I made this template based on thesisdown 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 regulome 4
    1.1.1 The regulome 4
    1.1.2 Defining epigenetics 5
    1.1.3 Histone modifications 5
    1.1.4 DNA methylation 6
  1.2 Use of DNA methylation in a population setting 8
  1.3 Epigenome-wide association studies 9
  1.4 Using methods from genetics to help inform future EWAS 10
    1.4.1 The GWAS Catalog 10
    1.4.2 Total variance captured by all sites measured genome-wide 11
    1.4.3 Inferring biology from signals 12
    1.4.4 Establishing causality 14
  1.5 Overview of thesis aims 15

Chapter 2: Methods **17**

Chapter 3: The EWAS Catalog: a database of epigenome-wide association studies **19**
  3.1 Abstract 19
  3.2 Introduction 20
  3.3 Methods 22
    3.3.1 Implementation 22
    3.3.2 Overview of publication data extraction 22
    3.3.3 Overview of GEO data extraction 23
    3.3.4 EWAS methods 24
      Avon Longitudinal Study of Parents and Children (ALSPAC) 24

GEO datasets 24
3.4 Results 25
3.4.1 Database interface and use 25
3.5 Discussion 27

Chapter 4: Properties of EWAS 29

Chapter 5: m2 31

Chapter 6: EWAS-GWAS comparison 33

Chapter 7: DNAm-lung cancer MR 35
7.1 Abstract 35
7.1.1 Background 35
7.1.2 Methods 35
7.1.3 Results 36
7.1.4 Conclusions 36
7.2 Introduction 36
7.3 Methods 38
7.3.1 EWAS study details 38
European Prospective Investigation into Cancer and Nutrition-Italy (EPIC-Italy) 39
Melbourne Collaborative Cohort Study (MCCS) 39
Norwegian Women and Cancer (NOWAC) 40
Northern Sweden Health and Disease Study (NSHDS) . . 41
7.3.2 EWAS Meta-analysis 42
7.3.3 Mendelian randomization 43
Sample 1: Accessible Resource for Integrated Epigenomic Studies (ARIES) 43
7.3.4 Sample 2: Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium . . . 45
7.3.5 Supplementary analyses 47
Assessing the potential causal effect of *AHRR* methylation: one sample MR 47
Tumour and adjacent normal methylation patterns 48
mQTL association with gene expression 48
7.4 Results 48
7.4.1 EWAS meta-analysis 49
7.4.2 Mendelian randomization 49

Potential causal effect of AHRR methylation on lung cancer risk: one sample MR 50
7.4.3 Tumour and adjacent normal lung tissue methylation patterns 51
7.4.4 Gene expression associated with mQTLs in blood and lung tissue 52
7.5 Discussion 52

Conclusion **57**

Appendix A: The First Appendix **59**

Appendix B: The Second Appendix, for Fun **61**

References **63**

List of Tables

List of Figures

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. 26

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

This thesis focuses on epigenome-wide association studies (EWAS), which assess the association between DNA methylation changes throughout the genome and traits of interest. Advances in technology and an increasing number of samples has lead to hundreds of EWAS having been performed to date (REFERENCE EWAS CATALOG CHAPTER). Goals of these EWAS included using DNA methtlation marks as predictors, diagnostic factors, and modifiable mediators of traits.

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. Further, to improve biological inference of sites identified in EWAS, causality (or lack of) should be established. IS THIS TOO VAGUE?

NON-VAGUE THING TO SAY – read over both and decide what to add: - Key things which are not currently understood: - Why certain DNA methylation sites/regions of the genome are more prevalent in EWAS to date - How much trait variation DNA methylation tends to associate with - Whether EWAS can add to current biological understanding in addition to other study designs - Whether DNA methylation-trait associations in EWAS are causal

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

1.1 DNA methylation as part of the regulome

The research questions of this thesis pertain to a specific study type, EWAS, of a specific epigenetic mark, DNA methylation. DNA methylation is one of many epigenetic marks that are involved in gene regulation (the regulome), so here I briefly outline some of these regulatory marks and discuss DNA methylation in the context of this complex biological process.

1.1.1 The regulome

Gene expression is a tightly controlled process and only in the right circumstances will RNA polymerase II be able to bind the correct site, initiate and finally complete the transcription process (REFS). 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 100s of unique cell types (REF). As these cells arise from a single progenitor they must contain identical genetic sequences (bar 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 its products (REFs). Indeed it was in developmental biology that we first began to understand the processes that may regulate gene expression (REFs). There are a plethora of methods cells use to regulate gene and protein expression post-transcriptionally (REFs), but these are beyond the scope of this thesis. So I'll continue to describe the regulome just in the context of epigenetic marks.

1.1.2 Defining epigenetics

The definition of epigenetics is much debated amongst those who study cellular and molecular processes (REFs). Waddington first coined the term and he originally defined it as **X** (REF). 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 simply mean 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 genome-bound proteins composed of four subgroups that are present twice each in a single histone octamer. Modifications can occur to any of the histone monomers and these have been associated with both positive and negative changes in gene expression (REFs). Histone modifications are numerous and complex in nature. To briefly describe the complexity, there are at least **X** types of histone modifications, each of the histone monomers can be modified across many different sites, and for any one site mutiple of the same modification can occur and it is the combination of all of this that plays a role in gene expression regulation (REFs). Furthermore, histone modifications are subject to rapid change upon environmental stimulus to help induce or repress gene expression (REFs). This has meant, that for only a few histone marks are we certain of the role (REFs) and that taking a snapshot of the histone modifications present in cells may give a poor summary of how gene regulation is occuring. Very few population-based studies have assessed histone modifications because of these, plus some other technical difficulties (REFs). 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

Good review: Peter Jones 2012 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 only measure CpG methylation, so that will be my focus here. The function of this epigenetic mark was proposed back in 1975 (REFs), where two papers suggested it represses gene expression and ever since then many papers have shown the correlation between DNA methylations around genes and a reduction in expression of those genes. Unfortunately, it's not that simple and even now the role of DNA methylation is being debated, with a paper recently suggesting that any DNA methylation changes associated with changes in gene expression were just a by-product of other regulatory processes such as transcription factor binding (REF). Regardless, DNA methylation is closely linked with genomic function and correlates with various aspects in different ways (**FIGURE**).

CpG sites are not randomly distributed throughout the genome, but are often found in clusters and so if DNA methylation does have an effect on things like gene expression it's thought that methylation and de-methylation of CpG sites in groups is what drives their regulatory function. To support this, there are clear biological processes that regulate DNA methylation at nearby sites together, for example, CpG sites at transcription factor binding sites will be de-methylated as a group when the transcription factor binds (REF), and further nearby sites are often statistically correlated. 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 driving regulatory function. This is something I explore a little more in (REFERENCE h2ewas CHAPTER!).

One 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 demethylate the DNA, for example the ten-eleven translocation (TET) enzymes, but ultimately cell division is required for full demethylation of a DNA molecule. Biologically, this suggests DNA methylation might be involved in long-term repression of gene expression, which can be seen for X-inactivation, and practically it makes studying the epigenetic mark easier as it prevents large temporal variations in DNA methylation (though they may occur!), so fewer samples are needed.

Figure here:

Figure should contain these diagrams: 1. DNA methylation at TSS and no DNA methylation in gene body (repressed transcription) 2. No DNA methylation at TSS and DNA methylation in gene body (active transcription) 3. DNA methylation in centromeres 4. DNA methylation at transposable elements

1.2 Use of DNA methylation in a population setting

The relationship between a large number of traits and DNA methylation have now been studied, from childhood adversities (REFs) to smoking (REFs), and one of the appealing factors of doing this research is that DNA methylation is modifiable. Thus, in theory, negative effects of something like childhood adversity may be captured in DNA methylation and could potentially be, in part, reversed by altering the methylome in some way. This view that clinicians may soon have the power to do anything like this has likely lead to a lot of poor and inconsequential EWAS being published. Unfortunately, there are massive complications to this research. Firstly, it should be recognised that DNA methylation has the same undesirable properties of every other complex phenotype, it correlates with a lot of things, making any associations suceptible to confounding and reverse causation, so determining causality is tricky (this will be discussed in 1.4). Secondly, current methods used to modify DNA methylation mostly rely on broad remodelling of the DNA methylome (e.g. 5-AZA), which is unviable in the clinical setting for treatment of most things. Granted, there are methods that can be used to increase specificity of targetting DNA methylation changes in the genome (REFs), these have not yet been adapted for use in the clinic and given the brevity of their existence, it's unlikely this will occur soon. Finally, DNA methylation varies between tissue and cell types. Not only does this pose the issue of confounding when pooling cell types from a tissue, it also leads to large complexities in the interpretation of DNA methylation studies. Most studies are conducted in readily available tissues, such as blood and saliva, for obvious reasons, but there is no way of knowing which tissue is best suited for studying DNA methylation for many complex traits. DNA methylation does correlate between tissues (REFs), which may mean a readily available tissue can act as a surrogate, but this also means targeting DNA methylation in the tissue measured for the EWAS, may have no impact on your phenotype. All together these problems mean that establishing causality is difficult, then targeting specific DNA methylation sites of interest to effect change is also challenging (and untested in humans), and finally the tissue that needs to be targetted may be impossible to target anyway. Establishing causality in EWAS is a focus of **7**.

These difficulties do not exclude DNA methylation as a trait that is important to study in a population setting; it is not necessary to establish causality to use DNA methylation as a diagnostic biomarker or predictor of future disease. Further, important biological information can be gained from studying DNA methylation with relation to many traits of interest, but one must be careful how they approach and interpret their studies.

1.3 Epigenome-wide association studies

Epigenome-wide associaiton studies are the flagship study design for estimating the association of DNA methylation with a trait of interest in a population setting. The premise is simple, DNA methylation is measured across parts of the whole genome, often using an array such as the Illimnum HumanMethylation 450 Beadchip (HM450), then simple regressions are conducted to estimate the association of DNA methylation at each site with a trait of interest.

ALSO ADD IN THAT DMR ANALYSIS IS COMMON IN EWAS –> WANT TO REFERENCE H2EWAS PAPER AGAIN!

- Describe what an EWAS is and explain how it is conducted
- Describe the current successes that have come from EWAS
- Describe the current problems with EWAS

1.4 Using methods from genetics to help inform future EWAS

There is a clear correlary of the EWAS in genetics, the genome-wide association study (GWAS), which also measures markers across the whole genome and assesses whether each of these markers associates with the phenotype of interest. GWAS have been around for far longer than EWAS and a huge amount of effort has been put into understanding what information is provided by these studies, what information can be discovered by these studies and how to use these results to inform other research.

1.4.1 The GWAS Catalog

The GWAS Catalog is a manually curated database of publically available GWAS data, developed by the EBI and made openly available to the public (REFs). 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 corralary database for EWAS may provide equal opportunity for epigenetic epidemiologists. Catalogs such as EWASdb (REF) and the EWAS Atlas (REF) are currently available, but fall short of some researcher requirements including ease of use and access to full summary statistics. The development of a new database is the focus of **Chapter 3**.

1.4.2 Total variance captured by all sites measured genome-wide

If at all possible, it's important to have an understanding of whether your 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. It is possible to estimate the proportion of phenotypic variance that is attributable to genetic factors by estimating heriability (H^2). The vast proportion of H^2 for any given complex trait is made up of the the additive effects h^2, or narrow-sense heritability. Without measuring any genotypes it is possible to estimate h^2 for a trait given the relatedness of individuals in a sample and an ability to minimise environmental influences, twin studies are an example of this. Simply, if monozygotic twins, who share 100% of their DNA sequence, tend to be more similar for the trait of interest than dizygotic twins, who share 50%, then the trait would be estimated to have some additive genetic component ($h^2 > 0$). The degree of difference between the estimates would provide an estimate of h^2 and so the contribution of all additive genetic effects to the variance of the trait. Heritability analyses have been conducted for a huge amount of traits, which provide adequate evidence that when searching for genetic effects on those traits, there is something to find!

After the largely unsuccessful attempt of geneticists using candidate gene studies to identify genetic sequence variation that influences traits (REF), a new study design was proposed using arrays to measure hundreds of thousands of genetic variants across the genome in a hypothesis-free approach to identify sites, GWAS. Initially, these studies 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 heriability estimates (<1%) (REF). 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 (REF). 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). To this end, Yang et al. developed a method to estimate the contribution of all the single nucleotide polymorphisms (SNPs) to phenotypic variance (SNP-heriability or h^2_{SNP}) (REF). SNP-heriability was subsequently shown not to be inconsequential for complex traits and so warranted continuing use of GWAS and acquisition of larger samples to conduct these GWAS.

Lack of associations and lack of phenotypic variance captured by DNA methylation sites has largely been the case in EWAS, yet currently there is no corralary test 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, but lots of small associations across the entire genome is also a possibility. 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, which can help inform future EWAS designs, like SNP-heritability studies helped influence the trajectories of genetic epidemiology.

1.4.3 Inferring biology from signals

A common goal of GWAS and EWAS is to understand the underlying biology of complex traits. Several techniques are often applied post-GWAS (REF) to attempt to accomplish this for genetic variants identified. A very common method is to map sites identified to genes and assess whether these genes are enriched for any particular biological pathway defined by various ontologies (geneset enrichment analysis), for example the Gene Ontology (GO) (REF) or KEGG (REF). This method has also been adapted to EWAS and applied several times (REFs).

In geneset enrichment analysis, the signal of interest is mapped to a gene or genes and these are mapped to pathways. Then using some statistical test, such as the hypergeometric test or Fisher's exact test, evidence that the gene(s) are present in the pathways more than expected by chance is estimated. The null distribution from which the observed numbers are tested against, is usually estimated from the total number of genes tagged by the array of choice or by performing permutation tests.

As discussed previously in 1.2 and in **??**, establishing causality from DNA methylation signal is difficult. Thus, when applying pathway enrichment analyses to EWAS signals, the pathways identified may actually be related to a particular confounder rather than the trait of interest. However, it is still of interest to perform the analysis because 1. it may give evidence that an EWAS signal is true (for example if an EWAS of diabetes picked up insulin pathways) and 2. it can provide some evidence for pathways involved in a trait if corroborated by other forms of evidence.

Further, as DNA methylation changes may be as a result of the trait (unlike genetic variation), pathway analysis for EWAS studies could capture biological consequences of the trait. In **Chapter @ref()** I compare overlap of GWAS and EWAS signals in this context and discuss whether both study designs are likely to be useful in discovering the entire underlying biology of complex traits.

1.4.4 Establishing causality

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 these limitations is Mendelian randomization (MR) (REFs). MR uses genetic variants as proxies for the exposure and outcome of interest in an instrumental variable framework (illustrated in **Figure ??**). Using genetic variants as proxies ensures there is no reverse causation, as the variants are inherited, and reduces chances of confounding as individual genetic variants tend not to be associated with a broad range of phenotypes (REF). Furthermore, genetic variants are randomly segregated at conception, so in the abscence of assortative mating and dynastic effects, should be associated randomly with respect to any confounders. BLAH – Need to read some MR papers again to get the wording right!

TWO SAMPLE MR NEEDS EXPLAINING HERE

PLEIOTROPY NEEDS EXPLAINING HERE

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. By utilizing large trait GWAS and publically available mQTL databases (REFs) in a two-sample MR framework, good statistical power can be attained to estimate the effect of DNA methylation changes on various traits. Unfortunately, for each DNA methylation site few independent mQTLs have been identified, and of these trans-mQTLs (mQTLs over 1Mb from the DNA methylation site), are thought to have a higher probablity of being pleiotropic. This means that ruling out any associations that have arisen because of pleiotropy is difficult. However, applying MR in this context can still add evidence to the observational analyses that have been conducted. In **Chapter ??** I use this principle to investigate the effect of DNA methylation on lung cancer.

Might be an idea to add detail of some of these methods into appendices (e.g. mathematical basis of snp-heritability)

- Give a little intro on what problems this chapter focuses on
- Describe the GWAS catalog and what it has done for genetic epi and how this can be adapted for EWAS
- Describe and explain heritability (and how it can be applied to EWAS)
- Describe pathway analysis??
- Describe MR

1.5 Overview of thesis aims

BELOW IS TERRIBLE BUT HELPFUL AS IT GIVES STRUCTURE

EWAS might be an effective study technique to help provide biomarkers for better diagnosis, prognosis and prediction of disease and traits and further it could provide insight into trait aetiology as well as the downstream effects of a trait. However, with the field in it's infancy, there has yet to be a comprehensive look at what information has been gained from EWAS, what information is still to gain from EWAS in it's current state and whether or not the sites identified in EWAS are likely to be causally related to traits.

In Chapter 3 the aim is to produce a database that brings together all the information currently published (as of X-date) together along with new EWAS with full summary statistics. This vast database will then be explored in Chapter 4, giving a detailed picture of what sites throughout the genome have been discovered in EWAS and exploring why they might have been discovered. NEED TO ADD IN HOW!

After exploring the information already gained from EWAS, Chapter 5 explores the amount of information still to gain from EWAS. NEED TO ADD IN HOW!

The database from Chapter 3 will again be used to extract large EWAS datasets for Chapter 6, where potential biological information from these EWAS will be compared to that of GWAS from the same traits. NEED TO ADD IN HOW!

Finally as an initial means of assessing whether DNA methylation might be causally associated with traits of interest, Chapter 7 will aim to use MR to help infer if DNA methylation identified in EWAS of lung cancer are likely to be causing lung cancer.

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 moleclute 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 $P < 1x10^{-4}$) 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 disaese 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 (Mill & Heijmans, 2013; Rakyan, Down, Balding, & Beck, 2011; Relton & Davey Smith, 2010). These associations may then be used for disease diagnosis or prediction (Mill & Heijmans, 2013; Rakyan et al., 2011; Relton & Davey Smith, 2010). Also, unlike genetic variants, changes in DNA methylation are responsive to the environment and so may be targeted for treatment. EWAS of smoking (Joehanes et al., 2016), body mass index (BMI) (Wahl et al., 2017) and aging (Horvath, 2013) 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 (Banos et al., 2018). 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 (Buniello et al., 2019) 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 (Li et al., 2019). Other databases are available but are limited to certain diseases (e.g. MethHC (Huang et al., 2015)). 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) (Boyd et al., 2013; Fraser et al., 2013) 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 (Li, 2011) 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 $P < 1x10^{-4}$. 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 (Relton et al., 2015). 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

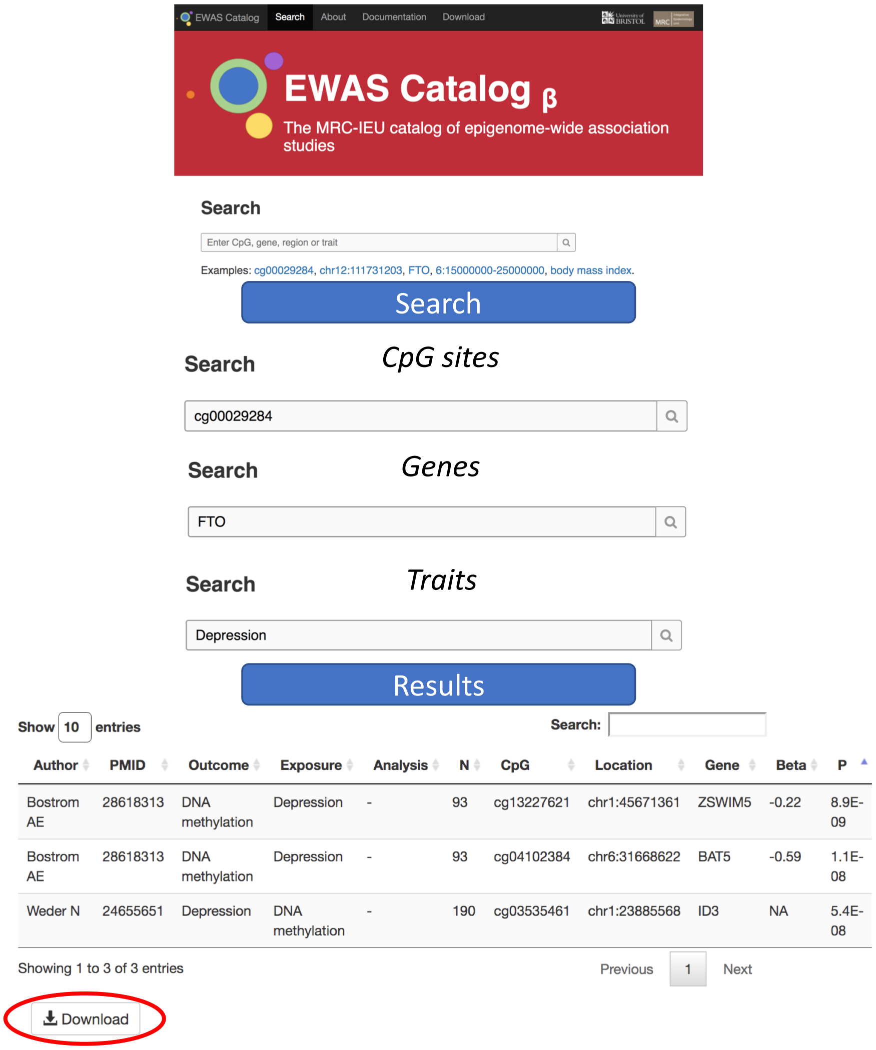
Supplementary Table 1). For all traits, linear regression models were fitted with DNA methylation as the outcome and the phenotype as the exposure as for the ARIES data. Twenty surrogate variables were included as covariates. Other covariates were considered, but surrogate variables only were used for 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 (Chen et al., 2017) 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 (Buniello et al., 2019), 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.



EWAS Catalog β

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

cg00029284

Search

FTO

Search

Depression

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

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: (Relton & Davey Smith, 2010)

Chapter 5

m2

Chapter 6

EWAS-GWAS comparison

Chapter 7

DNAm-lung cancer MR

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 (Ferlay et al., 2013). Several DNA methylation changes have been recently identified in relation to lung cancer risk (Baglietto et al., 2017; Fasanelli et al., 2015; McCarthy et al., 2016). Given the plasticity of epigenetic markers, any DNA methylation changes that are causally linked to lung cancer are potentially appealing targets for intervention (Jones & Baylin, 2002; Strathdee & Brown, 2002). However, these epigenetic markers are sensitive to reverse causation, being affected by cancer processes (Jones & Baylin, 2002), and are also prone to confounding, for example by socio-economic and lifestyle factors [Borghol et al. (2012); Elliott2014].

One CpG site, cg05575921 within the aryl hydrocarbon receptor repressor (*AHRR*) gene, has been consistently replicated in relation to both smoking (Joehanes et al., 2016) and lung cancer (Baglietto et al., 2017; Bojesen, Timpson, Relton, Smith, & Nordestgaard, 2017; Fasanelli et al., 2015) and functional evidence suggests that this region could be causally involved in lung cancer (Zudaire et al., 2008). 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 (Richmond, Hemani, Tilling, Davey Smith, & Relton, 2016), including residual confounding after adjustment for self-reported smoking behaviour (Fewell, Davey Smith, & Sterne, 2007; Munafò et al., 2012). 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 (Baglietto et al., 2017; Fasanelli et al., 2015).

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 (Davey Smith & Ebrahim, 2003; Davey Smith & Hemani, 2014). 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 (Relton & Davey Smith, 2012; Relton et al., 2015; Richardson et al., 2017) with the use of single nucleotide polymorphisms (SNPs) that correlate with methylation of CpG sites, known as methylation quantitative trait loci (mQTLs) (Gaunt et al., 2016).

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 methlyation sites associated with lung cancer. DNA methylation in each EWAS was assessed using the Illumina Infinium® HumanMethylation450 BeadChip. 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 (Baglietto et al., 2017) 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 (Baglietto et al., 2017), 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 (Baglietto et al., 2017). 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 (Baglietto et al., 2017). 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 (Baglietto et al., 2017). 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 (Leek et al., 2016) 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 (Houseman et al., 2012). 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 (Inoue & Solon, 2010; Pierce & Burgess, 2013).

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.mqtldb.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 (2, 3). 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/`) (4). 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$^{\text{TM}}$ 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 (5) using the Touleimat and Tost (6) algorithm to reduce the non-biological differences between probes. Methylation data in ARIES were rank-normalised to remove outliers, and then Matrix eQTL software (7) 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) (8).

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 Génotypage (CNG, Paris, France). Individuals were excluded based on non-European ancestry, missingness, relatedness, gender mismatches and heterozygosity. PLINK (v1.07) (9) 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 $P > 5x10^{-7}$, 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) and used to obtain mQTL-lung cancer estimates (β_{GD}) (McKay et al., 2017).

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 (Thomas, Lawlor, & Thompson, 2007). 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 (Bowden, Smith, & Burgess, 2015). Differences between the observational and MR estimates were assessed using a Z-test for difference.

START HERE!! 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 (Zhan, Hu, Li, Abecasis, & Liu, 2016). 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 (10). Genetic imputation was performed on these samples using the Haplotype Reference Consortium (HRC) Panel (release 1) (11) through the Michigan Imputation Server (12). 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) (McKay et al., 2017) 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) (Furberg et al., 2010) 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 (Baglietto et al., 2017; Fasanelli et al., 2015) 2, 3, we performed a one-sample MR analysis (Haycock et al., 2016) 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). Details of the phenotypic, methylation and genetic data, as well as the linked lung cancer data, are outlined in the Supplementary Methods.

An allele score of mQTLs located with 1Mb of cg05575921-*AHRR* was created and its association with AHRR methylation tested (Supplementary Methods). We investigated 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). We next performed MR analyses 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 (Teschendorff et al., 2015) 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 (GTEx Consortium, 2013) 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 (Hemani et al., 2016) 33. An adjusted P value that limited the FDR was calculated using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995) 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 I2=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 (r2<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 x 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>=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 (Joehanes et al., 2016) 9 and six were highlighted in a previous study as being associated with lung cancer (Baglietto et al., 2017) 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 (Fasanelli et al., 2015) 2. This previous study used methods which are sensitive to residual confounding and measurement error that may have biased results [Richmond et al. (2016); Hemani2017] 12, 35. These limitations are largely overcome using MR (Richmond et al., 2016) 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 (Burgess & Thompson, 2011) 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 repliacted, 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 (Shi et al., 2014) 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) (Freeman, Chu, Hsu, & Huang, 2016) 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 (Chen et al., 2017) 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 (Richmond et al., 2016) 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) (Gao, Zhang, Breitling, & Brenner, 2016) 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 readibility and/or setup.

In the main Rmd file

In Chapter ??:

Appendix B

The Second Appendix, for Fun

References

Baglietto, L., Ponzi, E., Haycock, P., Hodge, A., Bianca Assumma, M., Jung, C. H., … Severi, G. (2017). DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *International Journal of Cancer*, *140*(1), 50–61. `http://doi.org/10.1002/ijc.30431`

Banos, D. T., McCartney, D. L., Battram, T., Hemani, G., Walker, R. M., Morris, S. W., … Robinson, M. R. (2018). Bayesian reassessment of the epigenetic architecture of complex traits. *bioRxiv*, 450288. `http://doi.org/10.1101/450288`

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. `http://doi.org/10.1111/j.2517-6161.1995.tb02031.x`

Bojesen, S. E., Timpson, N., Relton, C., Smith, G. D., & Nordestgaard, B. G. (2017). AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax*, no pagination. `http://doi.org/10.1136/thoraxjnl-2016-208789`

Borghol, N., Suderman, M., Mcardle, W., Racine, A., Hallett, M., Pembrey, M., … Szyf, M. (2012). Associations with early-life socio-economic position in adult DNA methylation. *International Journal of Epidemiology*. `http://doi.org/10.1093/ije/dyr147`

Bowden, J., Smith, G. D., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, *44*(2), 512–525. `http://doi.org/10.1093/ije/dyv080`

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., … Smith, G. D. (2013). Cohort profile: The 'Children of the 90s'-The index offspring of the avon longitudinal study of parents and children. *International Journal of Epidemiology*, *42*(1), 111–127. `http://doi.org/10.1093/ije/dys064`

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*. `http://doi.org/10.1093/nar/gky1120`

Burgess, S., & Thompson, S. G. (2011). Avoiding bias from weak instruments in mendelian randomization studies. *International Journal of Epidemiology*. `http://doi.org/10.1093/ije/dyr036`

Chen, J., Behnam, E., Huang, J., Moffatt, M. F., Schaid, D. J., Liang, L., & Lin, X. (2017). Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. *BMC Genomics*. `http://doi.org/10.1186/s12864-017-3808-1`

Davey Smith, G., & Ebrahim, S. (2003). 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? `http://doi.org/10.1093/ije/dyg070`

Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, *23*(R1), R89–98. `http://doi.org/10.1093/hmg/ddu328`

Fasanelli, F., Baglietto, L., Ponzi, E., Guida, F., Campanella, G., Johansson, M., … Vineis, P. (2015). Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nature Communications*, *6*, 10192. `http://doi.org/10.1038/ncomms10192`

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., … Bray, F. (2013). GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase. No. 11 [Internet]., *11*, http://globocan.iarc.f. `http://doi.org/10.1016/j.ucl.2013.01.011`

Fewell, Z., Davey Smith, G., & Sterne, J. A. C. (2007). The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *American Journal of Epidemiology*. `http://doi.org/10.1093/aje/kwm165`

Fraser, A., Macdonald-wallis, C., Tilling, K., Boyd, A., Golding, J., Davey smith, G., … Lawlor, D. A. (2013). Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *International Journal of Epidemiology*. `http://doi.org/10.1093/ije/dys066`

Freeman, J. R., Chu, S., Hsu, T., & Huang, Y.-t. (2016). Epigenome-wide association study of smoking and DNA methylation in non-small cell lung neoplasms. *Oncotarget*, *7*(43). `http://doi.org/10.18632/oncotarget.11831`

Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N., Ardissino, D., … Sullivan, P. F. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*. `http://doi.org/10.1038/ng.571`

Gao, X., Zhang, Y., Breitling, L. P., & Brenner, H. (2016). Tobacco smoking and methylation of genes related to lung cancer development. *Oncotarget*. `http://doi.org/10.18632/oncotarget.10007`

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., … Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, *17*(1), 61. `http://doi.org/10.1186/s13059-016-0926-z`

GTEx Consortium, T. G. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–5. `http://doi.org/10.1038/ng.2653`

Haycock, P. C., Burgess, S., Wade, K. H., Bowden, J., Relton, C., & Smith, G. D. (2016). Best (but oft-forgotten) practices: The design, analysis, and interpretation of Mendelian randomization studies. *American Journal of Clinical Nutrition*. `http://doi.org/10.3945/ajcn.115.118216`

Hemani, G., Zheng, J., Wade, K. H., Laurin, C., Elsworth, B., Burgess, S., … Haycock, P. C. (2016). MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv*. Retrieved from `http://biorxiv.org/content/early/2016/12/16/078972.abstract`

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10), R115. `http://doi.org/10.1186/gb-2013-14-10-r115`

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., … Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. `http://doi.org/10.1186/1471-2105-13-86`

Huang, W. Y., Hsu, S. D., Huang, H. Y., Sun, Y. M., Chou, C. H., Weng, S. L., & Huang, H. D. (2015). MethHC: A database of DNA methylation and gene expression in human cancer. *Nucleic Acids Research*. `http://doi.org/10.1093/nar/gku1151`

Inoue, A., & Solon, G. (2010). Two-Sample Instrumental variables estimators. *The Review of Economics and Statistics*, *92*(3), 557–561. `http://doi.org/10.1016/0304-4076(90)90061-W`

Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., … London, S. J. (2016). Epigenetic Signatures of Cigarette Smoking. *Circulation: Cardiovascular Genetics*, *9*(5), 436–447. `http://doi.org/10.1161/CIRCGENETICS.116.001506`

Jones, P. a, & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews. Genetics*, *3*(6), 415–28. `http://doi.org/10.1038/nrg816`

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., & Storey, J. D. (2016). sva: Surrogate Variable Analysis.

Li, H. (2011). Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*. `http://doi.org/10.1093/bioinformatics/btq671`

Li, M., Zou, D., Li, Z., Gao, R., Sang, J., Zhang, Y., … Zhang, Z. (2019). EWAS Atlas: A curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Research*. `http://doi.org/10.1093/nar/gky1027`

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*. `http://doi.org/10.1038/ng.3643`

McKay, J. D., Hung, R. J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D. C., … Amos, C. I. (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature Genetics*, *49*(7), 1126–1132. `http://doi.org/10.1038/ng.3892`

Mill, J., & Heijmans, B. T. (2013). From promises to practical strategies in epigenetic epidemiology. *Nature Reviews Genetics*, *14*(8), 585–594. `http://doi.org/10.1038/nrg3405`

Munafò, M. R., Timofeeva, M. N., Morris, R. W., Prieto-Merino, D., Sattar, N., Brennan, P., … Davey Smith, G. (2012). Association between genetic variants on chromosome 15q25 locus and objective measures of Tobacco exposure. `http://doi.org/10.1093/jnci/djs191`

Pierce, B. L., & Burgess, S. (2013). Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, *178*(7), 1177–1184. `http://doi.org/10.1093/aje/kwt084`

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, *12*(8), 529–541. `http://doi.org/10.1038/nrg3000`

Relton, C. L., & Davey Smith, G. (2010). Epigenetic Epidemiology of Common Complex Disease: Prospects for Prediction, Prevention, and Treatment. *PLoS Medicine*, *7*(10), e1000356. `http://doi.org/10.1371/journal.pmed.1000356`

Relton, C. L., & Davey Smith, G. (2012). Two-step epigenetic mendelian randomization: A strategy for establishing the causal role of epigenetic processes in pathways to disease. *International Journal of Epidemiology*, *41*(1), 161–176. `http://doi.org/10.1093/ije/dyr233`

Relton, C. L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., … Davey Smith, G. (2015). Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *International Journal of Epidemiology*, *44*(4), 1181–1190. Retrieved from `http://dx.doi.org/10.1093/ije/dyv072`

Richardson, T. G., Zheng, J., Davey Smith, G., Timpson, N. J., Gaunt, T. R., Relton, C. L., & Hemani, G. (2017). Mendelian Randomization Analysis Identifies CpG Sites as Putative Mediators for Genetic Influences on Cardiovascular Disease Risk. *American Journal of Human Genetics*. `http://doi.org/10.1016/j.ajhg.2017.09.003`

Richmond, R. C., Hemani, G., Tilling, K., Davey Smith, G., & Relton, C. L. (2016). Challenges and novel approaches for investigating molecular mediation. *Human Molecular Genetics*, ddw197. `http://doi.org/10.1093/hmg/ddw197`

Shi, J., Marconett, C. N., Duan, J., Hyland, P. L., Li, P., Wang, Z., … Landi, M. T. (2014). Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nature Communications*. `http://doi.org/10.1038/ncomms4365`

Strathdee, G., & Brown, R. (2002). Aberrant DNA methylation in cancer: potential clinical interventions. *Expert Reviews in Molecular Medicine*. `http://doi.org/10.1017/s1462399402004222`

Teschendorff, A. E., Yang, Z., Wong, A., Pipinikas, C. P., Jiao, Y., Jones, A., … Widschwendter, M. (2015). Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. *JAMA Oncology*, *1*(4), 476–85. `http://doi.org/10.1001/jamaoncol.2015.1053`

Thomas, D. C., Lawlor, D. A., & Thompson, J. R. (2007). Re: Estimation of Bias in Nongenetic Observational Studies Using "Mendelian Triangulation" by Bautista et al. `http://doi.org/10.1016/j.annepidem.2006.12.005`

Wahl, S., Drong, A., Lehne, B., Loh, M., Scott, W. R., Kunze, S., … Chambers, J. C. (2017). Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, *541*(7635), 81–86. `http://doi.org/10.1038/nature20784`

Zhan, X., Hu, Y., Li, B., Abecasis, G. R., & Liu, D. J. (2016). RVTESTS: An efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics*, *32*(9), 1423–1426. `http://doi.org/10.1093/bioinformatics/btw079`

Zudaire, E., Cuesta, N., Murty, V., Woodson, K., Adams, L., Gonzalez, N., … Cuttitta, F. (2008). The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers. *The Journal of Clinical Investigation*. `http://doi.org/10.1172/JCI30024`

Abbreviations

EWAS - epigenome-wide assoctation study **GWAS** - genome-wide assoctation study **h<sup>2</sup>** - narrow-sense heritability **h<sup>2</sup><sub>SNP</sub>** - SNP-heritability **H<sup>2</sup>** - broad-sense heritability **MR** - Mendelian randomization **mQTL** - methylation quantitative trait loci **SNP** - single nucleotide polymorphism