NAV2</i>	chr11	19955509	19955640	Exon	-	cg15659420	Gene body	+
<i>RASAL1</i>	chr12	113573301	113573980	Promoter; TSS	-	cg24445165	Gene body	-
<i>SDK1</i>	chr7	4308985	4309123	TTS	+	cg09424595	Gene body	-
<i>TBX1</i>	chr22	19771606	19771687	TTS	-	cg10047755	Gene body	-

Table 9.18. Table showing overlaps between results of combined air and dietary PAH8 exposure, and results from published smoking EWAS. Overlaps were identified by looking for exact CpG probes and by looking for probes with the same genes.

Gene	Study	CpG	Direction	Tissue	CpG	Direction
<i>ACAP3</i>	Joehanes, 2016 ⁴¹⁵	cg27185793	+	Blood	cg19695266	+
<i>ACCN1</i>	Joehanes, 2016 ⁴¹⁵	cg02423064	+	Blood	cg10060065	-
		cg19942495				
<i>ADAMTS2</i>	Joehanes, 2016 ⁴¹⁵	cg17359265	-	Blood	cg05184550	-
	Joubert, 2016 ⁴¹⁶	cg10997906	+			
<i>ANK1</i>	Joehanes, 2016 ⁴¹⁵	cg12634208	+	Blood	cg15706250	-
	Joubert, 2016 ⁴¹⁶	cg01453458				
		cg27619646				
<i>APCDD1L</i>	Dogan, 2014 ⁴²¹	cg00950473	-	Blood	cg17181362	-
<i>ARHGEF10L</i>	Dogan, 2014 ⁴²¹	cg21696055	-	Blood	cg21908208	-
<i>BTBD3</i>	Joubert, 2016 ⁴¹⁶	cg00592643	+	Blood	cg07436991	-
		cg24562149				
<i>C11orf2</i>	Joubert, 2016 ⁴¹⁶	cg13626866	+	Blood	cg01981334	-
<i>C14orf4</i>	Besingi, 2014 ³⁸⁰	cg02583546	+	Whole blood	cg02583546	+
<i>C6orf154</i>	Joubert, 2016 ⁴¹⁶	cg19687985	+	Blood	cg26590603	-
		cg26590603				
<i>C6orf27</i>	Guida, 2015 ⁴²⁵	cg19868593	-	Blood	cg00686197	-
	Joehanes, 2016 ⁴¹⁵	cg08409562	+			
	Joubert, 2016 ⁴¹⁶	cg24065328	-			
<i>CACNA1C</i>	Joehanes, 2016 ⁴¹⁵	cg02959759	+	Blood	cg09576209	-
<i>CCND1</i>	Lee, 2016 ⁴⁵⁰	cg09520904	-	Blood	cg15233880	-

<i>CHFR</i>	Joubert, 2016 ⁴¹⁶	cg16482759 cg23432930	-	Blood	cg23432930	-
<i>CILP2</i>	Joubert, 2016 ⁴¹⁶	cg07942040	+	Blood	cg07356415	-
<i>COL18A1</i>	Joubert, 2016 ⁴¹⁶	cg05349624 cg07279557 cg14903689	-	Blood	cg09443102	-
<i>CPNE7</i>	Joubert, 2016 ⁴¹⁶	cg02500990 cg16616467	+	Blood	cg26916410	+
<i>CTNNA2</i>	Richmond, 2015 ⁴³⁸	cg27629977	+	Cord blood	cg23665824	-
<i>CYP51A1</i>	Allione, 2015 ⁴¹⁷	cg10655371	-	Whole blood	cg03308706	-
<i>DACT2</i>	Dogan, 2014 ⁴²¹	cg21223803	-	Blood	cg12256206	-
<i>DAGLA</i>	Dogan, 2014 ⁴²¹	cg18766608	-	Blood	cg12472022	-
<i>DDAH2</i>	Ivorra, 2015 ⁴²⁷ Joubert, 2016 ⁴¹⁶	cg15264752 cg26111283	+	Cord blood Blood	cg21286967	-
<i>DIRC3</i>	Joubert, 2016 ⁴¹⁶	cg01396774 cg15912082	-	Blood	cg05511924	+
<i>DPP10</i>	Chhabra, 2014 ⁴²⁰	cg22670147	-	Lung	cg12448298	+
<i>EDN2</i>	Guida, 2015 ⁴²⁵ Joehanes, 2016 ⁴¹⁵ Joubert, 2016 ⁴¹⁶	cg16736826 cg16736826 cg16736826	-	Blood	cg21862529	-
<i>EGFL8</i>	Allione, 2015 ⁴¹⁷	cg10502563	-	Whole blood	cg23973371	-
<i>FBXL16</i>	Joehanes, 2016 ⁴¹⁵ Joubert, 2016 ⁴¹⁶	cg05542681 cg02713960 cg02958327 cg26804595	+	Blood	cg01003448 cg07482202	+

	Markunas, 2014 ⁴³³	cg26804595	-			
FBXL5	Joubert, 2016 ⁴¹⁶	cg02630888 cg15175162	+	Blood	cg17304168	+
GAK	Guida, 2015 ⁴²⁵	cg06154597	-	Blood	cg20536207	-
	Joehanes, 2016 ⁴¹⁵	cg01552919	+			
GJD3	Joubert, 2016 ⁴¹⁶	cg05568941 cg05930207 cg06949812 cg11758793	+	Blood	cg02216727	-
	Markunas, 2014 ⁴³³	cg05568941 cg06949812				
GPR124	Joehanes, 2016 ⁴¹⁵	cg20272648 cg01226742		Blood	cg06009497	-
	Joubert, 2016 ⁴¹⁶	cg05552035 cg12424646 cg12869334	-			
GRID1	Joubert, 2016 ⁴¹⁶	cg04422256	-	Blood		
GSC2	Dogan, 2014 ⁴²¹	cg02917246	-	Blood	cg11985360	+
HTR2B	Joehanes, 2016 ⁴¹⁵	cg04042861	-	Blood	cg04042861	-
	Sun, 2013 ⁴⁴²	cg06096336				
IRF2BP1	Joehanes, 2016 ⁴¹⁵	cg08097614	-	Blood	cg12610917	-
JAZF1	Joubert, 2016 ⁴¹⁶	cg02010481 cg14491535 cg22938901 cg26438325	+	Blood	cg04493169	-
KCNG1	Dogan, 2014 ⁴²¹	cg03027241	-	Blood	cg10187707	-

<i>KCNQ2</i>	Joubert, 2016 ⁴¹⁶	cg13379325	-	Blood	cg09576415	-
		cg03387585	+			
<i>KIRREL3</i>	Joubert, 2016 ⁴¹⁶	cg09737499	-	Blood	cg20051949	+
		cg18434848	-			
		cg04445570	+			
		cg12322672	+			
<i>KNDC1</i>	Joubert, 2016 ⁴¹⁶	cg01258050	-	Blood	cg14494090	-
<i>LRP5</i>	Besingi, 2014 ³⁸⁰	cg04265051	+	Whole blood	cg19204693	-
		cg21611682	-			
	Dogan, 2014 ⁴²¹	cg06989074	+	Blood		
		cg09578155				
	Guida, 2015 ⁴²⁵	cg10420527				
		cg14624207	-	Blood		
		cg21611682				
		cg21746120				
	Joehanes, 2016 ⁴¹⁵	cg09578155	-			
		cg10420527	-			
		cg14624207	-	Blood		
		cg21611682	-			
cg24051242		+				
cg09578155						
Joubert, 2016 ⁴¹⁶	cg21611682					
	cg21916461	-	Blood			
	cg22151881					
	cg23949925					
Kupers, 2015 ⁴²⁹	cg21611682	-	Cord blood			
Li, 2018 ⁴³¹	cg21611682	-	Blood			
Shenker, 2013 ⁴⁴⁰	cg21611682	+	Blood			

	Tsaprouni, 2014 ⁴⁴⁴	cg21611682	-	Blood		
	Zeilinger, 2013 ⁴⁴⁸	cg14624207	-	Blood		
		cg21611682				
	Zhu, 2016 ⁴⁴⁹	cg09578155	-	Leukocytes		
		cg10420527				
		cg14624207				
		cg21611682				
MGC72080	Markunas, 2014 ⁴³³	cg20174893	-	Blood	cg00870778	+
MGMT	Allione, 2015 ⁴¹⁷	cg14312783	-	Whole blood	cg14677612	+
		cg27483317				
	Joubert, 2012 ⁴²⁸	cg09993459		Cord blood		
	Guida, 2015 ⁴²⁵	cg07381806		Blood		
		cg15187398				
		cg06896207				
	Joehanes, 2016 ⁴¹⁵	cg07381806	-	Blood		
		cg15187398				
		cg17931529				
MOBKL2A	Joubert, 2016 ⁴¹⁶	cg06896207		Blood	cg19299952	-
	Li, 2018 ⁴³¹	cg15187398		Blood		
	Zeilinger, 2013 ⁴⁴⁸	cg07381806		Blood		
		cg15187398		Blood		
	Zhu, 2016 ⁴⁴⁹	cg07381806		Leukocytes		
		cg15187398		Leukocytes		
MSX1	Joubert, 2016 ⁴¹⁶	cg01785568	-	Blood	cg25144207	-
		cg11078084				
MYH10	Dogan, 2014 ⁴²¹	cg06557376	+	Blood	cg25930644	+
	Joehanes, 2016 ⁴¹⁵	cg09975715				

NAV1	Joubert, 2016 ⁴¹⁶	cg08883485 cg14920846	-	Blood	cg04226892	-
	Guida, 2015 ⁴²⁵	cg04039799	-	Blood		
NAV2	Ivorra, 2015 ⁴²⁷	cg01249134 cg03529555	+	Cord blood, Blood	cg15659420	+
	Joehanes, 2016 ⁴¹⁵	cg12535090 cg03220447	+	Blood		
	Joubert, 2016 ⁴¹⁶	cg04039799 cg12711760	-	Blood		
	Zeilinger, 2013 ⁴⁴⁸	cg04039799	-	Blood		
NEURL1B	Joehanes, 2016 ⁴¹⁵	cg00327072	-	Blood	cg11190434	+
NFATC1	Dogan, 2014 ⁴²¹	cg05944967 cg24538512	+		cg15704408	-
	Joehanes, 2016 ⁴¹⁵	cg05302701 cg05753993	-	Blood		
	Joubert, 2016 ⁴¹⁶	cg06784563 cg15363134	-			
OVOL1	Allione, 2015 ⁴¹⁷	cg10604040	-	Whole blood	cg04293602	-
PC	Joubert, 2016 ⁴¹⁶	cg03229682	+	Blood	cg04174538	-
PCYT1A	Joubert, 2016 ⁴¹⁶	cg22221575	+	Blood	cg00121562	-
PDGFA	Kupers, 2015 ⁴²⁹	cg05556923	-	Cord blood	cg01740202	-
PHF11	Dogan, 2014 ⁴²¹	cg22924269	-	Blood	cg23137039	-
PHLDB1	Joubert, 2016 ⁴¹⁶	cg20110707	-	Blood	cg12781794	-
PLCG1	Joubert, 2016 ⁴¹⁶	cg24961795	-	Blood	cg06457011	+
	Allione, 2015 ⁴¹⁷	cg09934852	-	Whole blood	cg16574191	-

PODNL1	Joehanes, 2016 ⁴¹⁵	cg18547299	+	Blood		
PPP1R15B	Ivorra, 2015 ⁴²⁷	cg00093900	+	Cord blood, Blood	cg04662939	-
PRDM15	Joehanes, 2016 ⁴¹⁵	cg18151030	-	Blood	cg08073527	+
	Allione, 2015 ⁴¹⁷	cg00109293	-	Whole blood		
		cg25372239				
	Dogan, 2014 ⁴²¹	cg00068377	+	Blood		
		cg15386853	-			
		cg00806481				
		cg03126058				
		cg12297125	-			
	Joehanes, 2016 ⁴¹⁵	cg22510139		Blood		
		cg25618424				
		cg04134748	+			
		cg10493186				
PRDM16		cg01261194			cg02767788	-
		cg01418153			cg26913155	+
		cg01431482			cg06182121	+
		cg03254465				
		cg04134748				
		cg05804170				
	Joubert, 2016 ⁴¹⁶	cg08262220	-	Blood		
		cg11138362				
		cg11731671				
		cg12133962				
		cg12408250				
		cg12436196				
		cg13388191				

	cg13393782				
	cg15090440				
	cg17001566				
	cg17445936				
	cg17940849				
	cg18369939				
	cg18509466				
	cg19243842				
	cg19904265				
	cg21848084				
	cg22122862				
	cg22510139				
	cg22726349				
	cg22729726				
	cg25618424				
	cg26425711				
	cg12441214				
	cg24939838	+			
	Kupers, 2015 ⁴²⁹	cg252153667	-	Cord blood	
	Allione, 2015 ⁴¹⁷	cg16059943	-	Whole blood	
	Dogan, 2014 ⁴²¹	cg09180820	-	Blood	
		cg23629792			
<i>PRKCZ</i>	Freeman, 2016 ⁴²⁴	cg11345323	-	Lung adenocarcinoma, Lung squamous cell	cg24937768
		cg22865720			
	Joehanes, 2016 ⁴¹⁵	cg24842354	-	Blood	
	Joubert, 2016 ⁴¹⁶	cg02393699	+	Blood	
		cg09225489	-		

		cg27264462	+			
	Allione, 2015 ⁴¹⁷	cg07305000	-	Whole blood		
	Besingi, 2014 ³⁸⁰	cg15340709	+	Whole blood		
	Dogan, 2014 ⁴²¹	cg14743683	-	Blood		
		cg00566158	+			
	Joehanes, 2016 ⁴¹⁵	cg02223801	-	Blood		
		cg05433557	+			
<i>PTPRN2</i>		cg23385492	+			
	Joubert, 2012 ⁴²⁸	cg02356647	-	Cord blood	cg01731811	-
		cg02637474				
		cg02660277				
	Joubert, 2016 ⁴¹⁶	cg14338779	-	Blood		
		cg17748769				
		cg18064706				
		cg19350216				
<i>RASAL1</i>	Joubert, 2016 ⁴¹⁶	cg07140497	+	Blood	cg24445165	-
		cg19721065				
<i>RNF207</i>	Joubert, 2016 ⁴¹⁶	cg17515966	-	Blood	cg00030047	-
		cg19694465	+			
	Allione, 2015 ⁴¹⁷	cg21289763	-	Whole blood		
		cg26469982				
	Dogan, 2014 ⁴²¹	cg02933375	-	Blood		
		cg17872658				
<i>RPTOR</i>		cg01498832	+			
	Joehanes, 2016 ⁴¹⁵	cg15228441	-	Blood	cg18576374	+
		cg01561259	+			
		cg18780100	+			
	Joubert, 2016 ⁴¹⁶	cg16541275	-	Blood		
		cg26360197				

		cg01498832				
		cg03794617				
		cg07126783				
		cg08939850				
		cg15230985	+			
		cg16638092				
		cg18780100				
		cg27511181				
<i>SDK1</i>	Joehanes, 2016 ⁴¹⁵	cg05642264	+	Blood		
	Joubert, 2012 ⁴²⁸	cg21005410	-	Cord blood	cg09424595	-
	Joubert, 2016 ⁴¹⁶	cg16639880	-	Blood		
		cg26180191				
<i>SEMA4C</i>	Joubert, 2016 ⁴¹⁶	cg11344744	-	Blood	cg22591002	-
<i>SHF</i>	Dogan, 2014 ⁴²¹	cg22496377	+	Blood	cg01908020	+
<i>SLC17A9</i>	Joubert, 2016 ⁴¹⁶	cg15677087	-	Blood	cg00199007	+
<i>SLC4A11</i>	Joubert, 2016 ⁴¹⁶	cg26130864	+	Blood	cg11004890	-
<i>SPON1</i>		cg09191626				
	Joubert, 2016 ⁴¹⁶	cg20693209	-	Blood	cg12575434	+
		cg22805485				
<i>ST6GALNAC2</i>	Shenker, 2013 ⁴⁴⁰	cg14385325	-	Blood	cg16548154	-
<i>SYTL2</i>	Allione, 2015 ⁴¹⁷	cg11773367	-	Whole blood	cg10337956	+
	Joubert, 2016 ⁴¹⁶	cg27312916		Blood		
<i>TBC1D22A</i>	Allione, 2015 ⁴¹⁷	cg13554549	-	Whole blood	cg17145402	-
<i>TBCD</i>		cg07602659				
	Joehanes, 2016 ⁴¹⁵	cg16755833	+	Blood	cg00830755	-
	Joubert, 2016 ⁴¹⁶	cg07769421	-			

		cg11183935				
		cg23352492				
TBX1	Joubert, 2016 ⁴¹⁶	cg02948624	-	Blood	cg10047755	-
		cg10647101				
TNFAIP8L3	Monick, 2012 ⁴³⁴	cg02233197	+	Alveolar macrophage	cg23735339	-
	Allione, 2015 ⁴¹⁷	cg09022230	-	Whole blood		
	Besingi, 2014 ³⁸⁰	cg09022230	-	Whole blood		
	Dogan, 2014 ⁴²¹	cg09794469	+	Blood		
	Guida, 2015 ⁴²⁵	cg09022230	-	Blood		
TNRC18	Joehanes, 2016 ⁴¹⁵	cg09022230	-	Blood	cg09505513	+
	Joubert, 2016 ⁴¹⁶	cg09022230	-	Blood		
	Li, 2018 ⁴³¹	cg09022230	-	Blood		
	Philibert, 2012 ⁴³⁶	cg23268879	-	Lymphocyte, female		
	Zhu, 2016 ⁴⁴⁹	cg09022230	-	Leukocytes		
TRPM4	Dogan, 2014 ⁴²¹	cg19017254	-	Blood	cg17583504	-
	Monick, 2012 ⁴³⁴	cg10951975		Lymphoblast		
TSSC1	Allione, 2015 ⁴¹⁷	cg03661054	-	Whole blood	cg14950134	-
	Dogan, 2014 ⁴²¹	cg21557724	+	Blood		
WDR35	Joubert, 2016 ⁴¹⁶	cg01055594	+	Blood	cg22939193	-
		cg23811268	-			
	Allione, 2015 ⁴¹⁷	cg26361535	-	Whole blood		
	Guida, 2015 ⁴²⁵	cg26361535	-	Blood		
ZC3H3	Joehanes, 2016 ⁴¹⁵	cg26361535	-	Blood	cg09961689	-
		cg21404980	+			
		cg12688965	-			
	Joubert, 2016 ⁴¹⁶	cg26361535	-	Blood		
		cg14740860	+			

	Zeilinger, 2013 ⁴⁴⁸	cg26361535	-	Blood		
<i>ZIC4</i>	Joubert, 2016 ⁴¹⁶	cg13897134	+	Blood	cg19516404	-
					cg24620761	+

