



I made this template based on thesistown to comply with the
University of Bristol regulations

Thomas Battram
MRC Integrative Epidemiology Unit
Bristol Medical School
Faculty of Health Sciences
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the
requirements for award of the degree of Population Health Sciences in the
Faculty of Health Sciences

Bristol Medical School, September 2020, Word Count:

Abstract

My abstract will go here and it will be a solid abstract. Full of the things that go in abstracts. Such as numbers, acronyms, other words, and lots of punctuation.

It will have multiple paragraphs too!

Acknowledgements

THIS IS WHERE YOU THANK PEOPLE!!

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed

Dated

Table of Contents

Preface	1
Chapter 1: Introduction	3
1.1 DNA methylation as part of the regulatory machinery	4
1.1.1 Gene expression regulation	5
1.1.2 Defining epigenetics	6
1.1.3 Histone modifications	6
1.1.4 DNA methylation	7
1.2 Use of DNA methylation in a population setting	9
1.2.1 The appeal of DNA methylation to epidemiologists	9
1.2.2 Epigenome-wide association studies	11
1.2.3 Problems for EWAS	11
Confounding	12
Cell type heterogeneity	12
Measuring DNA methylation	13
Complexity of regulatory mechanisms	15
Treatments	15
1.3 Using methods from genetics to help inform future EWAS	16
1.3.1 The GWAS Catalog	17
1.3.2 Total variance captured by all sites measured genome-wide	17
1.3.3 Inferring biology from signals	19
1.3.4 Establishing causality	20
Mendelian randomization	20
Availability of data for MR	21
Assumptions of MR	22
Applying MR in a DNA methylation context	23
1.4 Overview of thesis aims	25

Chapter 2: Methods	27
Chapter 3: The EWAS Catalog: a database of epigenome-wide association studies	29
3.1 Abstract	29
3.2 Introduction	30
3.3 Methods	32
3.3.1 Implementation	32
3.3.2 Overview of publication data extraction	32
3.3.3 Overview of GEO data extraction	33
3.3.4 EWAS methods	34
Avon Longitudinal Study of Parents and Children (ALSPAC)	34
GEO datasets	34
3.4 Results	35
3.4.1 Database interface and use	35
3.5 Discussion	37
Chapter 4: Properties of EWAS	39
Chapter 5: h2ewas	41
Chapter 6: EWAS-GWAS comparison	43
Chapter 7: Appraising the causal relevance of DNA methylation for risk of lung cancer	45
7.1 Abstract	45
7.1.1 Background	45
7.1.2 Methods	46
7.1.3 Results	46
7.1.4 Conclusions	46
7.2 Introduction	47
7.3 Methods	48
7.3.1 EWAS study details	48
European Prospective Investigation into Cancer and Nutrition-Italy (EPIC-Italy)	49
Melbourne Collaborative Cohort Study (MCCS)	49
Norwegian Women and Cancer (NOWAC)	50
Northern Sweden Health and Disease Study (NSHDS)	51
7.3.2 EWAS Meta-analysis	52
7.3.3 Mendelian randomization	53

	Sample 1: Accessible Resource for Integrated Epigenomic Studies (ARIES)	53
7.3.4	Sample 2: Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium . . .	55
7.3.5	Supplementary analyses	57
	Assessing the potential causal effect of <i>AHRR</i> methylation: one sample MR	57
	Tumour and adjacent normal methylation patterns	59
	mQTL association with gene expression	59
7.4	Results	60
7.4.1	EWAS meta-analysis	60
7.4.2	Mendelian randomization	60
	Potential causal effect of <i>AHRR</i> methylation on lung cancer risk: one sample MR	62
7.4.3	Tumour and adjacent normal lung tissue methylation patterns	63
7.4.4	Gene expression associated with mQTLs in blood and lung tissue	63
7.5	Discussion	64
	Conclusion	69
	Appendix A: The First Appendix	71
	Appendix B: The Second Appendix, for Fun	73
	References	75

List of Tables

List of Figures

1.1	Graphic of epigenetic marks taken from (5).	5
1.2	Simplified diagrams of the associations between DNA methylation and gene regulation. In <i>a</i> , CpG sites are methylated at the promoter region, but not in the gene body, which is associated with lack of gene expression. In <i>b</i> , the opposite is occurring.	9
1.3	Mendelian randomization. Mendelian randomization can be used to test the causal nature of exposure-outcome relationships provided the assumptions are met. Assumption 1. There is an association between the instrument and the exposure. Assumption 2. There are no associations between the instrument and outcome, except through the exposure. Assumption 3. The instrument is not associated with any factors that confound the exposure-outcome relationship. . . .	23
1.4	Flowchart showing questions being asked in the thesis and the corresponding chapters attempting to help answer them.	26
3.1	Using the EWAS catalog. At the top of the figures is the home page URL, ewascatalog.org. Below that are examples of three types of searches possible: 1. CpG sites, 2. genes and 3. traits. Finally, the results are displayed after searching the catalog for “Depression”. Circled in red is the download button, this button enables the user to download the results of their search as a tab-separated value file. This file will contain the information shown on the website as well as additional analysis information.	36

Preface

This template is based on (and in many places copied directly from) the Reed College LaTeX template, but hopefully it will provide a nicer interface for those that have never used TeX or LaTeX before. Using *R Markdown* will also allow you to easily keep track of your analyses in **R** chunks of code, with the resulting plots and output included as well. The hope is this *R Markdown* template gets you in the habit of doing reproducible research, which benefits you long-term as a researcher, but also will greatly help anyone that is trying to reproduce or build onto your results down the road.

Hopefully, you won't have much of a learning period to go through and you will reap the benefits of a nicely formatted thesis. The use of LaTeX in combination with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file is smaller than a Word file. While you may have never had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you've written before, taxing Word's capabilities. After working with *Markdown* and **R** together for a few weeks, we are confident this will be your reporting style of choice going forward.

Why use it?

R Markdown creates a simple and straightforward way to interface with the beauty of LaTeX. Packages have been written in **R** to work directly with LaTeX to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to LaTeX, *R Markdown* also allows you to read in your data, to analyze it and to visualize it using **R** functions, and also to provide the documentation and commentary on the results of your project. Further, it allows for **R** results to be passed inline to the commentary of your results. You'll see more on this later.

Who should use it?

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about the final appearance of their document should use *R Markdown*. Of particular use should be anyone in the sciences, but the user-friendly nature of *Markdown* and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should make it of great benefit to nearly anyone writing a thesis project.

For additional help with bookdown Please visit the free online bookdown reference guide.

Chapter 1

Introduction

DNA methylation is a dynamic, chemical modification to DNA that can occur at millions of sites across the genome. The modification may impact various traits as well as be an important biomarker or predictor. This thesis will focus on the most common study design for assessing the association between DNA methylation and complex traits, epigenome-wide association studies (EWAS).

To help inform design of future studies, there is a need to understand what has been discovered thus far in EWAS and what future EWAS are likely to add in the context of current EWAS and other study designs. A key element not understood is why DNA methylation at certain sites or regions of the genome tends to be associated with multiple traits. Potential reasons include technical artefacts, genetic factors or other strong confounding factors that are biasing results, or they could genuinely be of biological importance. Further, the proportion of trait variation captured by DNA methylation at the sites measured is unknown, which means one can only speculate at how much information can be gained from EWAS. Even if there is variation captured by DNA methylation sites, it's unclear whether identifying these

sites provide extra biological understanding of traits above what can be found from other methods. It is also imperative to fully explore whether these sites are causally related to the trait of interest before making claims about how DNA methylation influences it.

In this chapter I describe DNA methylation in the context of the regulatory processes in human cells, discuss its potential for use in population level studies and describe the current state of EWAS research. Then I explain how epigenetic epidemiologists can draw on the methods developed by geneticists to understand 1. what information has been gained from EWAS, 2. what information is left to gain from EWAS and 3. the causal nature of DNA methylation-trait associations identified in EWAS.

1.1 DNA methylation as part of the regulatory machinery

DNA methylation is highly correlated with gene expression regulation (1–4). This is the main mechanism by which it is thought DNA methylation can influence changes in phenotype. However, DNA methylation is just one of many epigenetic marks that are involved in gene regulation (**Figure 1.1**), so here I briefly outline some of these regulatory marks and discuss DNA methylation in the context of this complex biological process.

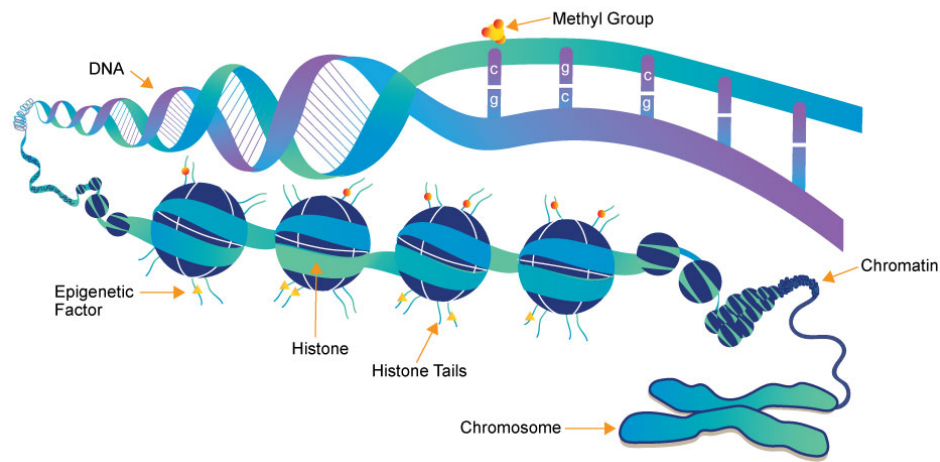


Figure 1.1: Graphic of epigenetic marks taken from (5).

1.1.1 Gene expression regulation

Gene expression is tightly controlled and only in the right circumstances will RNA polymerase II succeed in binding the correct site, initiating and finally completing the transcription process (6). The importance of this level of control is no more apparent than in the developmental stages of human life. Humans start as a single cell and after roughly nine months are transformed into a multicellular organism with trillions of cells, including hundreds of unique cell types. As these cells arise from a single progenitor they must contain identical genetic sequences (except a few somatic mutations). Therefore, the process by which the body is able to create such diversely functioning cells and tissues, must come from regulation of how the genetic sequence is read and from the regulation of its products. There are a plethora of methods cells use to regulate gene expression, protein expression and protein function post-transcriptionally (7–10), but these are beyond the scope of this thesis. So I'll continue to describe gene expression regulation just in the context of

epigenetic marks.

1.1.2 Defining epigenetics

The definition of epigenetics is much debated (11). In the 1940s the ‘epigenetic landscape’ was introduced by Waddington to describe how genes influence cell fates (12). Since then the term “epigenetics” has been used in many forms, so to avoid confusion, throughout the rest of the thesis, here is what I mean when using the term **epigenetics**: the study of mitotically heritable changes in gene expression that occur without changes in DNA sequence. When referring to **epigenetic marks** I’m referring to chemical changes to the genome and genome-bound proteins that are mitotically heritable and may influence gene expression without changing the DNA sequence.

1.1.3 Histone modifications

Histones are DNA-bound proteins that when modified can influence chromatin state and transcription factor interactions with the DNA. These proteins are contained in nucleosomes, which consist of four histone proteins, each of which is present twice in the nucleosome. Post-translational modifications can occur to any of the histone monomers and these have been associated with both positive and negative changes in gene expression (13–15). Histone modifications are numerous and complex in nature. To briefly describe the complexity, there are at least ten types of histone modification that can occur, each of the histone monomers can be modified across many different sites, and for any one site multiple of the same modification can occur. It is the combination of modifications across all histones that plays a role in gene expression regulation (14,15). Furthermore, histone modifications are subject to rapid change upon environmental stimulus to help induce or repress gene expression

(13). The complexity of histone modifications and the practical difficulties in collecting or assaying samples to assess these epigenetic marks remains a barrier to their wide-spread measurement for use in population-based analyses (16). They may become far more prominent in the future as our understanding and ability to measure the modifications in a meaningful way increases.

1.1.4 DNA methylation

DNA methylation is the addition of a methyl group to the DNA, which primarily occurs at the 5' cytosine of a CpG site. Little is known about the role of non-CpG site DNA methylation and current EWAS tend to only measure CpG methylation, so that will be my focus here. The function of this epigenetic mark was proposed back in 1975, where two papers suggested it represses gene expression (17,18). Since then, its association with various important intracellular processes such as X-inactivation, genomic imprinting and suppression of transposon action have been elucidated (18–20). Despite this, the role DNA methylation plays in regulating gene expression is not fully understood and research is still ongoing.

One thing research has revealed is that the location of DNA methylation is important to its relationship with gene expression. CpG sites are not randomly distributed throughout the genome, but are often found in clusters (4) and it's thought that methylation and de-methylation of CpG sites in groups is what drives their association with regulatory function. Location of the CpG sites relative to genes and other regulatory elements is also of importance. Several studies have shown that increased DNA methylation at transcription start sites tends to be associated with a decrease in gene expression (4,21,22), but gene body DNA methylation is positively correlated with expression (**Figure 1.2**) (23,24). This suggests that regulation of DNA methylation in clusters at specific sites relative to genes are what drive its relationship with

gene expression. To support this, there are clear biological processes that regulate DNA methylation at nearby sites together, for example, CpGs at transcription factor binding sites can be de-methylated as a group when the transcription factor binds (25). Further, nearby sites are often statistically correlated (26,27). However, there is no evidence to suggest that neighbouring sites do indeed act in tandem or whether it is likely one site from the group is driving regulatory function. This is something I explore in (**Chapter 5**).

These strong associations between DNA methylation and gene expression does not necessarily mean that the addition or removal of methyl groups will actively impact gene expression. Elucidating the causal nature of the association between DNA methylation and gene expression has been fraught with difficulties and has often provided conflicting results. One study showed an enzyme that catalyses the addition of methyl groups to the DNA, DNA methyltransferase 3A, is required in haematopoietic stem cells and for them to differentiate, suggesting gene expression changes required for differentiation were not possible without addition of methyl groups to the DNA (28). However, papers have provided evidence that DNA methylation is unlikely to initiate the ‘silencing’ of gene expression and may occur at transcription start sites of genes after they’ve already been repressed (29,30). To further complicate things, if DNA methylation does influence gene expression, the mechanism of action is unclear and may depend on the gene being examined. One study showed the presence of DNA methylation at the binding sites of the transcription factor, *MYC*, was inversely associated with its binding (31), but another study suggested the presence of DNA methylation didn’t have the same impact on the binding of the transcription factor, *SPI*. Although the body of work presented in this thesis does not aim to explore if and how DNA methylation influences gene expression, it is important to note the relationship between the two isn’t clear when

thinking of the implications of DNA methylation-trait associations, this will be discussed further in the following sections.

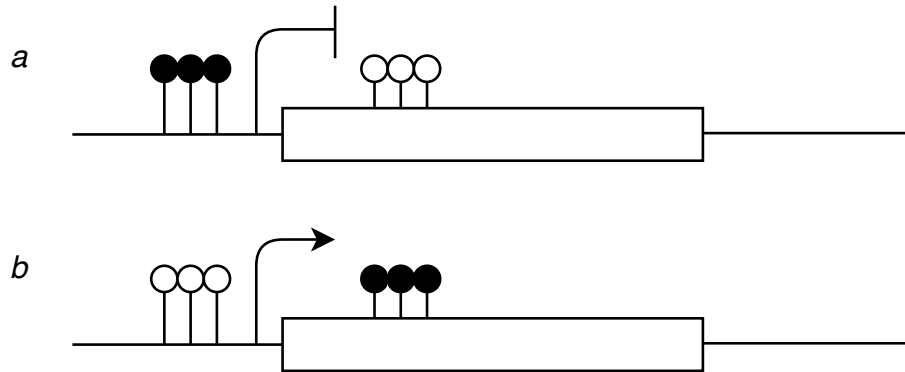


Figure 1.2: Simplified diagrams of the associations between DNA methylation and gene regulation. In *a*, CpG sites are methylated at the promoter region, but not in the gene body, which is associated with lack of gene expression. In *b*, the opposite is occurring.

1.2 Use of DNA methylation in a population setting

The importance of DNA methylation in disease has already been established in rare developmental disorders caused by aberrant imprinting patterns (32,33), but the relationship between DNA methylation and complex phenotypes is unclear and is the focus of this thesis. In this section I discuss the appeal of studying DNA methylation in this context, introduce the most common study design for doing so, EWAS, and overview some successes and complications of the work.

1.2.1 The appeal of DNA methylation to epidemiologists

All epigenetic modifications are of potential interest to those studying any phenotype. Arguably, epigenetics could underlie all phenotypic changes and be the difference between individuals who develop disease and those who don't (34). Further, epi-

genetic marks are modifiable, which means theoretically it would be possible to prevent or treat any disease by altering the epigenetic patterns of individuals (35). There are large practical issues with this, which are discussed in more detail in Problems for EWAS. However, even if targeting epigenetic marks is not easy, as long as it is possible to observe them they could be used as diagnostic biomarkers and predictors (16,34,36,37). Thus, the ability to measure, and the research in to understanding epigenetic mechanisms, could have broad consequences for public health.

A major difference between DNA methylation and other epigenetic marks, that was alluded to earlier, is that DNA methylation is far more stable. Enzymes do exist that can actively de-methylate the DNA, for example the ten-eleven translocation (TET) enzymes, but cell division or excision of the nucleotide is required for full de-methylation of a DNA molecule (38,39). Biologically, this suggests DNA methylation might be involved in long-term repression of gene expression, which is thought to be the case for X-inactivation (40), and practically it makes studying the epigenetic mark easier because stability ensures the marks are more resistant to changes after collection of samples. Also, even though it isn't clear that DNA methylation precedes gene expression regulation, the regulatory processes that govern whether genes are transcribed are linked. There are known examples of how DNA methylation tends to associate with other epigenetic marks, including positive correlation with the histone modification H3K9me3 (4) and histone deacetylation (4,41). This means measuring DNA methylation may capture regulatory information, even if addition or removal of methyl groups to the DNA would have little impact on gene expression. Recently, it's been shown that epigenetic marks can be used to predict each other with high accuracy (42), adding weight to the argument DNA methylation measurements capture far more than just DNA methylation itself.

1.2.2 Epigenome-wide association studies

EWAS are the most common study design for assessing the association between DNA methylation and a complex trait. They typically involve measuring hundreds of thousands of DNA methylation sites across the genome in a case-control or cohort setting and using linear models to assess the association between DNA methylation and the trait of interest. It's a study design that has been widely adopted over the past ten years and the relationship between a plethora of traits, from smoking to anthropometric measures to childhood adversities, and DNA methylation has now been studied (43–46). There are also large consortia that are pooling samples to gain power for these studies, for example the Pregnancy and Epigenetics Consortium (PACE) (47). A few traits have been identified as being associated with large variations in DNA methylation, one of which is smoking, where thousands of sites have been identified and many replicated in EWAS of smoking (43) and DNA methylation of a single locus can predict smoking status with high accuracy (48). It has also been revealed the direction of effect at some sites is likely from smoking to DNA methylation and in fact over time the DNA methylation changes caused by smoking may be (mostly) reversible by giving up smoking (49). Another complex trait shown to relate to large variation in DNA methylation across the genome is age, and like smoking, DNA methylation makes a highly accurate predictor for age (50–52). These studies have shown that large perturbations in the DNA methylome can be related to complex traits, highlighting the potential for EWAS for population health.

1.2.3 Problems for EWAS

Despite the promise of understanding the underlying biological processes related to traits, studying the relationship between DNA methylation and complex traits

provides many practical difficulties that often make the results of EWAS hard to interpret (16).

Confounding

One issue that plagues all observational epidemiology, including EWAS, is confounding. This is where the traits of interest share a common cause, and can bias effect estimates, hindering correct interpretation of the association between traits. Complex traits are strongly correlated with each other, often in clusters, which can lead to large amounts of measured and unmeasured confounding being present in EWAS. Of course, in order to produce therapies to prevent or treat disease by altering DNA methylation or other parts of the epigenome, causality must be established. Therefore, problems of confounding must be overcome in EWAS to use these results to start developing methods of targeting DNA methylation changes, this is discussed more in Establishing causality.

Cell type heterogeneity

As discussed, epigenetic factors guide differentiation of a single pluripotent cell to hundreds of cell types in human development. As these cell types can have large differences in morphology and function, it is clear that epigenetic marks, including DNA methylation, will vary between cell types (53,54). This poses two distinct problems for EWAS. Firstly, when collecting samples to measure DNA methylation, unless cells are purified, then a pool of cell types will be present in the samples, each with their own distinct DNA methylation patterns. This can lead to issues of confounding by cell type. For example, in a case control study, cases may be more likely to have increased numbers of CD4⁺ Th2 immune cells and these cells may on average have a higher level of DNA methylation at site X. In this scenario if

one were to take blood cells, measure DNA methylation, and assess the association between DNA methylation and the trait of interest, one might find an association between DNA methylation at site X and the trait, but this may just be function of the increased number of CD4+ Th2 cells present in cases and site X may have no causal relationship with the trait itself. There have been efforts to try and account for cell type heterogeneity in EWAS (53,55,56), but to completely prevent it's confounding effects, cells should be purified. The second problem arising from cell-type specific patterns of DNA methylation is the uncertainty that the cell type being studied is one in which DNA methylation covaries with the trait of interest. Non-invasive cells to collect, such as blood, skin, and saliva, are common amongst epidemiological studies, but it is unclear whether EWAS in these studies are relevant to a large proportion of complex traits. This is studied, with regards to blood, in **Chapter 5**. Studies have actually shown high levels of correlation between DNA methylation across cell types at many CpGs (57), but it is unknown whether the correlated sites are important to trait variation. Further, this correlation may complicate things with regards to epigenetic therapies. Some organs, such as the brain, are difficult to target with therapies so even if a promising epigenetic target is identified, it may be almost impossible to translate this into something clinically useful.

Measuring DNA methylation

Ideally, in every sample, DNA methylation would be measured across all sites in the genome. Unfortunately, this is not currently possible and sequencing technologies that offer something similar are often very expensive. There are three alternatives available to DNA methylation studies. Firstly, one could sequence a small portion of the genome if this section is of particular interest, however this candidate gene approach is unlikely to be profitable unless the genes targeted already have good

evidence for epigenetic variation with the trait of interest. Secondly, measuring DNA methylation on repeat sequences of the genome, such as long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs), can provide estimates for global DNA methylation changes (58). These measurements indicate if a trait is related to large perturbations of DNA methylation across the genome, but gives little mechanistic insight into what effects these changes may be having, as methylation at functional genes is not measured. Thirdly, an array approach covering DNA methylation genome-wide at selected sites. This last approach is the most common for population-based studies as it enables measurement of DNA methylation at hundreds of thousands of sites at a relatively cheap price per sample (59). The most commonly used array currently is the Illumina Infinium HumanMethylation450 (HM450) Beadchip, which measures DNA methylation at over 450,000 sites across the whole genome. The technology is always improving and now the Illumina Infinium HumanMethylationEPIC Beadchip, measuring over 800,000 sites is being adopted by various studies. However, these arrays only measure roughly 1.5-4% of all CpG sites in the genome and there is no empirical evidence to show that these sites capture a substantial proportion of complex trait variation. This will be examined in **Chapter 5**.

Batch effects and technical artefacts can also substantially bias results in EWAS (60). There has been considerable effort made to try and mitigate these effects and methods have been developed (some originally for use in RNA-based studies) to help correct for batch effects (61–63). Further, sequencing methods have been used to identify potentially unreliable probes (64,65). In **Chapter 4** I explore the extent to which batch effects tend to be removed in current EWAS and whether EWAS results are enriched for potentially faulty probes.

Complexity of regulatory mechanisms

EWAS identify single sites in the genome for which DNA methylation variation is associated with a trait of interest. As discussed, DNA methylation at a single site will likely be correlated with DNA methylation at neighbouring sites and other nearby epigenetic marks. This makes inferring mechanism of action very difficult. Differentially methylated region (DMR) analysis is often employed, which aims to determine if multiple neighbouring sites share an association with the trait of interest with the same direction of effect (26,27). These give evidence as to whether the sites covary similarly with the trait of interest, but don't provide evidence that the sites are acting independently or not. There are ways to circumvent the issues of biological complexity, but without additional gene expression data these often involve assuming the genes immediately adjacent to DNA methylation changes are of importance to the trait. However, no systematic evaluation of whether this assumption holds true for the majority of cases has been conducted.

Treatments

This is not a focus of my thesis, but for completion it is important to briefly note the complications of epigenetic-based therapies as they are an end goal of much epigenetic research. Currently there are therapies used in the clinic that target enzymes responsible for epigenetic alterations, for example DNA methyltransferase inhibitors and histone deacetylase inhibitors (66,67). They are primarily used to treat cancers, but as with many cancer treatments, are highly toxic. These therapies impact the epigenome globally and do not target any specific regions of the genome. This makes them highly undesirable for most diseases and as of yet there are no epigenetic therapies targeting specific regions of the genome. Methods, such as adapting CRISPR-cas9 enzymes, are being used in laboratories to alter DNA

methylation at specific sites (68), and some have even achieved *in vivo* targeted epigenetic modulation in mice (69). However, it is unclear whether these techniques can be scaled up for clinical use in humans and how long it may take to overcome the various complications.

In summary, there is great potential for EWAS to identify sites in the genome that could be targeted for treatment, but there are several challenges still to overcome. A great importance should be placed on using the data available to inform future designs of EWAS to maximise the potential of these studies.

1.3 Using methods from genetics to help inform future EWAS

The main commonality between EWAS and their counterpart in genetics, genome-wide association studies (GWAS), is the fact both assess the association between hundreds of thousands of markers genome-wide and complex traits. As germline genetic variants are fixed from conception, and these variants can act through pathways excluding DNA methylation changes, the genetic architecture of traits is likely vastly different to the epigenetic architecture. However, genetic epidemiologists have had to overcome various problems to help with the interpretation of GWAS, that are also present in EWAS. Further, genetic variants can be used as tools to augment the understanding of DNA methylation-trait associations. In this section I briefly describe some examples of these efforts and explain how they might be adapted to help inform future EWAS.

1.3.1 The GWAS Catalog

The GWAS Catalog is a manually curated database of publicly available GWAS data, developed by the European Bioinformatics Institute (EBI) and made openly available to the public (70). It has a broad range of applications for researchers, from replication of GWAS, to identifying overlapping GWAS signals between traits, to pooling the data to try and understand the genetic architecture of complex traits as a whole. Resources like this are invaluable to the genetic epidemiologist community and so developing a corollary database for EWAS may provide equal opportunity for epigenetic epidemiologists. Catalogues such as EWASdb (71) and the EWAS Atlas (72) are currently available, but fall short of some key researcher requirements including ease of use and access to full summary statistics. The development of a new database, The EWAS Catalog, is the focus of **Chapter 3**.

1.3.2 Total variance captured by all sites measured genome-wide

Before spending lots of time and resources on measuring the relationship between traits, it's important to have an understanding of the extent to which the exposure of interest covaries with the outcome of interest. Of course, if they are independent then studying the relationship between the two would be pointless. The phenotypic variance of a trait, σ_P^2 , is a function of the environmental variance, σ_E^2 , and genetic variance, σ_G^2 , that contribute to the trait:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 \quad (1.1)$$

σ_G^2 is mostly comprised of the variance due to additive genetic effects (σ_A^2), and the proportion of phenotypic variance accounted for by the additive genetic variance is

the narrow-sense heritability (h^2):

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \quad (1.2)$$

It is possible to estimate h^2 by taking advantage of pedigree information in families. This has been done for a plethora of traits to establish that they have some additive genetic component (i.e. $\sigma_A^2 > 0$ and $h^2 > 0$). This provides adequate evidence that genetic variation influences these traits and thus there are genetic effects to identify.

Initially, GWAS were conducted in hundreds or thousands of individuals and were identifying very few variants that could be said to reliably have an effect on the trait, and these sites explained an extremely small proportion of the heritability estimates (<1%) (73). Speculations were made about the reasons why this could be, for example the arrays were only capturing common genetic variation and it was rare genetic variation having the majority of the influence on phenotypes (74). This provided a need to derive individual aspects of h^2 to inform future study design. If indeed common genetic variation contributed little to h^2 for the majority of traits, the GWAS approach would need to be re-thought and technologies would need to be made to capture genetic variants not reliably captured by existing arrays (e.g. rare variants). Yang et al. developed a method to estimate the contribution of all the single nucleotide polymorphisms (SNPs) measured on SNP-arrays to phenotypic variance (SNP-heritability or h_{SNP}^2) (75). SNP-heritability was subsequently shown not to be inconsequential for complex traits and so warranted continuing use of GWAS and the acquisition of larger samples to conduct these GWAS (75).

Lack of associations and lack of phenotypic variance captured by DNA methylation sites has been observed in many EWAS, yet currently there is no corollary test to

SNP-heritability to enable increased understanding of why this may be. Of course, there could be a variety of reasons for lack of signal in EWAS, which include what was discussed earlier, e.g. measuring the wrong tissue. If there is covariation of DNA methylation with complex traits, it may just be that lots of associations with small effect sizes across the entire genome contribute to this covariation and so EWAS with large sample sizes are required to reliably identify these associations. In **Chapter 5**, I apply methods developed to assess SNP-heritability to estimate the proportion of phenotypic variance correlated with DNA methylation across a range of phenotypes.

1.3.3 Inferring biology from signals

As discussed, the complexity of cellular processes makes it difficult to infer what consequences a change in methylation at a specific site may have. Similarly, for the majority of SNPs identified in GWAS, the functional change that relates to an association between genetic variation at that site and the trait of interest is unclear. Traits themselves are the result of a large number of complex biological pathways that are determined by potentially thousands of gene products. It is often assumed that the signal from GWAS highlight genomic regions of importance to the trait and thus as a step to investigate the nature of the signal, sites identified are mapped to nearby genes. These genes can then be mapped to pathways and gene set enrichment analysis performed to assess whether the genes identified are present in any particular pathways more than expected by chance. This method has been adopted by epigenetic epidemiologists for use to examine EWAS signal (76).

Establishing causality from DNA methylation signal is difficult (Use of DNA methylation in a population setting, Establishing causality). Thus, when applying gene set enrichment analyses to identify prominent pathways in EWAS signals, the

pathways identified may actually be related to a particular confounder rather than the trait of interest. This is important to remember when interpreting the results of such an analysis, but it doesn't render the results useless. There is a huge body of work that characterises gene action and relationships of this gene action with various traits. By mapping EWAS signals to genes and pathways, a path between the trait (or a confounder) and changes in DNA methylation might become more clear. One example of this comes with EWAS of smoking, that have consistently identified DNA methylation at the *AHRR* gene (43,77–80). This gene is known to play a role in handling toxic substances found in tobacco smoke (81,82). Thus, large changes in DNA methylation related to this gene points towards epigenetic changes at that locus influencing the cellular response to smoking. This shows, that despite difficulties in interpreting EWAS findings and subsequent pathway analysis, EWAS can actually add to the pool of information about underlying trait biology when used in conjuncture with other evidence.

As DNA methylation changes may be a result of the trait (unlike genetic variation), pathway analysis for EWAS studies could capture biological consequences of the trait. This could allow for EWAS to identify facets of trait biology missed by GWAS studies, however it is unclear as to whether the analogous gene set enrichment design adopted by EWAS is adding to the information discovered by GWAS. In **Chapter 6** I compare overlap of GWAS and EWAS signals in this context and discuss the use of both study designs in elucidating underlying trait biology.

1.3.4 Establishing causality

Mendelian randomization

As discussed, population-based studies of DNA methylation suffer from the same limitations as any observational epidemiology study, namely confounding and

reverse causation. One method that aims to mitigate confounding is Mendelian randomization (MR) (83–85), which uses genetic variants as proxies for the exposure of interest in an instrumental variable framework (illustrated in **Figure 1.3**). Using genetic variants as instruments has the advantage that the direction of effect will always be from instrument to exposure and not *vice versa*, making interpretation of the studies simpler. Furthermore, unlike environmental phenotypes, that tend to be highly correlated and clustered into groups, genetic variants associated with a trait tend to be unconfounded (83,84). In the absence of assortative mating, genetic variants should be distributed randomly across the population, so in effect those grouped by genotype should exhibit differences in exposure, but confounding factors should not differ between genotype groups (84). Assortative mating has been reliably shown to occur with some traits such as social behaviours and anthropometric measures (86–88). There currently isn't evidence of assortative mating on DNA methylation variation and so is ignored for the purposes of this thesis.

Availability of data for MR

Another advantage of MR is the data it uses. Thousands of GWAS have been conducted giving researchers ample instruments for a wide variety of traits and many of these instruments are easily accessible through databases such as the GWAS Catalog (70) and IEU Open GWAS Project (89). Furthermore, it isn't necessary to use individual-level data to conduct MR studies; summary statistics from GWAS are all that is needed to provide data in a two-sample MR framework (90,91). This is especially valuable to conducting MR studies using DNA methylation data, as no cohorts have measured DNA methylation, genotype and a broad range of phenotypes in a large enough sample to provide adequate power to test DNA methylation-phenotype associations in a MR context.

Assumptions of MR

In order for MR analyses to be valid, they must satisfy three instrumental variable assumptions, these are illustrated in **Figure 1.3**. Testing assumption one, the instruments associate with the exposure of interest, is simple, but the other two assumptions can't technically be proven to be true. Horizontal pleiotropy, where genetic variants associate with more variables than just the exposure of interest, can lead to violations in assumptions two and three. Ideally, MR would be performed in the context where the genetic effect on the exposure had been characterised such that the mechanism of action was understood clearly. This would help give evidence against assumptions two and three being broken. Unfortunately, this is rarely possible. However, a plethora of methods have now been developed to test for pleiotropic effects, given the exposure of interest has multiple independent genetic variants reliably associated with it.

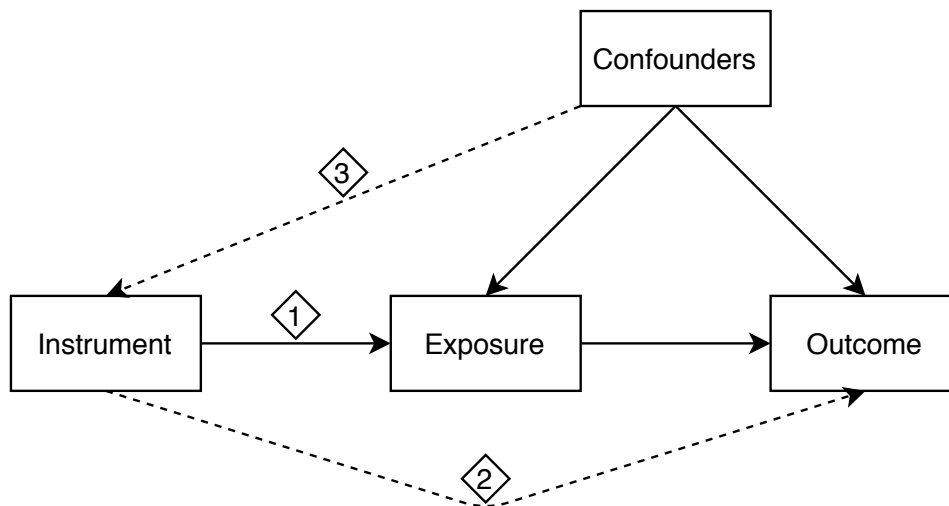


Figure 1.3: Mendelian randomization. Mendelian randomization can be used to test the causal nature of exposure-outcome relationships provided the assumptions are met. Assumption 1. There is an association between the instrument and the exposure. Assumption 2. There are no associations between the instrument and outcome, except through the exposure. Assumption 3. The instrument is not associated with any factors that confound the exposure-outcome relationship.

Applying MR in a DNA methylation context

MR can be applied to studies of DNA methylation by using methylation quantitative trait loci (mQTL), genetic variants associated with changes in DNA methylation levels, as proxies for DNA methylation variation (34,92,93). As mentioned previously, using a two-sample MR framework is especially useful to help increase power for these studies (34). Unfortunately, for each DNA methylation site few independent mQTLs have been identified. This prevents the use of various tests to examine whether the instruments are likely to be pleiotropic. Both cis-mQTLs (mQTLs within 1Mb of the DNA methylation site) and trans-mQTLs (mQTLs over 1Mb away from the DNA methylation site) have been identified in GWAS of DNA methylation variation. As genetic architecture of DNA methylation changes

is also not well understood, the mechanism of action for each mQTL can only be speculated at present. Cis-mQTLs are thought to be less likely to be pleiotropic than trans-mQTLs as the mechanism of action seems more likely to be related to the binding of regulatory machinery that may influence DNA methylation levels, for example a genetic variant may decrease the affinity of a transcription factor for that region and so the transcription factor will bind less frequently and/or for a shorter period, this would lead to increased methylation at that site. On the contrary, the mechanism of trans-mQTL action, especially those on separate chromosomes to the DNA methylation site of interest, is more likely to be pleiotropic, for example a trans-mQTL could influence gene expression of a transcription factor that binds many sites and alters their DNA methylation (94), this would make the trans-mQTL associate with multiple DNA methylation sites. Therefore, if one limits mQTLs to those in cis, this gives greater confidence that horizontal pleiotropy isn't biasing results.

As mentioned, DNA methylation varies both temporally and across cell types. If the instruments used to represent DNA methylation are viewed as influencing life-time differences in DNA methylation between individuals, then temporal variation can be ignored. However, cell type-specific effects are likely to occur for some mQTLs. One could imagine a scenario in which blood cells require a specific gene (gene A) to be expressed that is completely useless in adipose cells, which could mean all adipose cells have 100% DNA methylation at the promoter region of gene A. Thus, any genetic variants that associate with DNA methylation variation in blood would not also associate with variation in adipose cells.

With all this in mind, it's important to maintain the idea that making strong conclusions of causality in the context of DNA methylation is difficult, but triangulating evidence from multiple sources could be key to understanding the role of DNA

methylation in underlying trait biology (95).

1.4 Overview of thesis aims

DNA methylation has great potential for use in an epidemiological sense and as samples with DNA methylation data continue to grow it is important to understand the limitations of EWAS and how to maximise it's potential. My thesis aims to address this by exploring what information has been gained from EWAS (**Chapters 3 and 4**), what information is still to gain from EWAS (**Chapter 5**), whether EWAS might add to our biological understanding of complex traits above GWAS (**Chapter 6**) and by applying MR in a particular case, the potential for confounding in EWAS (**Chapter 7**). See the flowchart in **Figure 1.4** for a graphical depiction.

In **Chapter 3** a database of published EWAS is curated and made publicly available, which will be used in later chapters. The aim of **Chapter 4** is to analyse the vast database jointly to allow the discovery of commonalities across methylome-trait associations and provide a platform to explore what is driving these commonalities. Further, the chapter explores the extent to which published results are reliable by assessing replication rate and whether sites measured by unreliable probes are prominent.

After exploring the information already gained from EWAS, **Chapter 5** investigates the information still to gain from EWAS. The aim of the chapter is to apply methods developed to assess SNP-heritability to estimate the proportion of complex trait variation that is associated with sites commonly measured in EWAS.

Chapter 6 will then aim to assess whether the discoveries of EWAS may provide extra biological insight for traits of interest on top of those from GWAS. Tests will be applied to assess whether there is more overlap between the sites, genes

or pathways identified by some large EWAS ($N > 4500$) and their corresponding GWAS than expected by chance.

Finally, **Chapter 7** will apply MR to explore the causal nature of associations between DNA methylation and lung cancer. This application case-study will compare and contrast findings to conventional EWAS estimates to give an example of the potential residual confounding that can be present in EWAS.

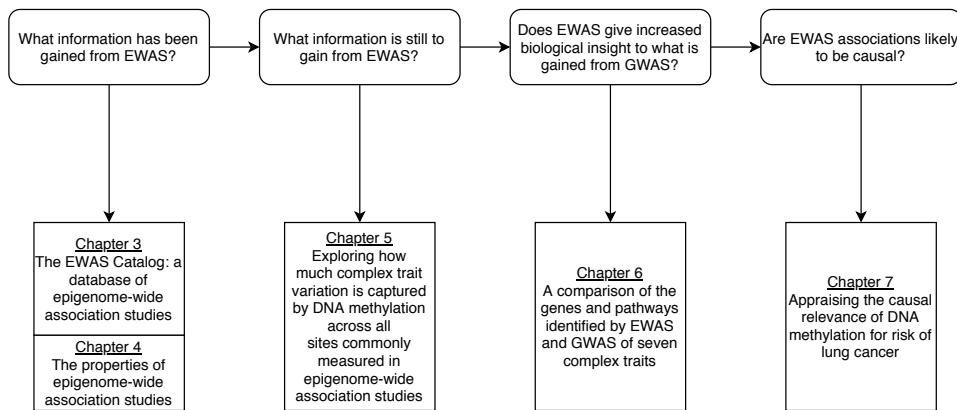


Figure 1.4: Flowchart showing questions being asked in the thesis and the corresponding chapters attempting to help answer them.

Chapter 2

Methods

Chapter 3

The EWAS Catalog: a database of epigenome-wide association studies

3.1 Abstract

Epigenome-wide association studies (EWAS) seek to understand the link between patterns of DNA methylation, the addition of a methyl group to a DNA molecule that may change how the molecule interacts with other cellular factors, at thousands or millions of sites across the genome to various traits and exposures. In recent years, the increase in availability of DNA methylation measures in population-based cohorts and case-control studies has resulted in a dramatic increase in the number of EWAS being performed and published. To make this rich source of molecular data more accessible, a manually curated database has been made containing CpG-trait associations (at $P < 1 \times 10^{-4}$) from published EWAS, each assaying over 100,000

CpGs in at least 100 individuals. The database currently contains these associations from over 150 published EWAS as well as full summary statistics for over 180 million association tests of 418 EWAS in the Avon Longitudinal Study of Parents and Children (ALSPAC) and the Gene Expression Omnibus (GEO). It is accompanied by a web-based tool and R package that allow these associations to be easily queried. This database will give researchers the opportunity to quickly and easily query EWAS associations to gain insight into the molecular underpinnings of disease as well as the impact of traits and exposures on the DNA methylome. The EWAS Catalog is available at: <http://www.ewascatalog.org>.

3.2 Introduction

Epigenome-wide association studies (EWAS) aim to assess the associations between phenotypes of interest and DNA methylation across the genome (34,96,97). These associations may then be used for disease diagnosis or prediction (34,96,97). Also, unlike genetic variants, changes in DNA methylation are responsive to the environment and so may be targeted for treatment. EWAS of smoking (43), body mass index (BMI) (45) and aging (50) have shown that various exposures are related to large perturbations in DNA methylation across the genome. Furthermore, a paper recently estimated that over 60% of the total proportion of BMI variation was captured by DNA methylation at about 150 CpG sites (98). In recent years, there has been a dramatic increase in the number of EWAS being performed and published due to technological advancements making it possible to measure DNA methylation at hundreds of thousands of CpG sites cheaply and effectively. Giving researchers easy access to EWAS outputs will help them gain insight into the molecular underpinnings of disease as well as the impact of traits and exposures on the DNA methylome. Furthermore, current collections of summary statistics have

already proven useful to various fields, for example the GWAS Catalog (70) has been cited over 2000 times in papers contributing to new methods and exploring the genetic architecture of a plethora of traits.

At the time of making the database, to our knowledge, there were no databases that had collected well-curated EWAS on all traits (no just diseases) in an online database accessible to researchers. During production one database fulfilled those metrics: EWAS Atlas (72). Other databases are available but are limited to certain diseases (e.g. MethHC (99)). The EWAS Atlas provides a simple-to-use website with annotated CpG sites and information on traits. Ideally a database of EWAS results will provide summary statistics, including betas, standard errors and p-values where provided from publications, in an easily accessible manner, this enables researchers to explore various aspects of the published data without having to retrieve the published article. For example, researchers might compare effect estimates between studies in the database or check to see if their results are replicated in another published study. At the time of writing the EWAS atlas platform did not enable users to download effect estimates and standard errors. Another caveat is that there is currently only published data on the platform, not full summary statistics from EWAS.

The EWAS Catalog aims to improve upon current databases to 1) allow easy and programmatic access to summary statistics for downstream analyses by researchers and 2) provide full summary statistics from a range of EWAS conducted in multiple cohorts. To this end we have produced The EWAS Catalog, a manually curated database of currently published EWAS, **NUMBER** (originally 378) EWAS performed in the Avon Longitudinal Study of Parents and Children (ALSPAC) (100,101) and **NUMBER** (originally 40) EWAS performed from data from the Gene Expression Omnibus (GEO) database. The process and data inclusion are

summarised in **FIGURE** (originally Supplementary Figure 1).

In this chapter, Dr James Staley built the original website, Dr Matthew Suderman has been key in development and maintenance of the website and there was a team to help gather and input the data. I helped develop and maintain the website, gather and input the data, run the EWAS within the ALSPAC cohort and on data from the GEO database. The team, led by myself, is continuing to develop and maintain the database. Full acknowledgements to the team can be found on the website: <http://www.ewascatalog.org/about/>.

3.3 Methods

3.3.1 Implementation

The EWAS Catalog web app was built using the Django Python package (<https://djangoproject.com>). The data is stored in a combination of MySQL databases and fast random access files (102) and can be queried via the web app or the R package (www.github.com/ewascatalog/ewascatalog-r/).

3.3.2 Overview of publication data extraction

To identify publications, periodic literature searches are performed in PubMed using the search terms: “epigenome-wide” OR “epigenome wide” OR “EWAS” OR “genome-wide AND methylation” OR “genome wide AND methylation”.

Our criteria for inclusion of a study into The EWAS Catalog are as follows:

1. The EWAS performed must contain over 100 humans
2. The analysis must contain over 100,000 CpG sites
3. The DNA methylation data must be genome-wide

4. The study must include previously unpublished EWAS summary statistics

CpG-phenotype associations are extracted from studies at $P < 1 \times 10^{-4}$. Variables extracted can be found in **TABLE** (NO ORIGINAL). All these criteria along with the variables extracted are documented on the website (www.ewascatalog.org/documentation). Experimental factor ontology (EFO) terms were mapped to traits to unify representation of these traits. These EFO terms were manually entered after looking up the trait in the European Bioinformatics Institute database (www.ebi.ac.uk/efo).

Based on these criteria, from 3rd July 2019, The EWAS Catalog contained 540,699 associations from 159 studies.

3.3.3 Overview of GEO data extraction

To recruit additional datasets suitable for new EWAS analysis, the *geograbi* R package (<https://github.com/yousefi138/geograbi>) was used to both query GEO for experiments matching The EWAS Catalog inclusion criteria (described above) and extract relevant DNA methylation and phenotype information. The query was performed on 20 March 2019 and identified 148 such experiments with 32,845 samples where DNA methylation and phenotype information could be successfully extracted. From these, the aim was to repeat the analyses performed in the publications linked by PubMed IDs to each GEO record. Thus, I looked up the corresponding full texts for each dataset and identified the main variables of interest. Of the 148 putative GEO studies, only 34 (23%) contained sufficient information to replicate the original analysis.

3.3.4 EWAS methods

Avon Longitudinal Study of Parents and Children (ALSPAC)

EWAS were conducted for **NUMBER** (originally 378) continuous and binary traits in peripheral blood DNA methylation of ALSPAC mothers in middle age (N = **NUMBER** (originally 940)), generated as part of the Accessible Resource for Integrated Epigenomics Studies (ARIES) project (103). The traits were extracted from the same time that blood was drawn for DNA methylation assays.

ADD IN QC STEPS AND COHORT INFO HERE (originally in Supplementary Material)

For all traits, linear regression models were fitted with DNA methylation at each site as the outcome and the phenotype as the exposure. DNA methylation was coded as beta values between 0 and 1. For a particular site, a beta value of 0 represents no methylation being detected in all cells measured and a value of 1 represents all cells being methylated at that site. Covariates included age, the top 10 ancestry principal components, and 20 surrogate variables.

ADD IN HOW PCS

GEO datasets

EWAS were performed using 30 datasets, containing 36 traits were extracted from GEO using the `geograbi` R package (<https://github.com/yousefi138/geograbi>).

ADD IN QC STEPS HERE (originally in Supplementary Material)

A list of all the traits with corresponding citations is provided in **TABLE** (originally Supplementary Table 1). For all traits, linear regression models were fitted with DNA

methylation as the outcome and the phenotype as the exposure as for the ARIES data. Twenty surrogate variables were included as covariates. Other covariates were considered, but surrogate variables only were used for two reasons: 1) to help automate the process and 2) because covariates used in the original EWAS were not included with many of the GEO datasets.

Statistical analyses were conducted in R (Version 3.3.3). The `smartsva` package (104) was used to create surrogate variables and the `ewaff` R package (<https://github.com/perishky/ewaff>) was used to conduct the EWAS, all p-values are two-sided.

3.4 Results

3.4.1 Database interface and use

There are two ways to access this large, curated database: through the main website www.ewascatalog.org or by using the R package “`ewascatalog`”. The website provides a simple user interface, which resembles that of the GWAS catalog (70), whereby there is a single search bar to explore the database and links to tabs that contain documentation on the contents and how to cite its use (Figure 1). Users may enter a CpG, gene, genome position or trait into the search bar and it will rapidly return detail for relevant EWAS associations, including CpG, trait, sample size, publication and association (effect and P value) (Figure 1). This information along with additional information such as ancestry, outcome, exposure units, and tissue analysed are available for download as a tab-separated value (tsv) text file. Unlike other EWAS databases, we provide the option of downloading summary results for both the user’s search and for the entire database.

The screenshot displays the EWAS Catalog homepage with a red header. Below the header, there are three search sections: 'CpG sites', 'Genes', and 'Traits'. Each section has a search input field and a 'Search' button. The 'Traits' section shows a search for 'Depression'. Below the search sections is a 'Results' button. The results section shows a table with 3 entries. At the bottom left, a 'Download' button is circled in red.

EWAS Catalog
The MRC-IEU catalog of epigenome-wide association studies

Search
Enter CpG, gene, region or trait
Examples: cg00029284, chr12:111731203, FTO, 6:15000000-25000000, body mass index.

Search CpG sites
cg00029284

Search Genes
FTO

Search Traits
Depression

Results

Show 10 entries Search:

Author	PMID	Outcome	Exposure	Analysis	N	CpG	Location	Gene	Beta	P
Bostrom AE	28618313	DNA methylation	Depression	-	93	cg13227621	chr1:45671361	ZSWIM5	-0.22	8.9E-09
Bostrom AE	28618313	DNA methylation	Depression	-	93	cg04102384	chr6:31668622	BAT5	-0.59	1.1E-08
Weder N	24655651	Depression	DNA methylation	-	190	cg03535461	chr1:23885568	ID3	NA	5.4E-08

Showing 1 to 3 of 3 entries Previous 1 Next

Download

Figure 3.1: Using the EWAS catalog. At the top of the figures is the home page URL, ewascatalog.org. Below that are examples of three types of searches possible: 1. CpG sites, 2. genes and 3. traits. Finally, the results are displayed after searching the catalog for “Depression”. Circled in red is the download button, this button enables the user to download the results of their search as a tab-separated value file. This file will contain the information shown on the website as well as additional analysis information.

The R package, along with installation instructions and examples are available at <https://github.com/ewascatalog/ewascatalog-r/>. Once installed, the database can be queried directly in R using the “ewascatalog()” function similar to the website: simply supply the function with a CpG site, gene, genome position or trait and the function returns the same output as is downloadable from the website.

3.5 Discussion

In this chapter, a database of previously published EWAS and the full summary statistics of 418 newly performed EWAS within ALSPAC and GEO has been established. This is freely available for all researchers to use and provides a platform to explore what information has been gained from EWAS as well as a platform that can be used to pool all existing data to gain new insights into both the EWAS study itself and how DNA methylation associates with traits. Despite the fact The EWAS Atlas has similar aims to The EWAS Catalog, latter provides full summary statistics, extra information and a user-friendly platform to enable more downstream analyses.

The EWAS catalog team will continue to collate and upload newly published EWAS and further increase the number of full summary statistics on the website by performing additional EWAS on available datasets and by inviting EWAS authors to provide full summary statistics. Currently work is ongoing to include additional functionality to allow users to easily and systematically compare their EWAS findings to EWAS in the database. With this full summary data, it is possible to make greater strides into discovering the epigenetic architecture of traits.

Therefore, despite the fact no extra information about EWAS was presented in

this chapter, a platform has been made that easily enables us to explore 1) what information has been gained from EWAS and 2) the relationship between DNA methylation and all traits. This will be explored in the next chapters.

Chapter 4

Properties of EWAS

Here is a reference to Caroline's paper: (34)

Chapter 5

h2ewas

Chapter 6

EWAS-GWAS comparison

Chapter 7

Appraising the causal relevance of DNA methylation for risk of lung cancer

7.1 Abstract

7.1.1 Background

DNA methylation changes in peripheral blood have recently been identified in relation to lung cancer risk. Some of these changes have been suggested to mediate part of the effect of smoking on lung cancer. However, limitations with conventional mediation analyses mean that the causal nature of these methylation changes has yet to be fully elucidated.

7.1.2 Methods

We first performed a meta-analysis of four epigenome-wide association studies (EWAS) of lung cancer (918 cases, 918 controls). Next, we conducted a two-sample Mendelian randomization analysis, using genetic instruments for methylation at CpG sites identified in the EWAS meta-analysis, and 29,863 cases and 55,586 controls from the TRICL-ILCCO lung cancer consortium, to appraise the possible causal role of methylation at these sites on lung cancer.

7.1.3 Results

16 CpG sites were identified from the EWAS meta-analysis ($FDR < 0.05$), 14 of which we could identify genetic instruments for. Mendelian randomization provided little evidence that DNA methylation in peripheral blood at the 14 CpG sites play a causal role in lung cancer development ($FDR > 0.05$), including for cg05575921-*AHRR* where methylation is strongly associated with both smoke exposure and lung cancer risk.

7.1.4 Conclusions

The results contrast with previous observational and mediation analysis, which have made strong claims regarding the causal role of DNA methylation. Thus, previous suggestions of a mediating role of methylation at sites identified in peripheral blood, such as cg05575921-*AHRR*, could be unfounded. However, this study does not preclude the possibility that differential DNA methylation at other sites is causally involved in lung cancer development, especially within lung tissue.

7.2 Introduction

Lung cancer is the most common cause of cancer-related death worldwide (105). Several DNA methylation changes have been recently identified in relation to lung cancer risk (106–108). Given the plasticity of epigenetic markers, any DNA methylation changes that are causally linked to lung cancer are potentially appealing targets for intervention (109,110). However, these epigenetic markers are sensitive to reverse causation, being affected by cancer processes (110), and are also prone to confounding, for example by socio-economic and lifestyle factors [(111); Elliott2014].

One CpG site, cg05575921 within the aryl hydrocarbon receptor repressor (*AHRR*) gene, has been consistently replicated in relation to both smoking (43) and lung cancer (80,106,107) and functional evidence suggests that this region could be causally involved in lung cancer (112). However, the observed association between methylation and lung cancer might simply reflect separate effects of smoking on lung cancer and DNA methylation, i.e. the association may be a result of confounding (85), including residual confounding after adjustment for self-reported smoking behaviour (113,114). Furthermore, recent epigenome-wide association studies (EWAS) for lung cancer have revealed additional CpG sites which may be causally implicated in development of the disease (106,107).

Mendelian randomization (MR) uses genetic variants associated with modifiable factors as instruments to infer causality between the modifiable factor and outcome, overcoming most unmeasured or residual confounding and reverse causation (83,84). In order to infer causality, three core assumptions of MR should be met: 1) The instrument is associated with the exposure, 2) The instrument is not associated with any confounders, 3) The instrument is associated with the outcome only through the

exposure. MR may be adapted to the setting of DNA methylation (92,103,115) with the use of single nucleotide polymorphisms (SNPs) that correlate with methylation of CpG sites, known as methylation quantitative trait loci (mQTLs) (116).

In this study, we performed a meta-analysis of four lung cancer EWAS (918 case-control pairs) from prospective cohort studies to identify CpG sites associated with lung cancer risk and applied MR to investigate whether the observed DNA methylation changes at these sites are causally linked to lung cancer.

In this chapter, Dr Rebecca Richmond performed analysis in the CCHS cohort (see section Potential causal effect of *AHRR* methylation on lung cancer risk: one sample MR of the results) and contributed to writing the introduction and discussion. I completed all the analysis and wrote each section except the methods and results sections for which I did not complete the analysis.

7.3 Methods

7.3.1 EWAS study details

A meta-analysis of four lung cancer case-control EWAS was conducted to identify DNA methylation sites associated with lung cancer. DNA methylation in each EWAS was assessed using the Illumina Infinium® HumanMethylation450 Bead-Chip. All EWAS are nested within prospective cohorts that measured DNA methylation in peripheral blood samples before diagnosis: EPIC-Italy (185 case-control pairs), Melbourne Collaborative Cohort Study (MCCS) (367 case-control pairs), Norwegian Women and Cancer (NOWAC) (132 case-control pairs) and the Northern Sweden Health and Disease Study (NSHDS) (234 case-control pairs). Study populations, laboratory methods, data pre-processing and quality control methods have been described in detail elsewhere (107) 3 and are outlined below.

At the various laboratory sites, samples were distributed into 96-well plates and processed in chips of 12 arrays (8 chips per plate) with case-control pairs arranged randomly on the same chip. Methylation data were pre-processed and normalized in each study, and probe filtering was performed as previously described (107), leaving 465,886 CpGs suitable for the analysis in EPIC-Italy, 485,330 CpGs in MCCS, 450,890 CpGs in NOWAC and 482,867 CpGs in NSHDS.

European Prospective Investigation into Cancer and Nutrition-Italy (EPIC-Italy)

EPIC-Italy includes 47,749 volunteers (32,579 women) aged 35–70 years at the time of recruitment (1992–1998). Anthropometric measurements and lifestyle variables including detailed information on smoking history were collected at recruitment through standardized questionnaires, together with a blood sample. Within EPIC-Italy we conducted a nested case-control study utilizing incident cases diagnosed within follow-up and healthy controls individually matched to cases by gender, date of birth (± 5 years), date of inclusion in the study and study centre. Analysis was performed for 185 incident cases diagnosed within follow-up and matched controls. Laboratory procedures were carried out at the Human Genetics Foundation (Turin, Italy) and DNA extracted from buffy coats as previously described (107). All participants signed an informed consent form, and the ethical review boards of the International Agency for Research on Cancer and of each local participating centre approved the study protocol.

Melbourne Collaborative Cohort Study (MCCS)

The MCCS is a prospective cohort study of 41,514 volunteers (24,469 women) aged between 27 and 76 years at baseline (1990-1994). At baseline attendance,

participants completed questionnaires that measured demographic characteristics and lifestyle factors. Height and weight were directly measured, and a blood sample was collected and stored. Incident cases of lung cancer were identified through linkage with the State and National Cancer Registries during follow-up up to the end of 2011. The MCCS sample included 367 cases and 367 matched controls selected from MCCS participants who were lung cancer free at the age of diagnosis of the matching case (density sampling). Matching variables included gender, date of blood collection (within 6 months), date of birth (within 1 year), country of birth (Australia and UK versus Southern Europe), type of biospecimen (lymphocyte, buffy coat and dried blood spot) and smoking status (never smokers; short-term former smokers: quitting smoking less than 10 years before blood draw; long-term former smokers: quitting smoking 10 years or more before blood draw; current light smokers: less than 15 cigarettes per day at blood draw; and current heavy smokers: 15 cigarettes or more at blood draw). For the MCCS, laboratory procedures were carried out at the Genetic Epidemiology Laboratory, the University of Melbourne according to manufacturers' protocols. DNA extraction from lymphocytes and buffy coats was performed as previously described (107). The Cancer Council Victoria's Human Research Ethics Committee approved the study protocol. Subjects gave written consent to participate and for the investigators to obtain access to their medical records.

Norwegian Women and Cancer (NOWAC)

The biobank of the NOWAC cohort was established in the years 2003-2006. Those who filled in an eight-page questionnaire and accepted the invitation to donate blood were sent blood drawing equipment together with a two-page epidemiological questionnaire. Around 50 000 women returned two tubes of blood to the Institute of

Community Medicine at UiT The Arctic University of Norway and data linkage to the National Cancer Registry of Norway was performed. During follow-up to the end of 2011, 132 eligible cases of lung cancer were identified and were used for the EWAS. For each case, one control with an available blood sample was selected and matched on time since blood sampling and year of birth in order to control for effects of storage time and ageing. The cases and the controls were processed together for all laboratory procedures in order to reduce any batch effect. Laboratory procedures were carried out at the Human Genetics Foundation (Turin, Italy). DNA extraction from buffy coats was performed as previously described (107). All participants gave informed consent. The study was approved by the Regional Committee for Medical and Health Research Ethics in North Norway. Data storage and linkage was approved by the Norwegian Data Inspectorate.

Northern Sweden Health and Disease Study (NSHDS)

NSHDS is an ongoing prospective cohort and intervention study intended for health promotion of the population of Västerbotten County in northern Sweden. All residents were invited to participate by attending a health check-up at their local health care centre at 40, 50 and 60 years of age. At the health check-up, participants were asked to complete a self-administered questionnaire covering various factors such as education, smoking habits, physical activity and diet. In addition, height and weight were measured and participants were asked to donate a blood sample. Incident lung cancer cases were identified through linkage to the regional cancer registry. One control was chosen at random for each lung cancer case from appropriate risk sets consisting of all cohort members alive and free of cancer (except non-melanoma skin cancer) at the time of diagnosis of the index case. Matching criteria were the same as for the MCCS except there was

no matching for type of biospecimens as DNA was extracted from whole blood for all samples. After quality control, a total of 234 incident lung cancer cases and 234 individually matched controls were available for this analysis. Laboratory procedures for NSHDS were carried out at two sites. DNA extraction from the buffy coat was conducted at Umeå University, Sweden, as previously described. Illumina Infinium HumanMethylation450 BeadChip analysis was conducted at the ALSPAC/IEU Laboratory at the University of Bristol. All study subjects provided written informed consent at time of the recruitment into the NSHDS.

7.3.2 EWAS Meta-analysis

To quantify the association between the methylation level at each CpG and the risk of lung cancer conditional logistic regression models were fitted for beta values of methylation (which ranges from 0 (no cytosines methylated) to 1 (all cytosines methylated)) on lung cancer status for the four studies. The cases and controls in each study were matched, details can be found above. Surrogate variables were computed in the four studies using the SVA R package (117) and the proportion of CD8+ and CD4+ T cells, B cells, monocytes, natural killer cells and granulocytes within whole blood were derived from DNA methylation (53). The following EWAS models were included in the meta-analysis: Model 1 – unadjusted; Model 2 – adjusted for 10 surrogate variables (SVs); Model 3 – adjusted for 10 SVs and derived cell proportions. EWAS stratified by smoking status was also conducted (never (N=304), former (N=648) and current smoking (N=857)). For Model 1 and Model 2, the case-control studies not matched on smoking status (EPIC-Italy and NOWAC) were adjusted for smoking.

An inverse-variance weighted fixed effects meta-analysis was performed of the EWAS (918 case-control pairs) using the METAL software. Direction of effect,

effect estimates and the I^2 statistic were used to assess heterogeneity across the studies in addition to effect estimates across smoking strata (never, former and current). All sites identified at a false discovery rate (FDR) < 0.05 in Model 2 and 3 were also present in the sites identified in Model 1. The effect size differences between models for all sites identified in Model 1 were assessed by a Kruskal-Wallis test and a post-hoc Dunn's test. There was little evidence for a difference ($P > 0.1$), so to maximize inclusion into the MR analyses we took forward the sites identified in the unadjusted model (Model 1).

7.3.3 Mendelian randomization

Two-sample MR was used to establish potential causal effect of differential methylation on lung cancer risk (90,91).

Sample 1: Accessible Resource for Integrated Epigenomic Studies (ARIES)

In the first sample, mQTL-methylation effect estimates (β_{GP}) for each CpG site of interest were identified in an mQTL database from the Accessible Resource for Integrated Epigenomic Studies (ARIES) (<http://www.mqtlldb.org>). Details on the methylation pre-processing, genotyping and quality control (QC) pipelines are outlined below.

DNA methylation data Samples were drawn from the Avon Longitudinal Study of Parents and Children [(101); Fraser2013]. Blood from 1,018 mother-child pairs were selected for analysis as part of the Accessible Resource for Integrative Epigenomic Studies (ARIES, <http://www.ariesepigenomics.org.uk/>) (103). There are three timepoints in children and two in their mothers, the timepoints with mean ages (in brackets) in ARIES are as follows for children: birth,

childhood (7.5), adolescence (17.1) and for mothers: during pregnancy (28.7), and at middle age (46.9). Following DNA extraction, samples were bisulphite converted using the Zymo EZ DNA Methylation™ kit (Zymo, Irvine, CA, USA). Following conversion, genome-wide methylation was measured using the Illumina Infinium HumanMethylation450 (HM450) BeadChip. Methylation data were normalised in R with the watermelon package (118) using the Touleimat and Tost (119) algorithm to reduce the non-biological differences between probes. Methylation data in ARIES were rank-normalised to remove outliers, and then Matrix eQTL software (116) was used to perform preliminary association analysis of SNPs with all CpG sites in the Illumina Infinium HM450 array with the exception of those failing QC, and those reported to map to more than one location (n=19,834) or to contain a genetic variant at the CpG site (n=74,182) (65).

Genetic data Children were genotyped using the Illumina HumanHap550 quad genome-wide SNP genotyping platform (Illumina Inc., San Diego, USA) by the Wellcome Trust Sanger Institute (WTSI, Cambridge, UK) and the Laboratory Corporation of America (LCA, Burlington, NC, USA). Individuals were excluded on the basis of incorrect gender assignment; abnormal heterozygosity (<0.320 or >0.345 for WTSI data; <0.310 or >0.330 for LCA data); high missingness (>3%); cryptic relatedness (>10% identity by descent) and non-European ancestry (detected by multidimensional scaling analysis). Following QC the final directly genotyped dataset contained 500,527 SNP loci.

Mothers were genotyped using the Illumina Human660W-quad genome-wide SNP genotyping platform (Illumina Inc., San Diego, USA) at the Centre National de Génomage (CNG, Paris, France). Individuals were excluded based on non-European ancestry, missingness, relatedness, gender mismatches and heterozygosity. PLINK (v1.07) (120) was used to carry out quality control measures on an initial

set of 10,015 subjects and 557,124 directly genotyped SNPs. Following QC the final directly genotyped dataset contained 526,688 SNP loci.

Imputation was performed to increase the SNP density for all genotyped mothers and children combined. Genotypes were phased together using ShapeIt, and then imputed against the 1000 genomes reference panel (phase 1 version 3, phased using ShapeIt version 2, December 2013, using all populations) using Impute (version 2.2.2). Genotypes were filtered to have Hardy-Weinberg equilibrium $P > 5 \times 10^{-7}$, MAF $> 1\%$ and imputation info score > 0.8 . Best guess genotypes were used for subsequent analysis. The final imputed dataset used for the analyses presented here contained 8,074,398 loci.

Written informed consent has been obtained from all ALSPAC participants. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and Local Research Ethics Committees. Please note that the study website contains details of all the data that is available through a fully searchable data dictionary: <http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/>

7.3.4 Sample 2: Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium

In the second sample, summary data was extracted from a GWAS meta-analysis of lung cancer risk conducted by the Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium (TRICL-ILCCO) (29,863 cases, 55,586 controls) for individuals genotyped using the Illumina Infinium OncoArray-500K BeadChip (Illumina Inc. San Diego, CA) and independent samples for which prior genotyping was performed (121). They were used to obtain mQTL-lung cancer estimates (β_{GD}).

For each independent mQTL ($r^2 < 0.01$), we calculated the log odds ratio (OR) per SD unit increase in methylation by the formula $\frac{\beta_{GD}}{\beta_{GP}}$ (Wald ratio). Standard errors were approximated by the delta method (122). Where multiple independent mQTLs were available for one CpG site, these were combined in a fixed effects meta-analysis after weighting each ratio estimate by the inverse variance of their associations with the outcome. Heterogeneity in Wald ratios across mQTLs was estimated using Cochran's Q test, which can be used to indicate horizontal pleiotropy (123). Differences between the observational and MR estimates were assessed using a Z-test for difference.

If there was evidence for an mQTL-CpG site association in ARIES in at least one time-point, it was assessed whether the mQTL replicated across time points in ARIES (FDR < 0.05, same direction of effect). Further, this association was re-analysed using linear regression of methylation on each genotyped SNP available in an independent cohort (NSHDS), using rvtests (124). The same NSHDS samples on which DNA methylation was measured were genotyped using the Illumina Infinium OncoArray-500k BeadChip (Illumina Inc. San Diego, CA) and quality control parameters were applied under the recently published TRICL-ILCCO GWAS study on lung cancer (121). Genetic imputation was performed on these samples using the Haplotype Reference Consortium (HRC) Panel (release 1) (108) through the Michigan Imputation Server (125). Replicated mQTLs were included where possible to reduce the effect of winner's curse using effect estimates from ARIES. We assessed the instrument strength of the mQTLs by investigating the variance explained in methylation by each mQTL (r^2) as well as the F-statistic in ARIES **Supplementary Table 1**. The power to detect the observational effect estimates in the two-sample MR analysis was assessed a priori, based on an alpha of 0.05, sample size of 29,863 cases and 55,586 controls (from TRICL-ILCCO) and calculated

variance explained (r^2).

MR analyses were also performed to investigate the impact of methylation on lung cancer subtypes in TRICL-ILCCO: adenocarcinoma (11,245 cases, 54,619 controls), small cell carcinoma (2791 cases, 20,580 controls), and squamous cell carcinoma (7704 cases, 54,763 controls). We also assessed the association in never smokers (2303 cases, 6995 controls) and ever smokers (23,848 cases, 16,605 controls) (121) 25. Differences between the smoking subgroups were assessed using a Z-test for difference.

We next investigated the extent to which the mQTLs at cancer-related CpGs were associated with four smoking behaviour traits which could confound the methylation-lung cancer association: number of cigarettes per day, smoking cessation rate, smoking initiation and age of smoking initiation using GWAS data from the Tobacco and Genetics (TAG) consortium (N=74,053) (126) 29.

7.3.5 Supplementary analyses

Assessing the potential causal effect of *AHRR* methylation: one sample MR

Given previous findings implicating methylation at *AHRR* in relation to lung cancer (106,107) 2, 3, we performed a one-sample MR analysis (127) 30 of *AHRR* methylation on lung cancer incidence using individual-level data from the Copenhagen City Heart Study (CCHS) (357 incident cases, 8401 remaining free of lung cancer). Copenhagen City Heart Study is a prospective study of the general population (128). Copenhagen residents were invited to complete a questionnaire and undergo a physical examination and are followed through a unique person identifier in the Danish health registries. All participants gave written informed consent, and a

Danish ethics committee approved the study (KF100.2039/91).

Phenotypic data Participants were asked whether they smoked at the day of attendance or previously. If they answered affirmative to either of these questions, they were asked about their current and former smoking behaviour, including age of smoking initiation, age of smoking cessation, and number of daily consumed cigarettes, cheroots, cigars, and weekly grams of pipe tobacco. Based on these answers, participants were categorized as never, former, and current smokers. In addition, participants reported on alcohol consumption, occupational exposure to dust and/or welding fumes, exposure to passive smoking, education, and familial cases of lung cancer. The answers were reviewed together with an examiner at the day of attendance. Body mass index was calculated as measured weight in kilograms divided by measured height (in meters) squared.

Methylation data At the physical examination, blood samples were drawn for DNA from which *AHRR* methylation extent was measured (80). The *AHRR* cg05575921 methylation extent was measured in duplicate samples of bisulphite treated DNA from peripheral blood from 9,234 individuals. We used a Taqman assay developed in our own laboratory, and included standard curves, as well as internal controls in each 384-well plate. Coefficients of variation at the methylation level of 71% varied from 5.0 to 6.7%. Laboratory technicians were blinded to smoking and disease status of the individuals. Results were validated with pyrosequencing on a subset of samples.

Genetic data Genotypes from the iCOGs array (129) and prospective data on lung cancer incidence were also available for these participants. Of the 9234 individuals, genotype data from iCOGS on 8778 were available. In short, DNA isolated from leukocytes was genotyped with a custom Illumina iSelect genotyping array, designed to test genetic variants related to breast, ovary and prostate cancer,

comprising roughly 211,000 SNPs after rigorous quality control (129).

Lung cancer data For lung cancer (ICD7, codes 1624 or 4624 until 1977, and ICD10, code C34 from 1978 and onwards), the date of first diagnosis was taken from the national Danish Cancer Registry from 1943 to December 2012.

Identification of mQTLs for CCHS one-sample MR mQTLs located within 1Mb of cg05575921 *AHRR* were identified in ARIES (FDR<0.05). Of those mQTLs which replicated within the CCHS, we performed an LD pruning step using a less stringent r^2 threshold of 0.2 and generated an unweighted allele score, calculated by coding and then summing the alleles to reflect the average number of methylation-increasing alleles carried by an individual. Associations between the allele score and several potential confounding factors (sex, alcohol consumption, smoking status, occupational exposure to dust and/or welding fumes, passive smoking) were investigated. Then MR analyses were performed using two-stage Cox regression, with adjustment for age and sex, and further stratified by smoking status.

Tumour and adjacent normal methylation patterns

DNA methylation data from lung cancer tissue and matched normal adjacent tissue (N=40 squamous cell carcinoma and N=29 adenocarcinoma), profiled as part of The Cancer Genome Atlas (TCGA), were used to assess tissue-specific DNA methylation changes across sites identified in the meta-analysis of EWAS, as outlined previously (130) 31.

mQTL association with gene expression

For the genes annotated to CpG sites identified in the lung cancer EWAS, we examined gene expression in whole blood and lung tissue using data from the gene-tissue expression (GTEx) consortium (131) 32.

Analyses were conducted in Stata (version 14) and R (version 3.2.2). For the two-sample MR analysis we used the MR-Base R package TwoSampleMR (132) 33. An adjusted P value that limited the FDR was calculated using the Benjamini-Hochberg method (133) 34. All statistical tests were two-sided.

7.4 Results

A flowchart representing our study design along with a summary of our results at each step is displayed in Figure 1.

(Figure 1 here)

7.4.1 EWAS meta-analysis

The basic meta-analysis adjusted for study-specific covariates identified 16 CpG sites which were hypomethylated in relation to lung cancer ($FDR < 0.05$, Model 1, Figure 2). Adjusting for 10 surrogate variables (Model 2) and derived cell counts (Model 3) gave similar results (Table 1). The direction of effect at the 16 sites did not vary between studies (median $I^2 = 38.6$) (Supplementary Table 2), but there was evidence for heterogeneity of effect estimates at some sites when stratifying individuals by smoking status (Table 1).

(Figure 2 here)

(Table 1 here)

7.4.2 Mendelian randomization

We identified 15 independent mQTLs ($r^2 < 0.01$) associated with methylation at 14 of 16 CpGs. Ten mQTLs replicated at $FDR < 0.05$ in NSHDS (Supplementary Table

3). MR power analyses indicated >99% power to detect ORs for lung cancer of the same magnitude as those in the meta-analysis of EWAS.

There was little evidence for an effect of methylation at these 14 sites on lung cancer ($FDR > 0.05$, Supplementary Table 4). For nine of 14 CpG sites the point estimates from the MR analysis were in the same direction as in the EWAS, but of a much smaller magnitude (Z-test for difference, $P < 0.001$) (Figure 3).

For 9 of out the 16 mQTL-CpG associations, there was strong replication across time points (Supplementary Table 5) and 10 out of 16 mQTL-CpG associations replicated at $FDR < 0.05$ in an independent adult cohort (NSHDS). Using mQTL effect estimates from NSHDS for the 10 CpG sites that replicated ($FDR < 0.05$), findings were consistent with limited evidence for a causal effect of peripheral blood-derived DNA methylation on lung cancer (Supplementary Figure 1).

(Figure 3 here)

There was little evidence of different effect estimates between ever and never smokers at individual CpG sites (Supplementary Figure 2, Z-test for difference, $P > 0.5$). There was some evidence for a possible effect of methylation at cg21566642-ALPPL2 and cg23771366-PRSS23 on squamous cell lung cancer ($OR = 0.85$ [95% confidence interval (CI) = 0.75, 0.97] and 0.91 [95% CI = 0.84, 1.00] per SD [14.4% and 5.8%] increase, respectively) as well as methylation at cg23387569-AGAP2, cg16823042-AGAP2, and cg01901332-ARRB1 on lung adenocarcinoma ($OR = 0.86$ [95% CI = 0.77, 0.96], 0.84 [95% CI = 0.74, 0.95], and 0.89 [95% CI = 0.80, 1.00] per SD [9.47%, 8.35%, and 8.91%] increase, respectively). However, none of the results withstood multiple testing correction ($FDR < 0.05$) (Supplementary Figure 3). For those CpGs where multiple mQTLs were used as instruments (cg05575921-*AHRR* and cg01901332-*ARRB1*), there was limited evidence for heterogeneity in MR effect

estimates (Q-test, $P > 0.05$, Supplementary Table 6).

Single mQTLs for cg05575921-*AHRR*, cg27241845-*ALPPL2*, and cg26963277-*KCNQ1* showed some evidence of association with smoking cessation (former vs. current smokers), although these associations were not below the $FDR < 0.05$ threshold (Supplementary Figure 4).

Potential causal effect of *AHRR* methylation on lung cancer risk: one sample MR

In the CCHS, a per (average methylation-increasing) allele change in a four-mQTL allele score was associated with a 0.73% [95% CI=0.56,0.90] increase in methylation ($P < 1 \times 10^{-10}$) and explained 0.8% of the variance in cg05575921-*AHRR* methylation (F-statistic=74.2). Confounding factors were not strongly associated with the genotypes in this cohort ($P \geq 0.11$) (Supplementary Table 7). Results provided some evidence for an effect of cg05575921 methylation on total lung cancer risk (HR=0.30 [95% CI=0.10,1.00] per SD (9.2%) increase) (Supplementary Table 8). The effect estimate did not change substantively when stratified by smoking status (Supplementary Table 8).

Given contrasting findings with the main MR analysis, where cg05575921-*AHRR* methylation was not causally implicated in lung cancer, and the lower power in the one-sample analysis to detect an effect of equivalent size to the observational results (power = 19% at $\alpha = 0.05$), we performed further two-sample MR based on the four mQTLs using data from both CCHS (sample one) and the TRICL-ILCCO consortium (sample two). Results showed no strong evidence for a causal effect of DNA methylation on total lung cancer risk (OR=1.00 [95% CI=0.83,1.10] per SD increase) (Supplementary Figure 5). There was also limited evidence for an effect of cg05575921-*AHRR* methylation when stratified by cancer subtype and

smoking status (Supplementary Figure 5) and no strong evidence for heterogeneity of the mQTL effects (Supplementary Table 9). Conclusions were consistent when MR-Egger (27) was applied (Supplementary Figure 5) and when accounting for correlation structure between the mQTLs (Supplementary Table 9).

7.4.3 Tumour and adjacent normal lung tissue methylation patterns

For cg05575921-*AHRR*, there was no strong evidence for differential methylation between adenocarcinoma tissue and adjacent healthy tissue ($P=0.963$), and weak evidence for hypermethylation in squamous cell carcinoma tissue ($P=0.035$) (Figure 4, Supplementary Table 10). For the other CpG sites there was evidence for a difference in DNA methylation between tumour and healthy adjacent tissue at several sites in both adenocarcinoma and squamous cell carcinoma, with consistent differences for CpG sites in *ALPPL2* (cg2156642, cg05951221 and cg01940273), as well as cg23771366-*PRSS23*, cg26963277-*KCNQ1*, cg09935388-*GFI1*, cg0101332-*ARRB1*, cg08709672-*AVPR1B* and cg25305703-*CASC21*. However, hypermethylation in tumour tissue was found for the majority of these sites, which is opposite to what was observed in the EWAS analysis.

(Figure 4 here)

7.4.4 Gene expression associated with mQTLs in blood and lung tissue

Of the 10 genes annotated to the 14 CpG sites, eight genes were expressed sufficiently to be detected in lung (*AVPR1B* and *CASC21* were not) and seven in blood (*AVPR1B*, *CASC21* and *ALPPL2* were not). Of these, gene expression of *ARRB1* could not be investigated as the mQTLs in that region were not present in

the GTEx data. rs3748971 and rs878481, mQTLs for cg21566642 and cg05951221 respectively, were associated with increased expression of ALPPL2 ($P=0.002$ and $P=0.0001$). No other mQTLs were associated with expression of the annotated gene at a Bonferroni corrected P value threshold ($P<0.05/19=0.0026$) (Supplementary Table 11).

7.5 Discussion

In this study, we identified 16 CpG sites associated with lung cancer, of which 14 have been previously identified in relation to smoke exposure (43) 9 and six were highlighted in a previous study as being associated with lung cancer (107) 3. This previous study used the same data from the four cohorts investigated here, but in a discovery and replication, rather than meta-analysis framework. Overall, using MR we found limited evidence supporting a potential causal effect of methylation at the CpG sites identified in peripheral blood on lung cancer. These findings are in contrast to previous analyses suggesting that methylation at two CpG sites investigated (in *AHRR* and *F2RL3*) mediated > 30% of the effect of smoking on lung cancer risk (106) 2. This previous study used methods which are sensitive to residual confounding and measurement error that may have biased results [(85); Hemani2017] 12, 35. These limitations are largely overcome using MR (85) 12. While there was some evidence for an effect of methylation at some of the other CpG sites on risk of subtypes of lung cancer, these effects were not robust to multiple testing correction and were not validated in the analysis of tumour and adjacent normal lung tissue methylation nor in gene expression analysis.

A major strength of the study was the use of two-sample MR to integrate an extensive epigenetic resource and summary data from a large lung cancer GWAS to appraise

causality of observational associations with >99% power. Evidence against the observational findings were also acquired through tissue-specific DNA methylation and gene expression analyses.

Limitations include potential “winner’s curse” which may bias causal estimates in a two-sample MR analysis towards the null if the discovery sample for identifying genetic instruments is used as the first sample, as was done for our main MR analysis using data from ARIES (134) 36. However, findings were similar when using replicated mQTLs in NSHDS, indicating the potential impact of this bias was minimal (Supplementary Figure 1). Another limitation relates to the potential issue of consistency and validity of the instruments across the two samples. For a minority of the mQTL-CpG associations (4 out of 16), there was limited replication across time points and in particular, 6 mQTLs were not strongly associated with DNA methylation in adults. Further, our primary data used for the first sample in the two-sample MR was ARIES, which contains no male adults. If the mQTLs identified vary by sex and time, then this could bias our results. However, our replication cohort NSHDS contains adult males. Therefore, the 10 mQTLs that replicated in NSHDS are unlikely to be biased by the sex discordance. Also, we replicated the findings for cg05575921 *AHRR* in CCHS, which contains both adult males and females, in a two-sample MR analysis, suggesting these results are also not influenced by sex discordance. Caution is therefore warranted when interpreting the null results for the two-sample MR estimates for the CpG sites for which mQTLs were not replicated, which could be the result of weak-instrument bias.

The lack of independent mQTLs for each CpG site did not allow us to properly appraise horizontal pleiotropy in our MR analyses. Where possible we only included cis-acting mQTLs to minimise pleiotropy and investigated heterogeneity where there were multiple independent mQTLs. Three mQTLs were nominally associated

with smoking phenotypes, but not to the extent that this would bias our MR results substantially. Some of the mQTLs used influence multiple CpGs in the same region, suggesting genomic control of methylation at a regional rather than single CpG level. This was untested, but methods to detect differentially methylated regions (DMRs) and identify genetic variants which proxy for them may be fruitful in probing the effect of methylation across gene regions.

A further limitation relates to the inconsistency in effect estimates between the one- and two-sample MR analysis to appraise the causal role of *AHRR* methylation. While findings in CCHS were supportive of a causal effect of *AHRR* methylation on lung cancer (HR=0.30 [95% CI=0.10,1.00] per SD), in two-sample MR this site was not causally implicated (OR=1.00 [95% CI=0.83,1.10] per SD increase). We verified that this was not due to differences in the genetic instruments used, nor due to issues of weak instrument bias. Given the CCHS one-sample MR had little power (19% at $\alpha = 0.05$) to detect a causal effect with a size equivalent to that of the observational analysis, we have more confidence in the results from the two-sample approach.

Peripheral blood may not be the ideal tissue to assess the association between DNA methylation and lung cancer. While a high degree of concordance in mQTLs has been observed across lung tissue, skin and peripheral blood DNA (135) 37, we were unable to directly evaluate this here. A possible explanation for a lack of causal effect at *AHRR* is due to the limitation of tissue specificity as we found that the mQTLs used to instrument cg05575921 were not strongly related to expression of *AHRR* in lung tissue. However, findings from MR analysis were corroborated by the lack of evidence for differential methylation at *AHRR* between lung adenocarcinoma tissue and adjacent healthy tissue, and weak evidence for hypermethylation (opposite to the expected direction) in squamous cell lung cancer tissue. This result may be

interesting in itself as smoking is hypothesized to influence squamous cell carcinoma more than adenocarcinoma. However, the result conflicts with that found in the MR analysis. Furthermore, another study investigating tumorous lung tissue (N=511) found only weak evidence for an association between smoking and cg05575921 *AHRR* methylation, that did not survive multiple testing correction ($P=0.02$) (136) 38. However, our results do not fully exclude *AHRR* from involvement in the disease process. *AHRR* and *AHR* form a regulatory feedback loop, which means that the actual effect of differential methylation or differential expression of *AHR/AHRR* on pathway activity is complex (104) 39. In addition, some of the CpG sites identified in the EWAS were found to be differentially methylated in the tumour and adjacent normal lung tissue comparison. While this could represent a false negative result of the MR analysis, it is of interest that differential methylation in the tissue comparison analysis was typically in the opposite direction to that observed in the EWAS. Furthermore, while this method can be used to minimize confounding, it does not fully eliminate the possibility of bias due to reverse causation (whereby cancer induces changes in DNA methylation) or intra-individual confounding e.g. by gene expression. Therefore, it doesn't give conclusive evidence that DNA methylation changes at these sites are not relevant to the development of lung cancer.

While DNA methylation in peripheral blood may be predictive of lung cancer risk, according to the present analysis it is unlikely to play a causal role in lung carcinogenesis at the CpG sites investigated. Findings from this study issue caution over the use of traditional mediation analyses to implicate intermediate biomarkers (such as DNA methylation) in pathways linking an exposure with disease, given the potential for residual confounding in this context (85) 12. However, the findings of this study do not preclude the possibility that other DNA methylation changes are causally related to lung cancer (or other smoking-associated disease) (137)

40.

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{ - }` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readability and/or setup.

In the main Rmd file

In Chapter ??:

Appendix B

The Second Appendix, for Fun

References

1. Suzuki MM, Bird A. DNA methylation landscapes: Provocative insights from epigenomics. 2008;
2. Siegfried Z, Eden S, Mendelsohn M, et al. DNA methylation represses transcription in vivo. *Nature Genetics*. 1999;
3. Bird A. DNA methylation patterns and epigenetic memory. 2002;
4. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* [electronic article]. 2012;13(7):484–492. (<http://www.nature.com/articles/nrg3230>)
5. Epigenetics: Fundamentals, History, and Examples | What is Epigenetics? (<https://www.whatisepigenetics.com/fundamentals/>). (Accessed June 15, 2020)
6. Kornberg RD. Eukaryotic transcriptional control. 1999;
7. Hafner A, Bulyk ML, Jambhekar A, et al. The multiple mechanisms that regulate p53 activity and cell fate. 2019;

8. Corbett AH. Post-transcriptional regulation of gene expression and human disease. 2018;
9. Wang YC, Peterson SE, Loring JF. Protein post-translational modifications and regulation of pluripotency in human stem cells. 2014;
10. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight? 2008;
11. Greally JM. A user's guide to the ambiguous word 'epigenetics'. 2018;
12. Stern CD, Conrad H. Waddington's contributions to avian and mammalian development, 1930-1940. *International Journal of Developmental Biology*. 2000;
13. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. 2011;
14. Berger SL. The complex language of chromatin regulation during transcription. 2007;
15. Jenuwein T, Allis CD. Translating the histone code. 2001;
16. Birney E, Smith GD, Greally JM. Epigenome-wide Association Studies and the Interpretation of Disease -Omics. *PLOS Genetics* [electronic article]. 2016;12(6):e1006105. (<http://dx.plos.org/10.1371/journal.pgen.1006105>)
17. Holliday R, Pugh J. DNA modification mechanisms and gene activity during development. *Science*. 1975;
18. Riggs AD. X inactivation, differentiation, and DNA methylation. *Cytogenetic*

- and Genome Research*. 1975;
19. Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*. 2000;
 20. Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. 1997;
 21. Ando M, Saito Y, Xu G, et al. Chromatin dysregulation and DNA methylation at transcription start sites associated with transcriptional repression in cancers. *Nature Communications*. 2019;
 22. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes and Development*. 2011;
 23. Wolf SF, Jolly DJ, Lunnen KD, et al. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: Implications for X chromosome inactivation. *Proceedings of the National Academy of Sciences of the United States of America*. 1984;
 24. Hellman A, Chess A. Gene body-specific methylation on the active X chromosome. *Science*. 2007;
 25. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics*. 2016;
 26. Jaffe AE, Murakami P, Lee H, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*. 2012;
 27. Suderman M, Staley JR, French R, et al. Dmrff: Identifying differentially methylated regions efficiently with power and control. *bioRxiv*. 2018;

28. Challen GA, Sun D, Jeong M, et al. Dnmt3a is essential for hematopoietic stem cell differentiation. *Nature Genetics*. 2012;
29. Lock LF, Takagi N, Martin GR. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell*. 1987;
30. Ohm JE, McGarvey KM, Yu X, et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature Genetics*. 2007;
31. Prendergast GC, Ziff EB. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science*. 1991;
32. Wrzeska M, Rejdach B. Genomic imprinting in mammals. 2004;
33. Nicholls RD. The impact of genomic imprinting for neurobehavioral and developmental disorders. 2000;
34. Relton CL, Davey Smith G. Epigenetic Epidemiology of Common Complex Disease: Prospects for Prediction, Prevention, and Treatment. *PLoS Medicine* [electronic article]. 2010;7(10):e1000356. (<http://dx.plos.org/10.1371/journal.pmed.1000356>)
35. Weaver ICG, Cervoni N, Champagne FA, et al. Epigenetic programming by maternal behavior. *Nature Neuroscience*. 2004;
36. Koch A, Joosten SC, Feng Z, et al. Analysis of DNA methylation in cancer: Location revisited. 2018;
37. Hentze J, Høgdall C, Høgdall E. Methylation and ovarian cancer: Can DNA methylation be of diagnostic use? (Review). *Molecular and Clinical Oncology*. 2019;

38. Cortellino S, Xu J, Sannai M, et al. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell*. 2011;
39. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. 2013;
40. Venolia L, Gartler SM. Comparison of transformation efficiency of human active and inactive X-chromosomal DNA. 1983;
41. Wade PA, Wolffe AP. ReCoGnizing methylated DNA. 2001;
42. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*. 2015;
43. Joehanes R, Just AC, Marioni RE, et al. Epigenetic Signatures of Cigarette Smoking. *Circulation: Cardiovascular Genetics*. 2016;
44. Breitling LP, Yang R, Korn B, et al. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *American Journal of Human Genetics*. 2011;
45. Wahl S, Drong A, Lehne B, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* [electronic article]. 2017;541(7635):81–86. (<http://dx.doi.org/10.1038/nature20784>)
46. Yang BZ, Zhang H, Ge W, et al. Child abuse and epigenetic mechanisms of disease risk. *American Journal of Preventive Medicine*. 2013;
47. Felix JF, Joubert BR, Baccarelli AA, et al. Cohort profile: Pregnancy and childhood epigenetics (PACE) consortium. *International Journal of Epidemiology*.

2018;

48. Philibert R, Dogan M, Beach SRH, et al. AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*. 2020;
49. Philibert R, Hollenbeck N, Andersen E, et al. Reversion of AHRR demethylation is a quantitative biomarker of smoking cessation. *Frontiers in Psychiatry*. 2016;
50. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biology*. 2013;
51. Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. 2015;
52. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. 2018;
53. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;
54. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*. 2014;
55. McGregor K, Bernatsky S, Colmegna I, et al. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biology*. 2016;
56. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: A review and recommendations. 2017;

57. Forest M, O'Donnell KJ, Voisin G, et al. Agreement in DNA methylation levels from the Illumina 450K array across batches, tissues, and time. *Epigenetics*. 2018;
58. Yang AS. A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. *Nucleic Acids Research*. 2004;
59. Dedeurwaerder S, Defrance M, Calonne E, et al. Evaluation of the Infinium Methylation 450K technology. *Epigenomics*. 2011;
60. Price EM, Robinson WP. Adjusting for batch effects in DNA methylation microarray data, a lesson learned. *Frontiers in Genetics*. 2018;
61. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*. 2007;
62. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;
63. Perrier F, Novoloaca A, Ambatipudi S, et al. Identifying and correcting epigenetics measurements for systematic sources of variation. *Clinical Epigenetics*. 2018;
64. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic acids research*. 2017;
65. Naeem H, Wong NC, Chatterton Z, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics*. 2014;

66. Piekarz RL, Bates SE. Epigenetic modifiers: Basic understanding and clinical development. 2009;
67. Nebbioso A, Tambaro FP, Dell'Aversana C, et al. Cancer epigenetics: Moving forward. 2018;
68. Pickar-Oliver A, Gersbach CA. The next generation of CRISPR–Cas technologies and applications. 2019;
69. Liao HK, Hatanaka F, Araoka T, et al. In Vivo Target Gene Activation via CRISPR/Cas9-Mediated Trans-epigenetic Modulation. *Cell*. 2017;
70. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*. 2019;
71. Liu D, Zhao L, Wang Z, et al. EWASdb: Epigenome-wide association study database. *Nucleic Acids Research*. 2019;
72. Li M, Zou D, Li Z, et al. EWAS Atlas: A curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Research*. 2019;
73. Maher B. Personal genomes: The case of the missing heritability. 2008;
74. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. 2009;
75. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*. 2010;
76. Phipson B, Maksimovic J, Oshlack A. MissMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*. 2016;

77. Philibert RA, Beach SRH, Brody GH. Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. *Epigenetics*. 2012;
78. Shenker NS, Polidoro S, Veldhoven K van, et al. Epigenome-wide association study in the European Prospective Investigation Into Cancer And Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Human Molecular Genetics*. 2013;
79. Zeilinger S, Kühnel B, Klopp N, et al. Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *PLoS ONE*. 2013;
80. Bojesen SE, Timpson N, Relton C, et al. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax*. 2017;
81. Haarmann-Stemmann T, Abel J. The arylhydrocarbon receptor repressor (AhRR): Structure, expression, and function. 2006;
82. Larigot L, Juricek L, Dairou J, et al. AhR signaling pathways and regulatory functions. 2018;
83. Davey Smith G, Ebrahim S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? 2003;
84. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics* [electronic article]. 2014;23(R1):R89–98. (<http://hmg.oxfordjournals.org/cgi/content/long/23/R1/R89>)
85. Richmond RC, Hemani G, Tilling K, et al. Challenges and novel approaches for investigating molecular mediation. *Human molecular genetics* [electronic

article]. 2016;ddw197. (<http://hmg.oxfordjournals.org/cgi/content/long/ddw197v2>)

86. Silventoinen K, Kaprio J, Lahelma E, et al. Assortative mating by body height and BMI: Finnish twins and their spouses. *American Journal of Human Biology*. 2003;
87. Maes HHM, Neale MC, Eaves LJ. Genetic and environmental factors in relative body weight and human adiposity. 1997;
88. Eaves LJ, Heath AC. Detection of the effects of asymmetric assortative mating. *Nature*. 1981;
89. Hemani G, Zheng J, Elsworth B, et al. The MR-base platform supports systematic causal inference across the human phenome. *eLife*. 2018;
90. Inoue A, Solon G. Two-Sample Instrumental variables estimators. *The Review of Economics and Statistics*. 2010;92(3):557–561.
91. Pierce BL, Burgess S. Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*. 2013;178(7):1177–1184.
92. Relton CL, Davey Smith G. Two-step epigenetic mendelian randomization: A strategy for establishing the causal role of epigenetic processes in pathways to disease. *International Journal of Epidemiology*. 2012;
93. Richardson TG, Haycock PC, Zheng J, et al. Systematic Mendelian randomization framework elucidates hundreds of CpG sites which may mediate the influence of genetic variants on disease. *Human Molecular Genetics*. 2018;
94. Bonder MJ, Luijk R, Zhernakova DV, et al. Disease variants alter transcrip-

- tion factor levels and methylation of their binding sites. *Nature Genetics*. 2017;
95. Lawlor DA, Tilling K, Smith GD. Triangulation in aetiological epidemiology. *International Journal of Epidemiology*. 2016;
 96. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nature Reviews Genetics* [electronic article]. 2013;14(8):585–594. (<http://www.nature.com/doifinder/10.1038/nrg3405>)
 97. Rakyan VK, Down TA, Balding DJ, et al. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics* [electronic article]. 2011;12(8):529–541. (<http://www.nature.com/articles/nrg3000>)
 98. Banos DT, McCartney DL, Battram T, et al. Bayesian reassessment of the epigenetic architecture of complex traits. *bioRxiv* [electronic article]. 2018;450288. (<http://biorxiv.org/content/early/2018/11/14/450288.abstract>)
 99. Huang WY, Hsu SD, Huang HY, et al. MethHC: A database of DNA methylation and gene expression in human cancer. *Nucleic Acids Research*. 2015;
 100. Fraser A, Macdonald-wallis C, Tilling K, et al. Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *International Journal of Epidemiology*. 2013;
 101. Boyd A, Golding J, Macleod J, et al. Cohort profile: The 'Children of the 90s'-The index offspring of the avon longitudinal study of parents and children. *International Journal of Epidemiology*. 2013;42(1):111–127.

102. Li H. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*. 2011;
103. Relton CL, Gaunt T, McArdle W, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *International Journal of Epidemiology* [electronic article]. 2015;44(4):1181–1190. (<http://dx.doi.org/10.1093/ije/dyv072>)
104. Chen J, Behnam E, Huang J, et al. Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. *BMC Genomics*. 2017;
105. Ferlay J, Soerjomataram I, Ervik M, et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase. No. 11 [Internet]. 2013;11:<http://globocan.iarc.f>. (<http://arxiv.org/abs/arXiv:1011.1669v3>)
106. Fasanelli F, Baglietto L, Ponzi E, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nature Communications* [electronic article]. 2015;6:10192. (<http://www.nature.com/doi/10.1038/ncomms10192>)
107. Baglietto L, Ponzi E, Haycock P, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *International Journal of Cancer*. 2017;140(1):50–61.
108. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*. 2016;
109. Stratthdee G, Brown R. Aberrant DNA methylation in cancer: potential clinical interventions. *Expert Reviews in Molecular Medicine*. 2002;

110. Jones P a, Baylin SB. The fundamental role of epigenetic events in cancer. *Nature reviews. Genetics* [electronic article]. 2002;3(6):415–28. (<http://www.ncbi.nlm.nih.gov/pubmed/16136652%7B/%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/12042769>)
111. Borghol N, Suderman M, Mcardle W, et al. Associations with early-life socio-economic position in adult DNA methylation. *International Journal of Epidemiology*. 2012;
112. Zudaire E, Cuesta N, Murty V, et al. The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers. *The Journal of clinical investigation*. 2008;
113. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *American Journal of Epidemiology*. 2007;
114. Munafò MR, Timofeeva MN, Morris RW, et al. Association between genetic variants on chromosome 15q25 locus and objective measures of Tobacco exposure. 2012;
115. Richardson TG, Zheng J, Davey Smith G, et al. Mendelian Randomization Analysis Identifies CpG Sites as Putative Mediators for Genetic Influences on Cardiovascular Disease Risk. *American Journal of Human Genetics*. 2017;
116. Gaunt TR, Shihab HA, Hemani G, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biology* [electronic article]. 2016;17(1):61. (<http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0926-z>)

117. Leek JT, Johnson WE, Parker HS, et al. sva: Surrogate Variable Analysis. 2016;
118. Pidsley R, Y Wong CC, Volta M, et al. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;
119. Touleimat N, Tost J. Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*. 2012;
120. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. 2007;
121. McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature genetics* [electronic article]. 2017;49(7):1126–1132. (<http://www.nature.com/doifinder/10.1038/ng.3892%7B/%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/28604730>)
122. Thomas DC, Lawlor DA, Thompson JR. Re: Estimation of Bias in Non-genetic Observational Studies Using "Mendelian Triangulation" by Bautista et al. 2007;
123. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* [electronic article]. 2015;44(2):512–525. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4469799/>)

124. Zhan X, Hu Y, Li B, et al. RVTESTS: An efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics*. 2016;32(9):1423–1426.
125. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nature Genetics*. 2016;
126. Furberg H, Kim Y, Dackor J, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*. 2010;
127. Haycock PC, Burgess S, Wade KH, et al. Best (but oft-forgotten) practices: The design, analysis, and interpretation of Mendelian randomization studies. *American Journal of Clinical Nutrition*. 2016;
128. Kaur-Knudsen D, Bojesen SE, Tybjaerg-Hansen A, et al. Nicotinic acetylcholine receptor polymorphism, smoking behavior, and tobacco-related cancer and lung and cardiovascular diseases: A cohort study. *Journal of Clinical Oncology*. 2011;
129. Bojesen SE, Pooley KA, Johnatty SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature Genetics*. 2013;
130. Teschendorff AE, Yang Z, Wong A, et al. Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. *JAMA oncology* [electronic article]. 2015;1(4):476–85. (<http://oncology.jamanetwork.com/article.aspx?articleid=2293216>)
131. GTEx Consortium TG. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* [electronic article]. 2013;45(6):580–5. (<http://www>.

ncbi.nlm.nih.gov/pubmed/23715323%7B/%%7D5Cnhttp://
www.pubmedcentral.nih.gov/articlerender.fcgi?artid=
PMC4010069)

132. Hemani G, Zheng J, Wade KH, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv* [electronic article]. 2016;(http://biorxiv.org/content/early/2016/12/16/078972.abstract)
133. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995;
134. Burgess S, Thompson SG. Avoiding bias from weak instruments in mendelian randomization studies. *International Journal of Epidemiology*. 2011;
135. Shi J, Marconett CN, Duan J, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nature communications*. 2014;
136. Freeman JR, Chu S, Hsu T, et al. Epigenome-wide association study of smoking and DNA methylation in non-small cell lung neoplasms. *Oncotarget*. 2016;7(43).
137. Gao X, Zhang Y, Breitling LP, et al. Tobacco smoking and methylation of genes related to lung cancer development. *Oncotarget*. 2016;

Abbreviations

EWAS - epigenome-wide association study **GWAS** - genome-wide association study **h^2** - narrow-sense heritability **h^2_{SNP}** - SNP-heritability **H^2** - broad-sense heritability **MR** - Mendelian randomization **mQTL** - methylation quantitative trait loci **SNP** - single nucleotide polymorphism