Considering the impact of smoking on DNA methylation in *Understanding Society*

A. D. Andrayas

A dissertation submitted for the degree of Master of Science (by Dissertation)

Department of Biological Sciences

University of Essex

October 2017

Abstract

Smoking is a huge issue for social health and consequently there has been much research considering the relationship between tobacco use and many biological processes. One interesting field of study has identified epigenetic signatures of smoking using DNA methylation profiles. Up to this point, these studies were mostly carried out using the 450K BeadChip technology from Illumina. The new Infinium EPIC array is capable of quantifying DNA methylation at almost double the number of CpG sites and was used on whole blood from around 1200 participants of *Understanding Society*. This allowed integration of the household study's rich smoking-related data with DNA methylation levels spanning the entire genome using linear modelling. The R package *limma* was used for this and allowed the identification of novel, smoking-associated loci in this study that were differentially methylated between smokers and non-smokers. These regressions also revealed a decrease in the number, and thus significance, of probes differentially methylated with smoking in former smokers compared to current smokers, supporting the idea than these changes are reversed upon cessation. Additionally, this study showed that DNA methylation levels within smokers varied with increasing dosage whereby duration of tobacco use appears to be more important than intensity in driving changes to the methylome caused by smoking. Furthermore, this differential methylation was reversed once a person had quit smoking and the degree of this decay increased with cessation years. Taking these findings, it was then possible to create two quantifiable DNA methylation-based biomarkers of smoking capable of predicting both years spent smoking in current smokers and years since quitting in former smokers. This may then prove to be a useful tool in characterising disease risk given the number of differentially methylated loci located in important health-related genes, especially if DNA methylation is indeed related to their expression.

Acknowledgments

I would like to thank my supervisor, Professor Leonard Schalkwyk, for his many words of wisdom and all the friends and family who so readily encouraged me during the writing of this dissertation.

Table of Contents

Table of Figures

1. Introduction

1.1. Smoking Epidemiology

In 1964 the United States Surgeon General elucidated to the dangers of smoking in a landmark report (NIH, 1964). Since then an overwhelming amount of evidence has led to the classification of smoking as the leading, preventable cause of morbidity and mortality worldwide and has been frequently linked to advanced aging and multiple cancers (Peto et al., 2000). The first link between smoking and lung cancer came from a group of German scientists in the late 1920's and led to the first anti-smoking campaign ever (Proctor, 1996). However, this was largely based on ideology rather than science and was discredited following World War 2 which may in part explain why many people continued smoking thereafter. In the present day, smoking is declining in most of the developed world however there are still an estimated 1.1 billion tobacco users worldwide and this clearly presents a huge problem for public health (Doll et al, 2004). Therefore, an enormous amount of research has gone into the biology of tobacco smoking and ways to prevent its use. This work has revealed at least 98 chemicals found within tobacco smoke that are known to have specific toxic properties, with some directly leading to DNA damage and subsequently cancer (Kastan, 2008).

Research into the genetics of smoking has disclosed a predisposition to nicotine dependence and this has now been firmly established (Ware et al., 2012). Nicotine exerts an overwhelming addictive effect on tobacco smokers but there is some variation in the extent of addiction between individuals. Twin studies have documented a heritability to nicotine dependence which gives some explanation to inter-individual differences in ability to quit smoking (Ingebrigtsen et al., 2011). Here, genetic variants can influence this risk where single nucleotide polymorphisms within genes coding for receptors and enzymes involved in neurotransmitter metabolism have been implicated through genome-wide association studies. The strongest associations came from a locus mapped to chromosome 15q24-25 that harbours a gene cluster coding for nicotinic acetylcholine receptor subunits. This is made up of *CHRNA5-CHRNA3-CHRNB4*, genes that mediate fast signal transmission at synapses (Hällfors et al.,

2017). Moving on from this came gene-environment interaction studies, with the best example being the interaction of smoking with apolipoprotein E variants. The ε4 variant is a much poorer antioxidant than other variants and this reduced anti-oxidative resilience can lead to several diseases, namely cardiovascular disease, when in conjunction with the excessive oxidative stress of smoking (Stephens et al., 2003 and Isik et al., 2007). Recently a surge of interest in epigenetics has thus far focused on the epigenome's response to pollutant, such as that found in cigarette smoke, and looked for signals common in all smokers. This differs to genetic research focusing on individual differences between smokers. The strongest signals here have involved the altered epigenetic modification to genes involved in the response to tobacco combustion products, such as *AHRR*. This was so robust in fact that measures of DNA methylation at such loci are able to distinguish between current, nascent and never smokers (Philibert et al., 2012). This also strengthens previous findings that smoke-less tobacco use was not associated with differential methylation at any sites in snuff users (Besingi et al., 2013). With this said, another finding suggested that even nicotine itself can alter the methylation levels of promotors in GABAergic neurons, those involved in the neurotransmission of GABA, whose principal role is to reduce neuronal excitability throughout the nervous system although the effects were fairly minor in comparison (Satta et al., 2008).

The current understandings in smoking-related changes to the epigenome act to strengthen the link between genes and health-related phenotypes that is only possible through epigenetic research. Thus, using the wealth of smoking epidemiology study available, this project aims to build upon this and more specifically better comprehend how DNA methylation changes relate to smoking phenotype. Findings of this nature will allow the creation of a biomarker of smoking, a useful tool in disease risk management and prevention. Often self-reported data on smoking is unreliable and even accurate measures of environmental exposure can fail to fully account for the internal dose a person has been exposed to. Thus, using biological exposure measures can offer a much more sensitive and reliable biomarker that is more accurately related to the final health outcome of a smoker.

1.2. DNA Methylation

Many biomarkers of tobacco smoke exposure have been utilized to characterize its biological effects and assess the subsequent impact on disease risk. One example is that of serum cotinine levels, an objective measure of nicotine exposure using a cutoff point of 14ng/mL but is one of many well-established biomarkers that fail to fully reveal the effects of past exposure. Epigenetic study however offers the potential for an early detection, sensitive and long term biomarker where measures of exposure, decades prior to the sample collection, can be observed (Zhang et al., 2016). This has enabled much better research into exposure induced risk of chronic diseases as it can reflect exposure to a variety of environmental factors linked to ill health and help in better understanding the underlying aberrant biology of smoking.

Epigenetics is defined as heritable changes that can affect gene architecture and expression without altering the genetic sequence itself. The three pillars of epigenetic regulation are DNA methylation, histone modifications and non-coding RNA species. Modifications of these affect almost all nuclear processes, including gene transcription and silencing, DNA repair and replication and telomere function to name a few. DNA methylation is one such mechanism involving the addition of a methyl (CH_3) group to DNA, often modifying the function of the genes at which it occurs (Lister et al., 2009). The addition of CH_3 at a 5-carbon of a cytosine ring, creating 5'-methylcytosine (5-mC), is the most widely characterized occurrence of any epigenetic mechanism. In mammalian DNA, this almost always occurs at 5'-CpG-3' dinucleotides, known as CpG sites. An exception of this occurs in embryonic stem cells where a lot of 5-mC is observed outside of CpG contexts (Ramsahoye et al., 2000).

Within the human genome, CpG sites located inside of clusters, termed CpG islands, are mostly unmethylated whereas other CpG sites remain largely methylated. This acts to, at least in part, separatee the genome into transcriptionally active and inactive zones. CpG islands make up approximately 1-2% of the genome and around 50-60% of all genes contain a CpG island, encompassing gene promoters or exons most of the time. These are related to gene expression and when CpG islands in the promotor region becomes methylated, expression can be repressed. Noted exceptions to this include imprinted genes and those on the inactive X chromosome (Moore et al., 2012). Methyl groups are added to DNA by a family of enzymes called DNA methyltransferases (DNMTs) that catalyze the transfer of methyl groups from S-

adenosylmethionine. DNMT1 is necessary to maintain already established DNA methylation patterns, whereas DNMT3a and DNMT3b seem to be required for the establishment of new or *de novo* DNA methylation patterns. Furthermore, deletion of any DNMT is lethal in murine and human cells, showing the indispensable functions methylation plays in mammals (Bestor et al., 2000).

In conjunction with this comes DNA demethylation, the removal of the methyl group. This process is required for epigenetic reprogramming of genes and has been implicated in some disease mechanisms, such as tumor progression. It was previously thought that DNA demethylation only occurs passively through dilution of methylation marks via *de novo* DNA synthesis by DNMT1. Today it is known that methylation marks can in fact be actively erased through the direct removal of the methyl group, or through a combination of the two (Ohno et al., 2013). In mammals, direct excision of 5mC paired with G does not seem possible, so instead the methylated base undergoes sequential modifications that are converted by ten-eleven translocation enzyme-mediated oxidation. This family of 5-mC hydroxylases include TET1, TET2 and TET3 and may promote DNA demethylation by binding to CpG rich regions, preventing DNMT activity. They work by producing 5-hydroxymethylcytosine (5-hmC) as the first intermediate and then further hydroxylating this intermediate to 5-formylcytosine (5-fC) and then 5-carboxylcytosine (5-caC). Thymine DNA glycosylase (TDG) can also directly excise 5-fC, allowing the subsequent base excision repair (BER) pathway to convert the modified cytosine back to its unmodified state (Bochtler et al., 2016). The biological significance of 5'-methylcytosine has been widely recognized and may reflect a global decrease of DNA methylation. This is likely a consequence of methyl-deficiency caused by a number of different environmental influences, including smoking, and quantification of global 5-mC could act as a molecular marker for disease (Robertson, 2005). Furthermore, a more recent study has shown small changes to levels of intermediate DNA methylation may be associated with complex disease phenotypes where these changes caused a cascade of events leading to altered glucocorticoid receptor (*NR3C1*) protein (Leenen et al., 2016). This is just two examples of a plethora of papers demonstrating not only the importance of DNA methylation but also how its role in disease phenotypes may be more far-reaching than previously thought.

1.3. Previous Findings

The first discovery of a smoking-related DNA methylation marker came from Breitling et al (2011) and has since been followed by the identification of thousands of individual CpG sites with differential methylation between smokers and non-smokers. Smoking has also been linked to a small global decrease in DNA methylation as a whole (Ambatipudi et al., 2016). These sites span all 23 chromosomes of the human genome and have varying degrees of methylation changes when compared to those who have never smoked. The significance of the top hits is staggering and in terms of this and effect size, the mean difference in methylation values between smokers and non-smokers, the strongest signals are those located in *AHRR* and the 2q37.1 region. Furthermore, the large majority of significantly associated sites show reduced methylation levels in smokers. A notable exception to this are the strong positive effect sizes seen in *MYO1G*. This gene is a plasma membrane-associated class I myosin, abundant in T and B lymphocytes and mast cells and aids in cell elasticity. Thus, *MYO1G* could be associated with smoking-related fibrosis in several tissues (Olety et al., 2010).

A large majority of smoking-related CpG sites are located within gene bodies. These were identified through epigenome-wide association studies (EWAS), and epidemiological studies have worked to further strengthen these findings and create biologically plausible associations with ill health. One way this has been done is by comparing DNA methylation at certain loci with disease phenotypes. A good example of this came with the discovery of the first smoking associated CpG site, cg0363183, located in the body of gene *F2RL3*, the coagulation factor II receptor-like 3 gene. The main function of this gene is to code for thrombin protease-activated receptor-4 (PAR-4). PAR-4 plays a role in platelet activation and cell signaling and is expressed in several tissues, including leukocytes and lung tissue. Thus, this could give some explanation as to why *F2RL3* methylation is found to be related to risks of cardiovascular diseases (CVD), lung cancer and even mortality (Zhang et al., 2015) although DNA methylation's role in this is still not well characterized. Perhaps stronger and more consistent associations have been made to CpG sites located in the genetic region of the aryl hydrocarbon receptor repressor (*AHRR*). This gene has been established as a possible tumor suppressor and it is suggested that smoking may affect the aryl hydrocarbon receptor when

tobacco smoking triggers the generation of polycyclic aromatic hydrocarbon (PAHs). These toxic chemicals can exert their effect through the aryl hydrocarbon receptor (AhR) and its binding partner, called aryl hydrocarbon receptor nuclear translocator (ARNT). AHRR competes with ARNT for binding to the AhR and can repress signal transduction. This then gives meaning to alterations in the methylation status and consequentially expression of the *AHRR* gene, suggesting its role as a mediator of PAH detoxification. Thus, *AHRR* may be involved in the metabolism of endogenous toxins found in cigarette smoke (Evans et al., 2008).

Other identified probes were in fact not located in the transcriptional regions of a gene and instead observed in other locations in the genome which may not directly impact the coding sequence of the gene. DNA methylation at these loci is often more tightly linked to transcriptional silencing than elements further downstream, whose methylation does not always impact the magnitude of gene expression (Brenet et al., 2011). Two examples of such sites, known to be associated with smoking, are cg19859270, located in the 1st exon of G-protein coupled receptor 15 (*GPR15*) and several in the intergenic region of 2q37.1. The loci on chromosome 2q37.1 are adjacent to an alkaline phosphatase gene cluster. One of these genes, *ALPPL2*, is responsible for dephosphorylation of many proteins and nucleotides and is beneficial as a biomarker for many cancers, having already been well established as a tumor marker in ovarian and testicular cancers and seminoma (Albrecht et al., 2004). However, this is only reliable in non-smokers as ALPPL2 enzyme serum concentrations have been found to increase up to tenfold in cigarette smokers (Schmoll et al., 2004) but nevertheless this still hints at an underlying mechanism by which DNA methylation changes increase smokers risk to cancer. Additionally, smoking associations of *GPR15* sites were first reported in a study by Wan et al. (2012), with suggestions about its correlation with current and long-term smoking. Afterwards, Tsaprouni et al. (2014) also showed that this gene was the only gene at the time that showed a clear trend of increased gene expression in smokers compared to non-smokers, and a negative correlation between gene expression and DNA methylation. This study then assumed that the decrease in DNA methylation at cg19859270 within *GPR15* seen in smokers would likely lead to an increase in transcription. This finding was confirmed in two papers, published by Bauer et al. (2015) and Koks et al. (2015). This gene regulates T-cell migration and immunity and thus might explain its role in chronic inflammatory diseases and as a

HIV co-receptor. Further to this, *GPR15* was also reported in interactions with ethnicity-dependent differentially prevalence of HIV, namely HIV2 in African Americans (Dogan et al., 2015), raising the possibility that, for certain loci, differential methylation could reflect a shift in blood cell mixture.

Other strong signals have also been seen in the 5' untranslated regions (5'UTR) of the *PRSS23* and *RARA* genes, although their biological role in smoking has not thus far been well characterized. 5'UTRs are cis-regulatory elements required to regulate translation. *PRSS23* codes for serine protease 23 and is a member of the trypsin family. Trypsin is formed when the proenzyme form of trypsinogen, produced by the pancreas, is activated thus proposing a role for these enzymes in smoker phenotypes when differentially methylated and also their impact in pancreatitis of which smoking is a risk factor (Lankisch et al., 2015). *RARA* codes for retinoic acid receptor alpha. This is a nuclear receptor that transduces retinoid signaling alongside the retinoid X receptor (RXR) forming RXR/RAR heterodimers. In the absence of ligand these heterodimers repress transcription by recruiting co-repressors. If ligand binds to the complex, a conformational change is induced that allows recruitment of histone acetyltransferase co-activators. This rearrangement of the *RARA* gene is a feature of acute promyelocytic leukemia (Vitoux et al., 2007).

Clearly a huge range of sites have been implicated in smoking-induced DNA methylation change, all with varying functions and importance in disease. As suggested in the studies mentioned, the genetic context of individually differentially methylated sites can help better understand pathophysiological processes that are activated or suppressed by changes in DNA methylation caused by smoking (Breitling et al., 2011). Consequently, such sites may contribute to a greater comprehension of smoking exposure by expounding smoking-related DNA methylation signatures or even help build a picture of an epigenetic mechanism that may lead to smoking-induced disease. In fact many of the mentioned sites have already been used in the construction of a reliable quantitative approach, with high specificity, for differentiating between the smoking status of individuals and this has been validated. One study created a predictor model of smoking exposure using bisulphite pyrosequencing of just four genomic loci that were differentially methylated between smokers and non-smokers. Combining these sites into a DNA methylation index gave a strong and positive prediction for previous smoking with an area under the curve (AUC) of 0.83 (Shenker et al., 2013).

However, given the huge variation of duration, intensity and years of cessation within smokers and ex-smokers, using sites identified through a simply comparison of smokers and non-smokers will not fully divulge the consequences of DNA methylation changes and this project aims to build a more comprehensive methylation signature that takes this into account.

When looking at dosage effects, much of the research up to this point in time has concentrated on the impact of pack years of smoking on the degree of DNA methylation changes within smokers. Pack years are calculated by multiplying the number of packs (20 cigarettes) smoked per day by the number of years smoked, and has been used to quantify lifetime cumulative exposure. This approach has achieved success, with one study by Zhang et al. (2015) disclosing a relationship between pack years and *AHRR* methylation in both current and lifetime exposure to tobacco smoke and even in smoking related mortality outcomes. Another by Ambatipudi et al. (2016) found that DNA methylation values fell in four loci within IER3 with an increasing number of pack years. However, there is some skepticism about this pack year model and some discrepancy in lung cancer incidence for those with the same amount of pack years. Two individuals may both have smoked for 20 pack years but if one of these persons has been smoking 0.5 packs per day for 40 years, they would likely have a hugely different risk compared to another person smoking 20 packs a day for 1 year even though in theory they have the same cumulative lifetime exposure. With this said, it would be better to include duration and intensity of smoking separately in any analyses looking into dosage effect (Peto et al., 2012). By doing so, some studies have found that intensity of smoking may be less proportional to incidences of smoking related disease than duration of smoking alone and genes differentially methylated in current smokers relative to never smokers are often significantly associated with duration of smoking as well (Ambatipudi et al., 2016). This suggests there may be more fundamental parameters, for instance age of onset or short periods of cessation, that should be taken into consideration to better characterize differential methylation and any successive health outcomes caused by tobacco use (Peto, 2012). One study supporting this theory found that an earlier age at smoking onset correlated with hypermethylation of *RASSF1A* which leads to a poor prognosis in primary non-small cell lung cancer (Kim et al., 2003). This all shows that further study of loci within other smoking related genes, and more complex modeling of smoking history is therefore necessary to explore precise dose response relationships with

DNA methylation and to give a more global understanding of smoking exposure in the hope to better understand the molecular mechanisms at play and this is another aim of this project.

Figure 1.3.1. shows a comparison of effect sizes for the top 90 CpG loci reported in more than five studies from Joehanes et al. (2016), the largest analysis of epigenetic signatures of smoking to date. This paper made use of 16 independent cohort studies. This figure compares the average effect size between current and never smokers against former and never smokers and shows that a consistent pattern of differential methylation, in the same direction, is observed for former smokers as it was seen for current smokers although to a much smaller degree. In general, this suggests a reversibility of smoking-related changes to the epigenome upon smoking cessation. This finding was also confirmed in many other studies including those by Ambatipudi (2016), Guida (2015), Tsaprouni (2014) and Lee (2016) where they all demonstrated patterns of reversibility in DNA methylation level changes. Thus, this suggests that alterations in DNA methylation caused by active smoking are site specific, dynamic and reversible and this can be seen in the varying degrees of differences between current and former smoking DNA methylation values in Figure 1.3.1. The clinical significance of smoking cessation has been well established and is known to reduce the increased risk smokers have of developing many diseases and cancers. It has been demonstrated that the ratio of lung cancer between current and former smokers increases sharply with time since quitting and it

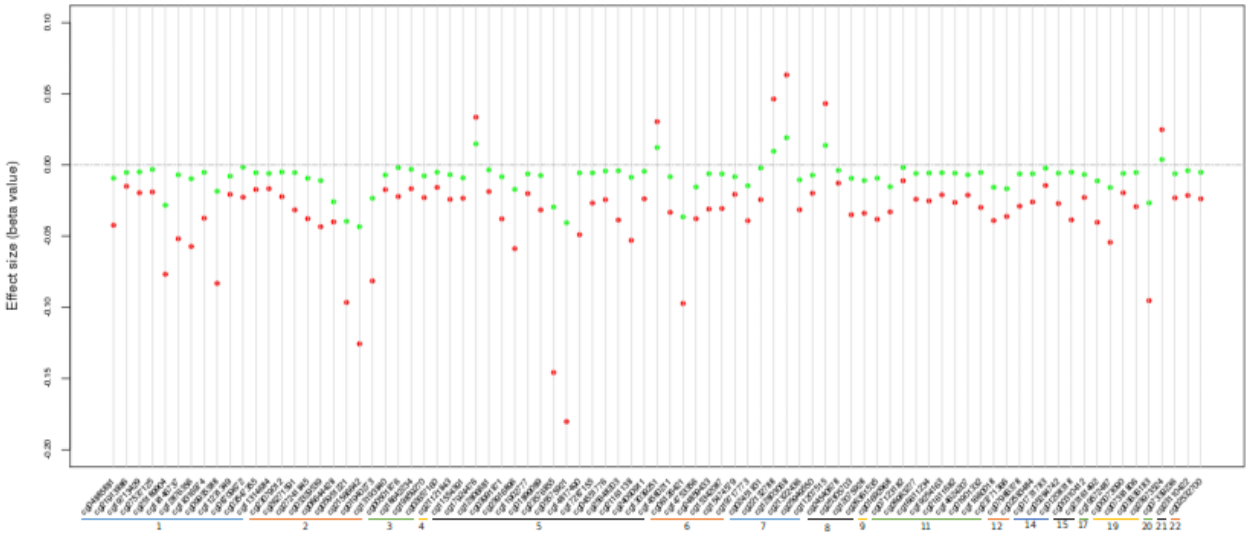

Figure 1.3.1.: Effect size distributions of the top 90 differentially methylated probes from Joehanes et al. (2016).Probes are ordered in relation to the chromosome it is located. All 90 sites were identified in more than five studies, from Joehanes et al (2016) and effect size here is a measure of difference in Beta-values between smokers and non-smokers for both current (red) and former (green) smokers.

is suggested that by stopping before middle age almost 90% of lung cancer risk attributed to tobacco use can be avoided later on in life (Peto et al., 2000). This effect remains even after adjusting for number of pack years (Vlaanderan et al., 2013). Furthermore, one paper showed an association of duration of smoking cessation with a significantly reduced risk of CIMP high colorectal cancer and suggested that quitting smoking induces a protective effect on the DNA methylation–related carcinogenesis pathway that leads to this cancer (Nishihara et al., 2013). However, the degree of this reduced risk may be over-estimated in many studies where risk is calculated by dividing the nearly constant smoker risk rate by the increasing non-smoker rate. When smoking ceases the rate of lung incidence does stops increasing steeply but still increases with age, much like most cancers where risk is often higher in older people (Peto et al., 2011). Some studies have also shown that it can take up to 20 years for some sites to reach full "methylation recovery" (Guida et al., 2015 and Zhang et al., 2014). Although demonstrating the reversibility of methylation marks is of great interest, it does not help explain what causes this site specificity, or the variation in DNA methylation. One aim of this project is to give a more in depth look into the decay of the DNA methylation signature. The hope is that this may garner a better idea of smoking induced methylome changes and the environmentally driven process that causes it. This, along with more complex dosage effect studies, may enable a quantification of smoking exposure that can better predict related disease risk.

Although the pathological mechanisms underplaying many of these sites is poorly understood, they still pose benefits in their correlation with smoking exposure. In fact, there has been efforts made to use methylation at one CpG site, cg05575921 within *AHRR*, as a quantifiable biomarker for smoking cessation. Here this site was seen to be sensitive and specific to smoking status with a characteristic, receiver operated area under the curve (AUC) of 0.99 and this kind of study is ongoing (Philibert, 2015). A measurable increase in methylation at this site, taken at regular time points, could be used to help physicians accurately monitor time since quitting with the hope that this would encourage successful cessation of smoking (Philibert et al., 2016).

Several studies have found other findings in their mining of smoking data that should be considered when studying variation in smoking-induced DNA methylation. Elliott (2014), Joehanes (2016), Lee (2016),

Zaghlool (2015) and Zhu (2016) all identified ethnic heterogeneity of smoking-related differential methylation patterns at several loci, such as cg05575921 in *AHRR* where differences in methylation were much higher in South Asian participants than those of a European origin. However, this may partly be down to generally higher statistically powered studies in European cohorts compared to other EWASs. Another study by Besingi et al (2013) investigated both tobacco and snuff smoking and showed that smokeless tobacco was not significantly involved in DNA methylation, indicating that the majority of epigenetic alterations may be caused by the burnt products of tobacco and not its basic components. Additionally, ppassive smoking may also impact DNA methylation. A recent paper has found that high levels (more than 10 hours per week) of recent indoor second hand smoke exposure may be associated with lowered DNA methylation of AHRR in human monocytes, albeit with weaker associations than active smoking (Reynolds et al., 2016). All these factors will add to the differences in DNA methylation levels seen in smokers and thus can help produce a more reliable and highly applicable epigenetic signature of smoking.

1.4. Understanding Society and EPIC Array

This study will make use of a recently created genome-wide DNA methylation resource as part of the UK Household Longitudinal Study (UKHLS), *Understanding Society*. This has been funded primarily by the Economic and Social Research Council (ESRC). *Understanding Society* builds on the success of the British Household Panel Survey (BHPS) that was heavily used by researchers, generating hundreds of scientific publications. However, *Understanding Society* aims to support a wider range of research than BHPS and this DNA methylation resource aims to help in doing so (Buck and McFall, 2011). Longitudinal studies of this nature can provide understanding of the trajectories of individual life histories and this project focuses on the detailed recorded smoking information available from participant surveys. This comprised of sex, age, smoking status, number of cigarettes smoked per day, age when starting smoking and age when last smoked. The data used throughout this project was collected in the main survey in wave 3. Furthermore, upon blood collection, approximately 5 months after the main survey questionnaire, nurses also asked participants three more smoking-related questions which included if they had smoked within the past 30 minutes, 24 hours and how many hours since last smoking. This will enable a more complex smoking

history to be elucidated that takes into account even short-term smoke exposure and is not limited to comparisons of smokers and non-smokers as is the case for many previous studies. Instead, models can be created that look into intensity, duration and years of smoking cessation and these factors alone could yield some interesting results that had not yet been fully established.

Another huge benefit of *Understanding Society* comes with its longitudinal nature. The participants used in this study were also involved in the BHPS before joining with *Understanding Society*. Participants were visited once a year and thus complete data on the smoking status and number of cigarettes smoked by them is available for at least 9 consecutive years, from wave I through to R of BHPS, then followed by wave 2 and 3 of *Understanding Society*. This allows a more precise smoking history to be deduced such as the differences in consistent, low smoking effects and those with lots of variation in the number of cigarettes smoked. Thus, *Understanding Society* will provide a great resource in the analysis of smoking and DNA methylation marks in this project.

2. Preprocessing and Modeling

2.1. EPIC Methylation Array

The participants used in this study were selected as they were part of BHPS before it was integrated into *Understanding Society*. This meant that the longest smoking histories prior to sample collection could be obtained and thus allowed the most comparisons to be made between this data and DNA methylation levels. To create this DNA methylation resource whole blood samples, taken from around 1200 participants in the wave 3 nurse visit of *Understanding Society* approximately 5 months after the main survey, were used.

Most studies detailed thus far have made use of BeadChip technology in Illumina methylation arrays to quantify methylation at thousands of genomic loci at single-nucleotide resolution due to their low cost and high-throughput capabilities. These studies on smoking-associated DNA methylation have used the older 450K array, however in this study methylome sequencing was carried out using the new Infinium Methylation EPIC BeadChip, with a coverage of over 850,000 CpG sites, almost double the size of its predecessor. These sites include over 90% of sites found in the 450K array and offer improved coverage of regulatory elements (Pidsley et al., 2016). The DNA from the samples was prepared, and the arrays processed, by Jon Mill's laboratory at Exeter University using the protocol detailed by the manufacturer. The technology used in such microarrays first involves a bisulfite conversion of the genomic DNA to convert unmethylated cytosine to uracil which is then subjected to whole genome amplification (WGA) using hexamer priming and Phi29 DNA polymerase. The DNA is then enzymatically fragmented and purified primers, enzymes and dNTPs are then applied to a chip. This chip contains two bead types for each CpG locus and each bead type is attached to a single stranded 50-mer DNA oligonucleotide that differ in sequence at the free end, making them allele specific. One bead type corresponds to the methylated cytosine and the other to the unmethylated cytosine which after conversion to uracil is amplified as thymine in previous steps (Weisenberger et al., 2008). The fragmented DNA products are then denatured to single strands and hybridized to the chip via allele specific annealing to either the methylation specific probe or the non-methylation probe. This step is followed by single-base extension with hapten labeled

dideoxynucleotides where ddCTP is labeled with biotin and the others (ddATP, ddUTP and ddGTP) are labeled with 2,4-dinitrophenol (Steemers et al., 2006). At this point multi-layered immunohistochemical assays are performed by repeatedly staining with a combination of antibodies that differentiate between the two types. After, the chip is scanned to obtain intensities of the unmethylated and methylated bead types (Bibikova et al., 2011). The system further analyzes this microarray data to normalize the raw data and reduce experimental variation effects (Staaf et al., 2008).

2.2. Pre-processing

Pre-processing, quality control and normalization were carried out in the statistical environment of R (R Core Team, 2017). The bioconductor package *bigmelon* was used which has many methods for working with Illumina arrays. This package extends the capabilities outlined by Pidsley et al in the R package *wateRmelon* by adapting methods from the *gdsfmt* package for efficient memory use and management, overcoming the overheads associated with data handing in R (Gorrie-Stone et al., 2017).

The entire dataset of 1187 samples was first normalized using the **dasen** function developed by Leonard Schalkwyk for the *wateRmelon* R package. This works by normalizing the methylated (M) and unmethylated (U) probe intensities. By doing so any technical variation can be more simply dealt with by adjusting these rather than the derived "raw" beta values which are the methylation level estimates calculated in the Illumina protocol with little normalization and adjustment. It also involves a combination of background adjustment of the M and U intensities and four separate, between-array quantile normalizations of methylated Type I, unmethylated Type I, methylated Type II and unmethylated Type II intensities (Pidsley et al., 2013). To further elucidate any samples which are grossly affected by this process, the function **qual**, also by Schalkwyk, is used to assess the degree of which the normalized and raw beta values differ. It calculates and outputs the root mean square deviation, sum of squared differences (SSD), sum of absolute squared differences and root mean square error (RMSD) for each sample. A cutoff of >0.05 RMSD or >0.05 SSD was used as this appeared to allow good identification of samples which

were very obviously altered when normalised. Using this, 8 outliers were identified with large differences in their raw and normalized values, one of which had a drastically larger RMSD and SSD.

After elucidating samples that had normalized badly, meaning those that showed discrepancies between the original and normalized intensities, the function **outlyx** was utilised to further compute data-outliers using a subset of probes from the large data set. This was developed by Tyler Gorrie-Stone for the *wateRmelon* package and first involves specifying the number of inter-quantile ranges to be discriminated from the upper and lower quantiles. This is identified from principal component analysis and in this case, was 2. These computed principal components are used to determine distance measures for each observation. Then, weights for location and scatter outliers are computed based on these distances and the combined weights are used to determine outliers (Filzmoser et al., 2008). In this case an arbitrary threshold of 0.15 for the final weight output was used meaning outliers were defined as samples with a combined weight of less than 0.15. Here, 6 outliers were identified and these were also observed in the previous step using qual. These were thus removed from further analysis alongside the other two samples that had normalized poorly.

The next step of quality control checked for sample quality with **bscon**, a function developed for *wateRmelon* by Louis El Khoury, Eilis Hannon and Leonard Schalkwyk. This function uses the green and red channel readings of the type I and type II bisulfite conversion data to return the median bisulfite conversion percentage value for each array. This quantity shows average conversion of unmethylated cytosine to uracil and is an important step in pre-processing the data as complete conversion is necessary for further study. It uses the intensities of sample-dependent controls included in the BeadChip to evaluate performance across arrays. Type I chemistry beta values are calculated by first dividing the first three control probes of the green channel and the second three control probes of the red channel by the sum of all six of these probes and the unconverted green channel probes (U4, U5, U6) and red channel probes (U4, U5, U6). Type II chemistry beta values are calculated by simply dividing the methylated red channels by the sum of methylated red and unmethylated green channels. This outputs a percentage value for bisulfite conversion. In general, most arrays achieved over 85% conversion however 4 samples had noticeably lower values than the rest of the data set and these were also removed from further analysis. It is important to note

here that updates to the EPIC array manifest removed control probes C6 and U6 but this thus far does not seem to hinder the function.

Next, an R implantation of Horvath's clock (Horvath, 2013), named **agep**, was used to predict the age of the samples. It forms a weighted average for measures of DNA methylation at 353 epigenetic markers on the human genome that were elucidated using an elastic net regression. This results in a linear regression model whose coefficients correspond to transformed age and this is used to predict "DNA methylation age" by plugging in the relevant beta values. Unfortunately, the Horvath clock was created using the older 450K microarray and thus 17 of the 353 CpG probe sites used in the DNA methylation-based age predictor are absent in the newer EPIC array but nevertheless prediction was still fairly accurate. However, one sample had a large age discrepancy of almost 30 years. Another quality check used was to visualize sex differences between samples. This was done by plotting principal component 1 against principal component 2 and showed that the sex of all participants was correctly matched. Finally, by plotting raw intensities per rack any obvious batch effects between the different plates are revealed. This showed some differences between plates and these also had varying methylation level (Beta-value) distributions suggesting some technical variation. It is then important to bear this in mind when carrying out downstream analyses and thus these have been counted for within the linear models, although normalization with **dasen** seems to correct for most of this variation. Furthermore, one sample shows very low log2 methylated intensity and this was also identified in previous quality control steps and thus removed from further analysis.

In total, 12 samples were removed from the dataset based on poor normalization and low bisulfite conversion during the making of the DNA methylation profile. This left a total of 1175 samples of good quality and that act normally when transformed. These were used in downstream analyses in conjunction with the smoking data. However, when creating smoking variables to be used in further analyses, 4 samples gave negative values for duration and a further 138 had incomplete smoking data and therefore these were also removed from the dataset to allow for the most comparative analyses. This left 1033 samples with reliable descriptive information to be used in linear models. When carrying these out, the ratio between the now quality checked methylated and unmethylated intensities were used to obtain an estimate of the

methylation level for each probe. This is calculated by dividing the methylated intensity with the sum of unmethylated and methylated intensity plus 100 (*M/U+M+100*). This is called a Beta-value where a value of 0 is equal to non-methylation, 1 equal to total methylation and a 0.5 value suggests one copy is methylated but not the other in the diploid human genome. Logit transformed Beta-values, termed M-values, have been used in the analyses outlined in this chapter. Beta-values have been shown to have high heteroscedasticity for sites that are highly methylated or unmethylated, meaning that the ability of a linear regression model to predict dependent variables are not consistent across all values of the dependent variable. Therefore, M-values are used throughout this project when elucidating differentially methylated probes and regions as these perform much better in terms of detection rate and true positive rate and their performance can be improved even further by applying a minimum threshold of difference whereas Beta-values cannot (Du et al., 2010). Beta-values were only used when looking at effect sizes, which in this case constitutes the difference in average methylation between two groups.

2.3. Linear Models

Linear models were carried out in the statistical environment R, using the Bioconductor software package *limma*. This package allows differential methylation analysis of large-scale microarray data and the identification of differentially methylated CpG sites. Although originally created for gene expression analyses, it has proved to be a valuable tool for studying DNA methylation arrays of which the package's linear modelling strategy works well and has been used by other studies (Ambatipudi et al, 2016). In this case, the package operates on a matrix of methylation values, the calculated and quality controlled M-values, where rows represent a probe for each genomic feature, here these are CpG sites, and each column represents the participant sample. The *limma* function **lmFit** then fits a linear model to each row of data, taking into account a specified design matrix that details relevant information related to each sample array, and specifies the hypothesis to be tested. Within this study, the treatment-contrasts parametrization method was used to construct design matrices using the **model.matrix** function. This includes a coefficient for the comparison of interest itself rather than extracting the contrast after using a contrast matrix. Although the data is from two colour oligonucleotide arrays, linear modeling that compares DNA methylation between

two groups is effectively the same as analysis of variance (ANOVA) or multiple regression except that a model is fitted for every probe (Ritchie et al., 2015).

For assessing differential methylation, *limma* uses an empirical Bayes method to moderate the standard errors of the estimated log-fold change and is implemented in the **eBayes** function.. This enables more stable inference and better statistically powered analysis. This is because the statistical methods implemented by the **lmFit** function can produce imprecise results when sample sizes are small but the data is of a high dimension, as with microarray data. Thus, the Empirical Bayes (EB) technique can offer gains in performance by leveraging information from the entire dataset, across the CpG sites in this case, when making assumptions about individual probes. EB in *limma* moderates genewise, or probe-wise, variance estimators by assuming a Bayesian hierarchical model for these variances and estimates the prior distribution from the marginal distribution of the inputted data itself rather than basing it on prior knowledge (Smyth et al., 2005, Phipson et al., 2016). The resulting fitted models, which have borrowed information across probes, were summarized using the **topTable** function to obtain a list of probes most likely to be differentially methylated between the groups specified in the design matrix. The resulting object also shows the *p*-values adjusted using the Benjamini-Hochberg method to control the expected false discovery rate (FDR) as well as the the corresponding log2-fold-change, average log2-expression for each probe over all arrays and channels and *t* and *F* statistics (Ritchie et al., 2015). The large majority of functions used in package were authored by Gordon Smyth.

In short, *limma* allows the construction of both simple baseline and complex experimental design matrices where both categorical and continuous variables may be handled in much the same way and where all included variables can be studied at the same time. It also expands this using its empirical Bayes method to average variability of DNA methylation over all CpG sites used in the comparisons of interest leading to hopefully truer variances. This makes it an accessible, powerful R package for differential methylation analysis.

2.4. Kernel Smoothing

To better understand the association of individual CpG sites found to be differentially methylated between smokers and non-smokers via *limma*, differentially methylation regions (DMRs) were identified using the Bioconductor package *DMRcate*. DMRs are stretches of DNA within the genome that involve a minimum of 2 adjacent sites or groups of sites in close proximity that have different methylation patterns between samples. DMRs are often of more value than single CpG sites that are not contextualized by the methylation status of neighbouring probes. *DMRcate* identifies these using kernel smoothing, a statistical method to estimate a function reflecting DMRs as the weighted average of neighbouring differential methylation signals identified using the same design matrices and corresponding M-values used within *limma.* These CpG sites are first annotated with their chromosome position and test statistic through the function **cpg.annotate**. This makes the method agnostic to other site annotations other than spatial ones and passes the square of the moderated *t* statistic, calculated for each EPIC probe to the next stage in the *DMRcate* protocol. To control for multiple testing an FDR cutoff of 0.05 was used to specify which CpG sites are individually significant, as done before. This is also used to index default thresholding in the following steps (Peters et al., 2015). The resulting object is then passed to **dmrcate**, the main function of the package, which compares two smoothed estimated per chromosome, one weighted with the test statistic and one not, for null comparison, to identify significantly differentially methylated regions. The recommended values for lambda, the number of nucleotides for the gaussian kernal bandwidth used in smoothed-function estimation, and C, the scaling factor for bandwidth, was used and set to 1000 and 2 respectively. This meant that half a kilobase represents 1 standard deviation of support which has been shown to hold near optimal prediction of sequencing-derived DMRs for 450K experiments. No empirical testing has yet been done for the best parameters when using EPIC array data and thus it is necessary to be cautious about results from this analysis or risk type I errors. Lambda here also informs the DMR bookend definition where gaps greater than or equal to lambda between significant CpG sites will be in separate DMRs. The Gaussian kernel is calculated where lambda over C is equal to sigma.

This approach then first applies standard linear modelling to the data and then applies gaussian smoothing to the resulting CpG sites test statistics using a given bandwidth. It then models these smoothed test statistics using the method outlined by Satterthwaite (Satterthwaite, 1946) and computes *P* values based on this model. These are then adjusted and a threshold employed, giving FDR-corrected significantly associated sites and those nearby each other are finally agglomerated using the specified bandwidth. The resulting DMRs were then obtained using the **extractRanges** function that takes the dmrcate output to create a GRanges object, annotating the DMRs to promotor overlaps using the specified reference genome hg19 (Peters et al., 2015).

3. Qualitative Smoking

To determine the differences in the DNA methylation of smokers, ex-smokers and non-smokers, whole blood was collected from participants of *Understanding Society*. Participants were then divided into three qualitative categories; never, former and current smokers, based on self-report data collected in the wave 3 questionnaires. For the purpose of these analyses a sample of participants were taken from each category to create three equal sized groups of 175 each and the general characteristics of those who contributed this data are shown in Table 3.1. This was done to prevent spurious associations that might occur from differing sample sizes of the smoking strata. At the time of blood collection, the 525 participants used were aged between 28 and 96, with a mean age of 57.04 ± 14.88. As the methylation profiles were carried out using whole blood, DNA methylation based estimates of the leukocyte subpopulations were calculated using the **estimateCellCounts** function from the *minfi* R package. This implements the reference-based algorithm created by Houseman and team (2012) using DNA methylation profiles from purified leukocyte samples by which the algorithm performs linear constrained projection (CP) to calculate the distribution of white blood cells within a sample. This showed that those who had quit smoking had lower proportions of CD8T cells and a higher proportion of monocytes than both current smokers and those who had never smoked. Smokers and former smokers also had slightly higher proportions of granulocytes and CD4T cells, and current

Table 3.1.: General characteristics of participants used in qualitative smoking analyses

Characteristic	All	Never	Former	Current
Sample size	525	175	175	175
Sex (Male:Female)	228:297	68:107	95:80	65:110
Age, years (Mean±SD)	57.04±14.88	56.79±13.98	62.37±15.56	51.95±13.20
CD8T (Mean±SD)	0.073±0.041	0.075±0.045	0.067±0.038	0.079±0.039
CD4T (Mean±SD)	0.123±0.059	0.119±0.054	0.123±0.064	0.126±0.059
NK (Mean±SD)	0.040±0.037	0.047±0.042	0.046±0.037	0.028±0.030
Bcell (Mean±SD)	0.052±0.027	0.051±0.025	0.051±0.029	0.054±0.026
Mono (Mean±SD)	0.040±0.021	0.040±0.019	0.042±0.023	0.036±0.019
Gran (Mean±SD)	0.686±0.084	0.682±0.080	0.685±0.089	0.691±0.081

smokers had much lower levels of natural killer cells than both never and former smokers. Given the effect smoking has on innate immunity, it makes sense to see such changes in white blood cell distributions (Mehta et al., 2008).

3.1. Current vs Never Smokers

To start considering the relationship between smoking and methylation within *Understanding Society*, a sample of 175 current smokers and 175 never smokers were first compared. To do so, linear regression analysis was used within the *limma* package in R to identify changes in DNA methylation levels between these two categories, coded as 0 for current smokers and 1 for never smokers. To regress out known confounders, age, sex, blood process day and counts for CD4+ and CD8+ T cells, natural killer cells, B cells, monocytes and granulocytes were included in the model. Furthermore, batch effects were observed between samples on different plates during pre-processing and thus this information was also included in the model although, as stated earlier, normalization of the data corrected for most of these differences.

To control for type 1 errors, or "false positives", the FDR-controlling procedure was used at a false discovery rate of 0.05 (Benjamini and Hochberg, 1995). All adjusted *P*-values below this point were deemed as significant and this revealed 5198 differentially methylated probes (DMPs) between current and never smokers. These consisted of 3058 hypomethylated and 2140 hypermethylated CpG sites spanning the whole genome at varying methylation states and degrees of difference. Of these, 3838 probes were annotated to 2640 genes found in the UCSC database, leaving 1360 unannotated sites contained within intergenic regions. The average effect size, or methylation difference, of each of these probes, along with their significance and chromosome location, are summarized in Figure 3.1.1. A more stringent cutoff of Bonferroni 5% level (Dunn, 1961), which in this case with 866,895 genomic loci being tested and compared is 5.76×10^{-8}, revealed 610 probes that were differentially methylated. The effect sizes seen in the DMPs between current and never smokers ranged from -0.256 to +0.154, with a huge surplus of negative over positive directions of change. This is in line with previous findings. These hypomethylated sites also tended to have the strongest significance values and largest effect sizes. Further still, looking at the average

methylation difference of all >850,000 probes, more than half were hypomethylated in smokers. A simple and interesting explanation of this could be the activation of a number of "clean up" systems to aid in the removal of harmful toxins found in cigarette smoke. If DNA methylation does indeed impact the gene expression and resulting protein products than hypomethylation of such genes would cause this activation.

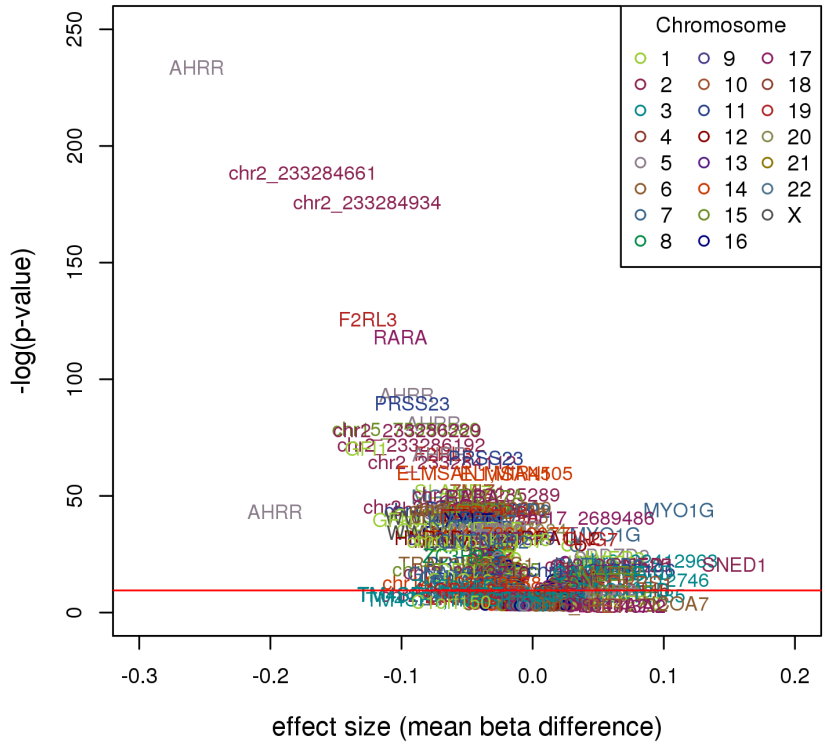

Figure 3.1.1.: Summary of 5198 significant CpG sites differentially methylated between current and never smokers at false discovery rate *P* < 0.05. Red line indicates Bonferroni threshold. Each CpG site is represented by significance as shown by their –log(P-value) values (y-axis) and effect size and direction (x-axis), the mean β value difference between groups. Associations are colour-coded in reference to the chromosome the CpG site is located on.

Over one quarter of the identified DMPs are located in non-coding regions of the genome which consist largely of repetitive DNA, including many tandem repeats such as mini- and microsatellites. These make up more than 50% of the genome and CpGs in these regions have been shown to remain mostly methylated. This includes transposable elements (TEs) that may change chromosome locations and alter the genetic identity of cells. TEs are often hypermethylated, silencing transcription of the TE-encoded enzymes needed for their transposition. DNA methylation thus acts to protect the genome's integrity and loss of DNA methylation at these regions thus may reactive transposable elements (Levin and Moran, 2011). Given that

DNA hypomethylation most commonly arises outside of promotors, in repetitive elements and peri-centric DNA, chromosomal rearrangements may play a large role in the aberrant nature of methylome changes and indeed that of smoking (Ehrlich, 2008). In fact, global genome-wide hypomethylation has also been seen to induce genome instability and is one of the earliest molecular abnormalities seen in cancer. It has also been described following carcinogen exposure and in cells with altered differentiation and proliferation states (Issa et al., 1999 and Lisanti et al., 2013).

Cigarette smoke exposure itself is one of the most powerful modulators of DNA methylation and may aid in the understanding of DNA hypomethylation seen in smokers. Firstly, carcinogens found in cigarette smoke, such as arsenic and polycyclic aromatic hydrocarbons (PAHs) can lead to DNA damage and the recruitment of DNMT1 where CpG sites adjacent to repaired nucleotides become methylated (Cuozzo et al., 2007). Secondly, nicotine can affect gene expression when it binds to nicotinic acetylcholine receptors, activating cAMP response element-binding protein. This pathway has been shown to downregulate DNMT1 (Satta et al., 2008). Thirdly, smoking may indirectly impact DNA-binding factors which in turn prevent *de novo* methylation of CpG sites in these motifs (Han et al., 2001). Lastly, cigarette smoke is known to induce hypoxia where carbon monoxide binds to haemoglobin and decreases oxygenation of many tissues. This then caused HIP-1α-dependent upregulation of methionine denosyltransferase 2A, known to synthesis S-adenosylmethionine, the major methyl donor used for DNA methylation (Liu et al., 2011). This together can help understand the cause of altered epigenetic landscapes found in smokers within this study and the large supply of hypomethylated sites.

Hypomethylation of some genes also makes sense biologically, for example the strongest smoking-associated CpG site cg05575921 is located in the *AHRR* gene on chromosome 2 which acts to detoxify toxins found in cigarette smoke. In this study a total of 39 DMPs were situated in this gene, including some novel probe associations from the EPIC array. These sites had FDR adjusted *P*-values ranging from 3.06×10^{-102} to 4.18×10^{-2} and all were featured in the gene body, between the ATG and stop codon, with the exception of cg26954197 found in the 3'UTR of *AHRR*, between the stop codon and poly A signal. This has been well established in a number of papers (Shenker et al., 2013, Philibert et al., 2012). DNA

methylation in gene bodies has even been associated with altered DNA transcription and is common in ubiquitously expressed genes (Suzuki and Bird, 2008). Additionally, the effect size, a measure of mean difference in beta values between current and never smokers, ranged widely from -0.256 to 0.061. This suggests CpG sites within *AHRR* become both hypermethylated and hypomethylated and may be explained with context within the gene as the hypomethylated sites tended to be closer to CpG islands which are known to be unmethylated for the most part.

Table 3.1.1.: Top 10 hypomethylated and hypermethylated DMPs associated with smoking (current vs never smokers)

Illumina Probe ID	Chr Number	Chromosome position (bp)	Design Type	UCSC Gene Name	UCSC Gene Region	Present in 450K Array	FDR Adjusted P Value	Mean Effect Size
Hypomethylated in current smokers								
cg05575921	5	373378	I	*AHRR*	Body	TRUE	3.06E-102	-0.26
cg21566642	2	233284661	I			TRUE	1.24E-82	-0.17
cg01940273	2	233284934	II			TRUE	3.89E-77	-0.13
cg03636183	19	17000585	II	*F2RL3*	Body	TRUE	2.17E-55	-0.13
cg17739917	17	38477572	II	*RARA*	5'UTR	FALSE	4.98E-52	-0.10
cg21161138	5	399360	II	*AHRR*	Body	TRUE	3.22E-41	-0.10
cg14391737	11	86513429	II	*PRSS23*	5'UTR;Body	FALSE	1.18E-39	-0.09
cg25648203	5	395444	II	*AHRR*	Body	TRUE	7.07E-36	-0.08
cg18110140	15	75350380	II			FALSE	6.96E-35	-0.10
cg22812571	2	233286229	II			FALSE	1.03E-34	-0.10
Hypermethylated in current smokers								
cg12803068	7	45002919	II	*MYO1G*	Body	TRUE	8.92E-20	0.11
cg05009104	7	45002980	II	*MYO1G*	Body	FALSE	2.82E-15	0.06
cg13039251	5	32018601	II	*PDZD2*	Body	TRUE	2.53E-11	0.06
cg15542713	1	42385581	II	*HIVEP3*	TSS1500	TRUE	8.29E-11	0.06
cg04414766	3	22412963	II			FALSE	1.85E-10	0.09
cg24049493	1	42385941	II	*HIVEP3*	TSS1500	TRUE	5.68E-10	0.05
cg22635676	2	241975971	I	*SNED1*	Body	TRUE	1.12E-09	0.15
cg08035323	2	9843525	II			TRUE	1.14E-09	0.06
cg26718213	2	241976080	II	*SNED1*	Body	TRUE	1.49E-09	0.08
cg11207515	7	146904205	II	*CNTNAP2*	Body	TRUE	1.54E-09	0.05

The top ten hypomethylated and hypermethylated probes, all showing more than a 5% difference in DNA methylation between current and never smokers, are shown in Table 3.1.1. These include well known smoking-associated methylation differences in the *AHRR*, *F2RL3* and *RARA* as well as those in the

intergenic region of 2q37.1. Here, both probes seen in studies using the older 450K array and new probes unique to the EPIC array were differentially methylated in these genes and thus their strong links with smoking has not only been replicated but strengthened by this study. The top 20 DMPs also included less reported genes such as the hypermethylation of CpG sites in *HIVEP3*, located on chromosome 1, but little is known about its role in smoking phenotypes. It is however a member of the human immunodeficiency virus type 1 enhancer-binding family of proteins and binds the recognition signal sequence for recombination of immunoglobulin and T-cell receptor gene segments (Mak et al., 1998) suggesting a role in immunity which is impacted by smoking. It is also of note that hypermethylated sites had much smaller significance values than the top hypomethylated loci despite their equally large effect sizes.

The second most commonly observed annotated gene, second only to *AHRR*, was *ZMIZ1*. With 18 smoking-related DMPs ranging in FDR adjusted *P*-values from 3.03×10^{-23} to 4.61×10^{-2} and effect sizes ranging between -0.042 and 0.061 it is definitely of interest. The association of *ZMIZ1* with smoking has been previously identified (Besingi et al., 2013) and its encoded protein has been shown to regulate the activity of some transcription factors, including the androgen receptor (AR). *ZMIZ1* interacts with the transactivation domain of AR and has been found to augment its transcriptional activity in human prostate cancer cells and moreover, it co-localizes with SUMO-1 to enhance sumoylation of AR. Thus, decreased methylation of *ZMIZ1* may decrease AR-mediated transcription and lead to a number of downstream consequences (Sharma et al., 2003). One study found that smokers had a significantly higher mean number of CAG repeats within the AR gene and lower levels of testosterone than non-smokers. It also found that the sperm of those who smoke had lower motility and increased morphological defects (Mitra et al., 2012).

ZMIZ1 is also a known diabetes susceptibility gene, a disease of which smoking is a risk factor, with suggestions that tobacco use may cause increased inflammation and even directly impact insulin resistance (Xie et al., 2009). Other DMPs were located in *ANPEP*, another gene linked to diabetes in which methylation was significantly associated with gene expression, suggesting that the four smoking DMPs located in this gene may then be linked to altered gene function. *ANPEP* codes for the alanine aminopeptidase enzyme, a membrane-associated peptidase involved in a broad range of cellular processes

which could be impacted by smoking (Ligthart et al., 2016). The presence of smoking-related DMPs in these two genes offers support for the increased risk of diabetes many smokers have and hints at the potential biological mechanisms effected in those with diabetes.

Of the 5198 significant CpG sites associated with smoking at FDR < 0.05, 2780 were unique to the Illumina EPIC array. Given the increased coverage of this BeadChip, this offers huge potential for the identification of novel associations between smoking and DNA methylation changes. Within the annotated sites, this includes many genes that were not previously covered in the older 450K array. The novel gene with the largest number of DMPs was *GNMT* on chromosome 6, seen in 8 probes differentially methylated between current and never smokers. Probe targets to the *GNMT* gene ranged little in their significance, from 1.12×10^{-4} to 4.73×10^{-2}. Although these may not be as strongly linked with smoking as other sites, these probes were located in gene regions of particular interest. Most were seen within 200 bases upstream of the transcription start site (TSS) or just downstream within the 5'UTR of the first exon. As stated previously, DNA methylation of sites in these components may be more closely linked to the transcription of genes like *GMNT* than those further downstream and thus are more likely to impact gene expression. *GMNT* codes for an enzyme called glycine N-methyltransferase which acts to regulate the ratio of S-adenosylmethionine (SAM), the methyl donor for DNMTs, and S-adenosylhomocystein (SAH). It does so by competing with tRNA methyltransferases for SAM, increasing SAH levels which in turn acts as an inhibitor for tRNA methyltransferases, reducing methylation. Thus, *GNMT* is important for maintaining cellular homeostasis and deficiency of this enzyme has been linked to liver disease and hepatomegaly (Luka et al., 2009). All 8 *GNMT* CpG sites became hypermethylated in smokers with relatively small effect sizes where differences in mean Beta-values between smokers and never smokers ranged from 0.009 to just 0.003. Hypomethylated *GNMT* could be linked to the aberrant and widespread hypomethylation seen in smokers which may be exaggerated if this gene became over expressed. Furthermore, *GNMT* has been suggested to detoxify environmental carcinogens such as polyaromatic by binding to these hydorcarbons and inhibiting the formation of DNA adducts (Yen et al., 2013) although this does not help explain why this is seen to be hypermethylated in smokers. However, given the difference in DNA methylation seen in smokers the

associations of *GNMT* with tobacco use make sense biologically and present a novel insight into the underlying mechanism causing such changes.

Stronger novel associations were observed in probes cg00045592 (FDR 2.71 x 10^{-22}) and cg04009575 (FDR 7.56 x 10^{-6}) which were not present in older microarrays by Illumina. These are located in CpG sites within the 5'UTR of the signalling lymphocytic activation molecule (*SLAMF7*) gene on chromosome 1. Expression levels of *SLAMF7* have been shown to correlate with smoking behaviour (Charlesworth et al., 2010) but its association with smoking-induced methylation changes has not been shown until now. This gene is a receptor present on immune cells and is known to be expressed on multiple myeloma cells. This finding lead to the production of elotuzumab, an anti-SLAMF7 antibody used in its treatment and works by enhancing natural killer cell activation leading to cell cytotoxicity of myeloma cells. *SLAMF7* also mediates both the inhibitory and activation effects on natural killer cells depending on the expression of its adaptor, sarcoma-associated transcript 2 (EAT-2). The association of this gene with smoking may then relate to the large amount of research on the consequences of cigarette smoke inhalation on inflammation and immune suppression (Guo et al., 2015 and Lee at al., 2012). Methylation change at *SLAMF7* may then in part relate to the drastically lower proportions of natural killer cells observed in current smokers compared to both never and former smokers within this study and the impact this would have on their innate immune system.

The genetic contexts of some novel probes identified in this study have already been previously reported in papers. These studies used the chromosomal coordinates of unannotated smoking DMPs, for instance those lying in within intergenic regions, to obtain their nearest corresponding gene symbol. However, given the wider coverage of the EPIC array compared to the previous 450K, confirmation of DNA methylation changes at some of these genes has been established in this study and strengthens their role in smoking-induced epigenetic landscapes. One such gene is *ELMSAN1* (Ambatipudi et al., 2016). This gene, located on chromosome 14, contained at least 6 of the identified smoking DMPs and codes for the ELM2 And Myb/SANT Domain Containing 1 protein. A histone deacetylase complex (HDAC) called the MiDAC complex was identified fairly recently and given its name, Mitotic Deacetylase Complex, after higher levels

of this complex were observed in cells arrested in mitosis. This complex contains histone deacetylase 1 and 2, a MIDEAS corepressor protein, otherwise known as ELMSAN1 and also DNTTIP1. The ELM-SANT domain units act to scaffold this HDAC and given that protein lysine acetylation plays a key role in controlling gene expression, *ELMSAN1* is thus also involved in transcription binding and activation (Itoh et al., 2015). Furthermore, HDACs can be targeted by selective histone deacetylase inhibitors to find those with anti-cancer and anti-inflammatory properties and present a therapeutic avenue of research (Bantscheff et al., 2011). This finding, along with the altered DNA methylation state of the *ELMSAN1* gene in smokers, may strengthen the role of epigenetic regulation in disease, especially in those caused by tobacco use.

The 610 DMPs meeting the Bonferroni genome-wide significance threshold were located in roughly 305 genes. To better understanding the biological mechanisms implicated by smoking, these were inputted into the STRING database. This web tool looks for association networks between the gene's protein products based on evidence of fusion, co-expression, co-occurrence, presence in the same neighbourhood or from text mining. This showed 134 interactions between the encoded proteins at the highest confidence threshold of 0.9, and this association network is shown in Figure 3.1.2.

STRING also shows any KEGG pathways in which the inputted proteins are involved and these relationships are displayed within the interaction network. One example includes the eight proteins of *GNG7*, *GNG12*, *GNAQ*, *KCNQ1*, *AKT3*, *CACNA1D*, *ITPR1* and *ADCY9* which are involved in the cholinergic synapse. The corresponding neurotransmitter, acetylcholine, plays a critical part in brain maturation and is detectable even before neurulation, much like nicotinic acetylcholine receptors (nAChRs). Acetylcholine also later promotes the transition from replication to differentiation and thus modulates neuronal development by promoting or preventing apoptosis. This continues into adolescence, a time most smokers begin using tobacco (Slotkin et al., 2004). Nicotine enhances cholinergic synaptic transmission by activating presynaptic nAChRs that then increase presynaptic calcium concentration which in part explains the excitatory effects of nicotine on the central nervous system and its highly addictive nature. Additionally, nicotine impacts cholinergic signalling within key nodes of the reinforcement circuitry and learning. This suggests differential methylation at these loci thus may be a product of nicotine and its

indirect promotion of acetylcholine and dopamine neurotransmitter release as most of these genes are also implicated at the dopaminergic synapse. *AKT3*, *KCNQ1*, *CACNA1D*, *GNAQ* and *ADCY9*, along with *SCN7A*, *MYH6*, *CACNA2D2* and *CACNA2D4* were also implicated in the adrenergic signalling in the cardiomyocyte pathway. Acute sympathetic stimulation of cardiac adrenergic receptors (ARs) is necessary for the appropriate output of the heart but when this becomes chronic a number of complications may arise that is detrimental to health and may induce apoptosis and cardiomyocyte hypertrophy (Lohse et al., 2003).

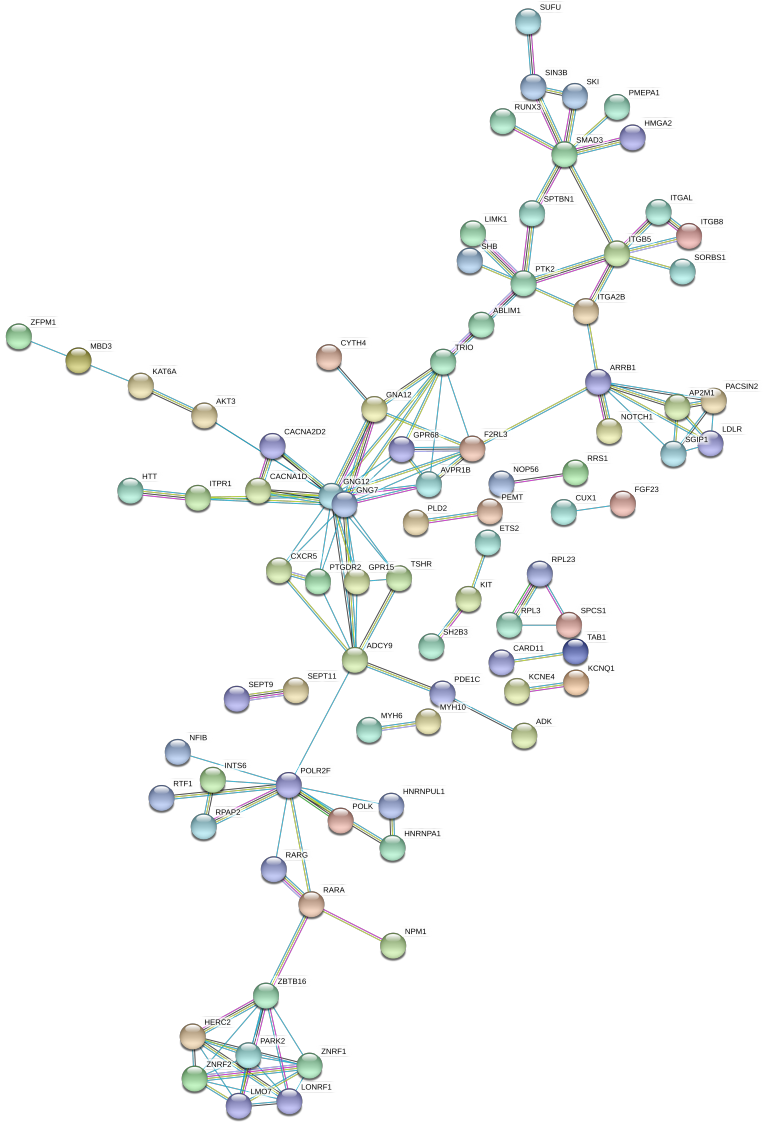

Figure 3.1.2.: STRING interaction network of the 134 interacting proteins encoded by the genes enclosing smoking-associated DMPs seen at a Bonferroni genome-wide significance threshold. Only shows interactions that meet the highest confidence threshold, with an interaction score of 0.9. Lines are coloured by the type of evidence the interaction is based on. Red - presence of fusion, Green - neighborhood evidence, Blue - cooccurrence, Purple -experimental, Yellow - textmining evidence, Light blue - database, Black – coexpression.

Decreased methylation of such sites may be linked to the increased risk of heart disease found in smokers. DNA methylation changes may then be able to reflect the signals stimulated by nicotine and tobacco use.

Another enriched KEGG pathway is that of platelet activation in which genes *ITGA2B*, *AKT3* and *F2RL3* are involved. Smoking is a major risk factor for coronary thrombosis and seems to impact the hemostatic process that maintains circulatory integrity after vascular injury. One way this occurs is through altered platelet function. Two distinct pathways occur during thrombus formation which may initiate platelet activation. The first occurs through exposure to subendothelial collagen, resulting in adhesion of platelets to the site of injury. The second involves tissue factors that establish a proteolytic cascade that produces thrombin. Thrombin then cleaves and activates receptors on the surface of platelets. Activated platelets in turn drive further thrombus formation. Cigarette smoke contributes to both pathways and the resulting clots seen in smokers appear to be more resistant to thrombolysis. Two key factors that contribute to the activation of these pathways is the free radical-mediated oxidative stress and loss of NO protection associated with smoking (Barua and Ambrose, 2013). Thus, differences in methylation in the *ITGA2B*, *AKT3* and *F2RL3* genetic loci of smokers may be related in the process of platelet activation when responding to the environmental stress of cigarette smoke.

With this said however, the biological processes affected by smoking are certainly not limited to the synapses and platelets and these are just two of the most strongly enriched pathways. The fact that just the top 664 genome-wide significant DMPs inputted into the database can show the severe and far-reaching impact of smoking on health suggests that changes in DNA methylation are important biomarkers of such diseases. It then makes sense that so many KEGG pathways, ranging from regulation of the actin cytoskeleton to Huntington's disease, be associated with smoking-related genes given the widespread impact it has on the methylome.

3.2. Current vs Former Smokers

A similar trend of smoking-induced methylome changes were seen when comparing same-sized groups of current and former smokers whereby 826 probes were differentially methylated at a cutoff of 0.05 after FDR adjustment for multiple testing. 137 DMPs also remained significant after bonferroni correction, with a cutoff of 5.76×10^{-8}. Figure 3.2.1. summarizes the adjusted *P*-values and effect sizes for the loci seen in this comparison, showing 513 hypomethylated and 313 hypermethylated CpG sites. Again, the strongest signals were seen in the *AHRR*, *RARA*, *F2RL3* genes, the 2q37.1 region and others that have been well characterized (Joehanes et al., 2016). However, for all of these sites, both their significance values and effect sizes are much smaller when using former smokers rather than current smokers, suggesting some reversal of DNA methylation change upon cessation.

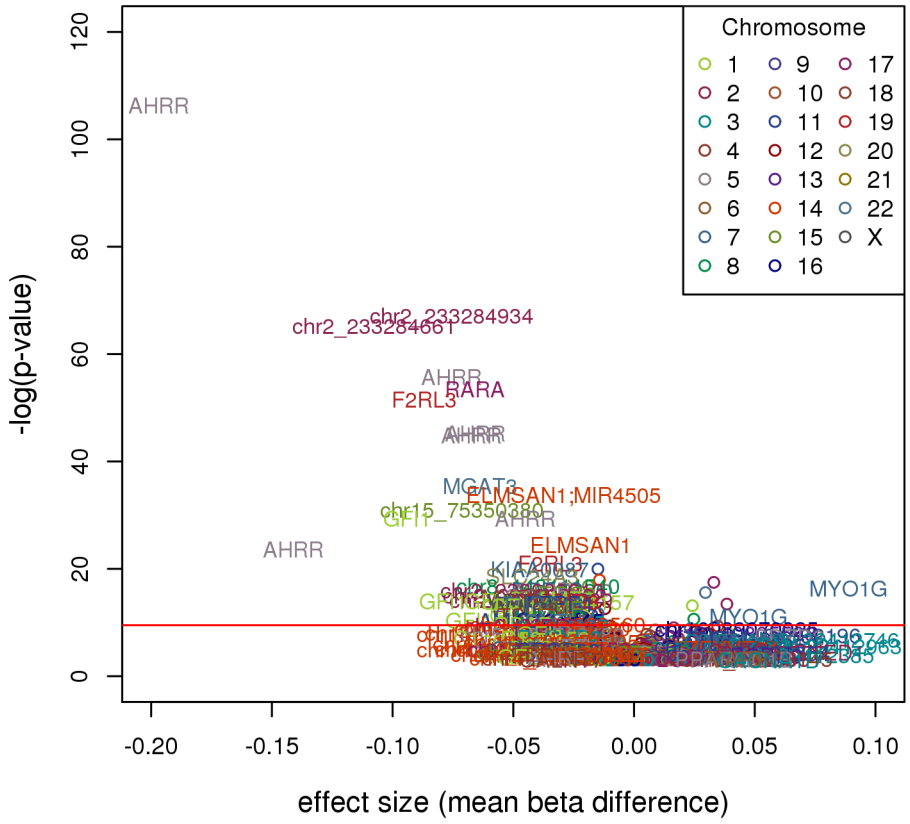

Figure 3.2.1.: Summary of 826 significant CpG sites differentially methylated between current and former smokers at false discovery rate P < 0.05. Red line indicates Bonferroni threshold. Each CpG site is represented by significance as shown by their –log(P-value) values (y-axis) and effect size and direction (x-axis), the mean β value difference between groups. Associations are colour-coded in reference to the chromosome the CpG site is located on.

Of the total 826 identified DMPs, 222 were not observed, i.e. were not statistically significantly differentially methylated, when comparing the same current smokers to never smokers. As seen in Figure

3.2.2., 191 of these sites showed greater differences in DNA methylation between current and ex-smokers than there was between current and never smokers. Here, sites that were hypomethylated in current smokers compared to never smokers are in fact hypermethylated in former smokers when compared to the same never smokers and vice versa. For these sites, perhaps smoking cessation not only caused a reversal of smoking-related changes but this may go so far as to surpass non-smoker levels. For the remaining 31 sites, little difference is observed in the methylation levels between current, former ad never smokers and effect sizes for these sites remain around zero. This suggests that these sites simply met the FDR threshold when comparing current and former smokers by chance.

In general, sites where this phenomenon was the most obvious were also the sites with the strongest effect sizes in the current and never smoker observation. This suggests that these sites are more susceptible to changes in DNA methylation as a response or reflection of smoking status. A notable outlier is cg11025972 located on chromosome 13. This site has a near zero effect size between current and never smokers but showed obvious differences in DNA methylation when comparing current and former smokers as well as former and never smokers. Thus, this site had similar Beta-values for current smokers and non-smokers but was hypermethylated in former smokers. This site lied within an unannotated region of the genome so, using the *humarray*

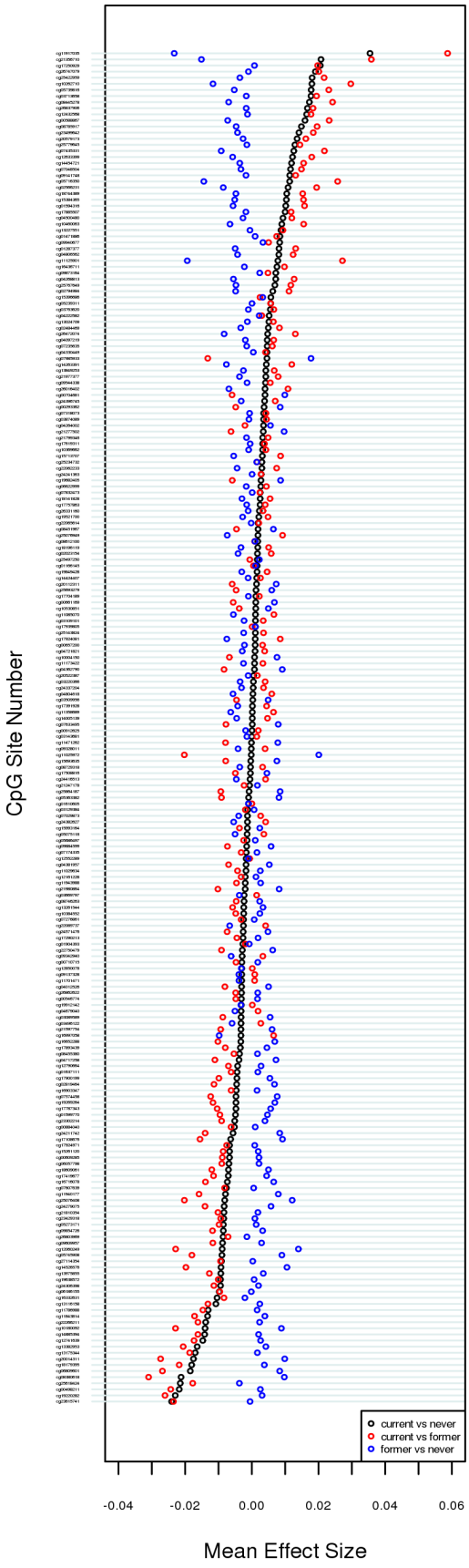

Figure 3.2.2.: Summary of effect sizes for 222 DMPs observed when comparing current and former smokers but not between current and never smokers.

package (Cooper, 2017), the closest gene symbol was found to be *SOX1*. This codes for a transcription factor involved in the neurogenesis of the central nervous system. It is particularly expressed in the striatum, a major component of the reward system, receiving glutaminergic and dopaminergic inputs (Guth & Wegner, 2008). This system is hugely impacted when a person ceases smoking and hypermethylation of this region may be related to the withdrawal symptoms ex-smokers experience.

3.3. Former vs Never Smokers

The decreased number of significantly differentially methylated sites seen when comparing current and former smokers as opposed to current and never smokers suggests that smoking-induced methylation is reversible upon smoking cessation and this has been shown in several studies (Tsaprouni et al., 2014 and Guida et al., 2015). To further test this, another model was run to compare methylation between the 175 former and 175 never smokers used previously. This revealed 17 CpG sites with significant DNA methylation differences, with only 7 meeting the Bonferroni significance threshold (Figure 3.3.1.). All of

Table 3.3.1.: 17 DMPs associated with former smoking (former vs never smokers)

Illumina Probe ID	Chr Number	Chromosome position (bp)	Design Type	UCSC Gene Name	UCSC Gene Region	Present in 450K Array	FDR Adjusted P Value	Mean Effect Size
cg14391737	11	86513429	II	PRSS23	5'UTR;Body	FALSE	1.30E-15	-0.07
cg21566642	2	233284661	I			TRUE	3.67E-13	-0.07
cg06644428	2	233284112	I			TRUE	2.02E-11	-0.05
cg01940273	2	233284934	II			TRUE	2.67E-11	-0.05
cg05575921	5	373378	I	AHRR	Body	TRUE	6.13E-11	-0.06
cg03636183	19	17000585	II	F2RL3	Body	TRUE	5.12E-05	-0.04
cg22812571	2	233286229	II			FALSE	7.74E-05	-0.05
cg00475490	11	86517110	II	PRSS23	5'UTR;Body	FALSE	6.43E-03	-0.02
cg16047567	1	12664243	II	DHRS3	Body	TRUE	7.95E-03	-0.02
cg01692968	9	108005349	II			TRUE	1.90E-02	-0.03
cg18110140	15	75350380	II			FALSE	1.90E-02	-0.03
cg17739917	17	38477572	II	RARA	5'UTR	FALSE	1.90E-02	-0.04
cg16841366	2	233286192	II			FALSE	2.41E-02	-0.05
cg25189904	1	68299493	II	GNG12	TSS1500	TRUE	2.41E-02	-0.04
cg15342087	6	30720209	II			TRUE	2.41E-02	-0.02
cg15420926	8	132929147	II	EFR3A	Body	TRUE	3.27E-02	-0.01
cg23771366	11	86510998	II	PRSS23	TSS1500	TRUE	3.45E-02	-0.03

these sites were also seen in the current vs never smoker model and this small number of significant hits suggest that almost all change to the epigenome caused by smoking becomes insignificant when quitting, especially as the effect sizes of the 17 DMPs were minimal, ranging from -0.069 to -0.014. Furthermore, no site had a higher significance than 1.30×10^{-15} after FDR adjustment as shown in Table 3.3.1.

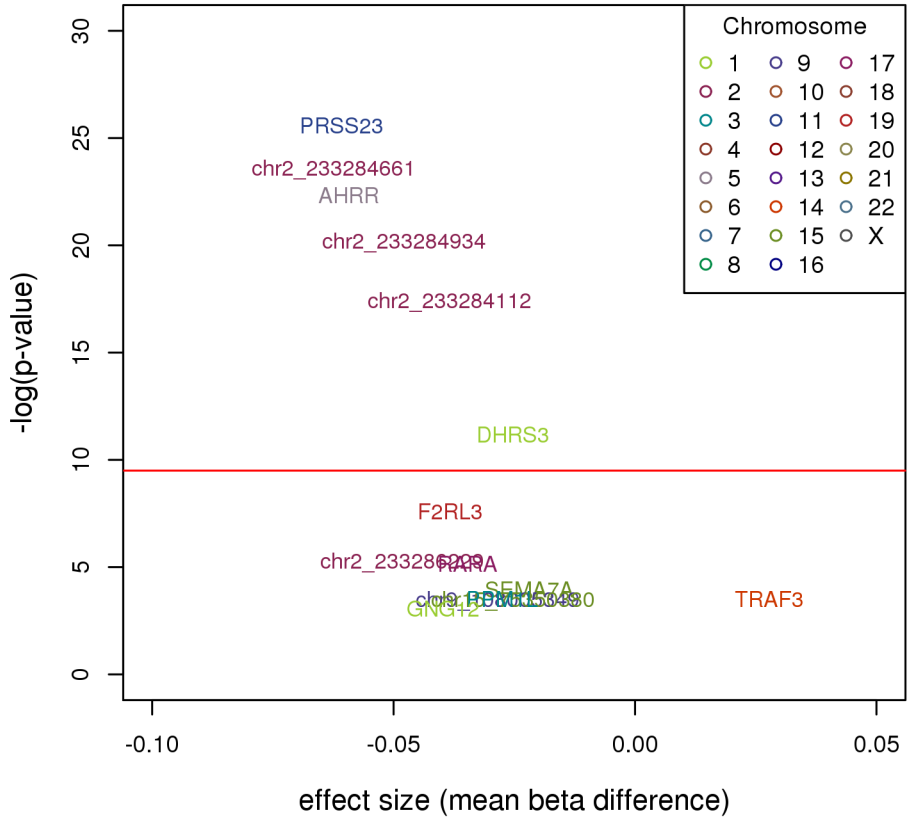

Figure 3.3.1.: Summary of 17 significant CpG sites differentially methylated between former and never smokers at false discovery rate P < 0.05. Red line indicates Bonferroni threshold. Each CpG site is represented by significance as shown by their –log(P-value) values (y-axis) and effect size and direction (x-axis), the mean β value difference between groups. Associations are colour-coded in reference to the chromosome the CpG site is located on.

However, this does show that there is a long-lasting effect on DNA methylation at certain sites that remains even once smoking has ceased. Also, given that all but one DMP was hypermethylated in former smokers, this suggests that DNA hypomethylation represents more persistent changes to the methylome than those involving hypermethylation. The strongest site observed here is cg14391737 on chromosome 11, located within the serine protease 23 (*PRSS23*) gene along with multiple other DMPs. *PRSS23* is a member of the trypsin family of serine proteases and coordinates many physiological functions, including the immune

response and blood coagulation. Changes in the methylation state of this gene may contribute to the known protease-antiprotease imbalance observed in the emphysema of smokers. This causes an increased number of neutrophils and macrophages to be induced, causing these cells to release proteolytic enzymes that are not completely inhibited by antiproteases. These in turn lead to damage of connective tissues within the lung and has also been implicated in chronic obstructive pulmonary disease (COPD), a disease very closely associated to tobacco use. There is also strong evidence for this hypothesis in those with antitrypsin deficiency as it is the main inhibitor of neutrophil elastase (Abboud and Vimalanatha, 2008). However, there is a potential role of macrophage proteases in those without such a deficiency and serine proteases have been shown to contribute to elastolysis by alveolar macrophages in vitro, although it is difficult to identify a single protease that causes such lung destruction (Russell et al., 2002). Although this analysis was carried out on whole blood samples, the fact that former smokers had high estimates of monocyte number may in part explain the strong significance of *PRSS23* when comparing former and never smokers. It may also suggest a prolonged impact of smoking on protease activity, even after smoking cessation.

Another point to note in this comparison is the presence of one site, cg16047567, located in the *DHRS2* gene on chromosome 1. When comparing current and never smokers this site did not seem very important, ranking much lower on the list of DMPs compared to stronger signals but seems to one of the few sites that remain significantly differentially methylated after cessation. Dehydrogenase/Reductase 2 (*DHRS2*) codes for the Hep27 protein and is involved in carbonyl reductase (NADPH) activity. Smoking has been shown to stimulate the production of reactive oxygen species (ROS) which in turn causes oxidative stress and consequently many other related pathologies of the cardiovascular system. Some ROS-generating NADPH oxidases have been suggested to produce ROS and the reductive reaction of the Hep27 enzyme may work to prevent the toxic action of such species by converting them to less toxic compounds and thus help quench the effects of oxidative stress (Monge et al., 2009). The hypomethylation of such a site in former smokers compared to never smokers may then be related to the process.

The presence of DMPs in *AHRR*, *F2RL3, RARA* and intergenic loci in the 2q37.1 region when comparing former and never smokers suggests that these genes also become permanently differentially methylated in

those who have smoked and represent a long-term biomarker of past exposure to smoking which has already been shown for *AHRR* and *F2RL3* (Shenker et al., 2013).

3.4. Differentially Methylated Regions

DNA methylation change at individual CpGs often has a high dependence on its regional context and thus levels of methylation tend to correlate with neighbouring CpG sites. This co-methylation occurs in sites within close proximity to eachother and this is strongest for sites up to 1000bp from one another and especially in the context of CpG islands (Eckhardt et al., 2006). Regional clusters of neighboring CpGs that are differentially methylated with smoking are termed differentially methylated regions (DMRs) and may help better understand the biological processes affected by smoking and how the process of global DNA methylation change might occur. Using the *DMRcate* package in R, 836 DMRs were identified between current and never smokers through kernel smoothing and these again spanned the entire genome, as

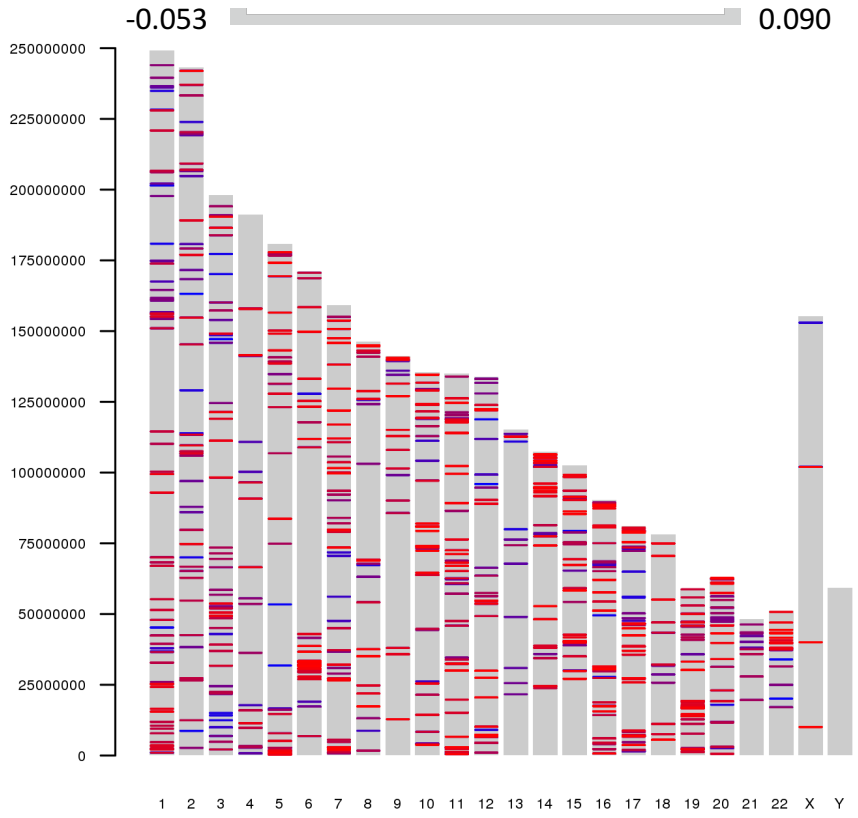

Figure 3.4.1. Genome-wide distribution of 836 differentially methylated regions between current and never smokers. DMRs are colour coded in relation to the mean beta fold change within the region with red representing negative associations and blue showing positive.

expressed in Figure 3.4.1. The 836 DMRs included 4806 individual CpG sites, 1457 of which were DMPs found to be differentially methylated between current and never smokers suggesting only a 28.03% coverage of the 5198 smoking-related DMPs in these DMRs. The chromosomal coordinates of the 836 DMRs were inputted into the Genomic Regions Enrichment of Annotations Tool (GREAT) online (McLean et al, 2010). This is particularly good for assigning biological meaning to non-coding genomic regions by analyzing annotations of nearby genes and is useful here given that almost a quarter of the identified DMPs were not annotated to a coding gene. GREAT also incorporates 20 different ontologies. Focusing on the enriched health-related ontologies, 3 diseases were found to be associated with the inputted smoking DMRs, including alpha 1-antitrypsin deficiency, crescentic glomerulonephritis and myelofibrosis.

Firstly, alpha 1-antitrypsin deficiency (A1AD) is a rare genetic condition in which decreased alpha-1 antitrypsin activity in the blood and lungs often leads to panacinar emphysema or COPD in adult life, especially in those exposed to cigarette smoke (Needham, 2004). Second, crescentic glomerulonephritis (RPGN) is a condition of the kidney characterized by rapid renal failure and with glomerular crescent formation involving layers of proliferating skin. In more than 50% of cases it is associated with another underlying disease such as Goodpasture syndrome (GPS). GPS is a rare autoimmune disease where the immune system attacks the basement membrane of the lung and kidneys and is likely caused by insults to the blood vessels connecting the lungs and heart, and one such possible insult is cigarette smoking (Greco et al., 2015). Lastly, myelofibrosis is a fairly rare cancer of the bone marrow where proliferation of an abnormal clone of hematopoietic stem cells leads to the replacement of marrow with scar tissue (Tefferi, 2014). Smoking can cause up to a 25% increase in peripheral blood leukocytes and has been shown to stimulate the bone marrow by shortening the transit time of polymorphonuclear leukocytes (Eeden and Hogg, 2000). Although not directly linked to smoking, these diseases all show some association between their pathology and cigarette smoke exposure and may help in understanding the downstream consequences on health caused by DNA methylation change in these particular DMRs. With that said, the lack of expected disease ontology is likely due to the small coverage of CpGs included within the microarray. The human genome has approximately 28 million CpG sites meaning the EPIC array only reflects a minute fraction of

5-mC DNA methylation. The accuracy of kernel smoothing in estimating real regions of differential methylation will thus be greatly hindered by this.

4. Dosage Effects

It is clear through the qualitative analysis of the previous chapter that smoking has a huge impact on DNA methylation, especially in current smokers. What is not clear are the factors that contribute to methylome variation in the same group of current smokers. Thus, in this chapter a more quantitative look at the impact of cigarette smoke on DNA methylation was used and considered the effects of dosage. This involved running linear models between transformed DNA methylation beta values, termed M-values, and data on number of cigarettes smoked per day and years spent smoking. This was done using 175 participants who currently smoked in wave 3 of *Understanding Society*, the same used in the qualitative linear models, as DNA methylation changes are known to be reversed in former smokers and thus may negatively impact the statistical strength of the model. These data were inputted as continuous, numeric variables in the design matrix with age, sex, white blood cell counts and batch as covariates. Methylation data was also restricted to the 5198 significant DMPs identified between current and never smokers to prevent any spurious associations.

4.1. Duration

The measure of duration used in these analyses was calculated by subtracting the age participants started to have first started smoking, a single time-point variable from wave 3, from their age at the time of blood collection. This gave the number of total years spent smoking which was then used in the linear model alongside the other stated variables. This initially only gave 1 significant association between duration and DNA methylation change, in the cg2030125 probes on chromosome 5, with an FDR adjusted *P*-value of just 0.02.

However, when looking at the interdependence between duration and other variables within the model a high amount of correlation was found, at 0.94, between years spent smoking and age. This makes sense given that most people start smoking in their teens, where the mean age of smoking initiation for current smokers was 16.78 ± 4.72. Thus, the older a smoker is the longer they have smoked and this hinders the

duration model greatly. To counteract this, age was then removed from the model as the two variables supply quite redundant information when together given their correlation and thus removal did not drastically impact the R-squared value. Once this was done 1331 significant DMPs were found to be associated with years spent smoking, although it is not completely clear whether these sites are differentially methylated with age instead or in fact both.

4.2. Epigenetic Age and Smoking Duration

To understand which of the 5198 smoking associated probes, seen when comparing current and never smokers, are in fact associated with age an age-only model was run using the 175 non-smokers used throughout these analyses. This yielded 2020 sites with an FDR adjusted *P*-value below 0.05, compared to 1422 age-associated sites seen when using current smokers. Furthermore, the top sites seen in the current smoker age model were much the same as those seen in the duration model but not the same as in the never smoker age model. For example, the most significantly associated probe with duration, cg19965693, located in the IFIH1 gene, was much further down the list, being the 486th strongest age-associated site in never smokers. This suggests that differential methylation at these sites are in fact more a consequence of smoking duration and not age and thus works to demonstrate the huge impact of smoking on the methylome whereby tobacco is a stronger driver of epigenetic modifications.

To further study the relationship of DNA methylation with age and smoking duration, the age of the 175 current smokers was predicted using DNA methylation data. Many different measures to do this have been created and used within the literature using a wide range of tissues but the most accurate thus far, especially in whole blood, seems to be Horvath's epigenetic clock (Horvath, 2013). The predicted ages and actual ages of the currently smoking participants had a correlation of 0.88. The age acceleration of these same participants, or difference in actual age and DNA methylation age, was then calculated and ranged widely from -15.46 to 23.53 at the two extremes, although most estimations were much closer to the chronological age of the participants. Here, quite a strong negative trend of -0.69 was observed between age acceleration and actual age where younger participants tended to have underestimated predicted ages and older

participants had overestimated predicted ages. Interestingly, this occurrence was more pronounced in non-smokers who had a correlation of -0.76 between age acceleration and actual age and more still in former smokers who had a correlation of -0.79. It is unclear at present why this occurs and may suggest that the Horvath clock does not fit well on EPIC methylation array data or that the missing probes from the 450K array on which the clock was built might have some impact. Another reasoning might be an issue within the *Understanding Society* data itself. In the latter, perhaps a "survival bias" is present where those that have survived to old age are more likely to be healthier and generally biologically younger than others who had not reached old age. With this reasoning perhaps smokers do not tend to reach older age and thus do not display survival bias to the same degree as non-smokers, explaining the lower correlation of smoker's age with age acceleration.

When comparing age acceleration with years spent smoking a similar negative trend was observed between the two. This would initially suggest that the DNA methylation age acceleration of early smokers is reversed with duration but this clearly is not the case but a consequence of the strong correlation between duration and age. Taken together it is clear that age and duration are highly intertwined but given the strong associations of smoking with DNA methylation it is acceptable to assume that by dropping age from the model, duration-associated sites can be observed and insights into the impact of dosage on the methylome can be made.

4.3. Duration-related DMPs

The FDR adjusted *P*-values and estimated log2-fold changes of the 1331 duration-associated DMPs are summarized in Figure 4.3.1. and the top 10 hypomethylated and 10 hypermethylated sites are shown in Table 1. The top hit, with an adjusted *P*-value of 4.91×10^{-15} was cg19965693, located upstream of the TSS in the *IFIH1* gene on chromosome 2. The *IFIH1* gene encodes the Interferon Induced with Helicase C Domain 1 protein and is also known as Melanoma Differentiation-Associated protein 5. It functions as a pattern recognition receptor (PRR) that senses viruses and is consequently an essential part of the immune system (Takeuchi and Akira, 2008). Antibodies against MDA5 are associated with amyopathic

dermatomyositis with rapidly progressive interstitial lung disease (Fiorentino et al., 2011), providing a link between the hypomethylation of this CpG site and smoking duration. Furthermore, inhibiting DNA methylation has been shown to induce the response of such interferons in cancer through dsRNA sensors (Chiappinelli et al., 2015). Given that both age and smoking are risk factors for a number of cancers, differential methylation observed at this gene may then relate to interferon function and thus strengthens the importance of DNA methylation, among other mechanisms, in controlling health. Furthermore, as this differential methylation is highly associated with smoking duration, perhaps this correlates with the extent of health impacts caused by cigarette smoke.

Table 4.3.1.: Top 10 hypomethylated and hypermethylated probes associated with smoking duration

Illumina Probe ID	Chr Number	Chromosome position (bp)	Design Type	UCSC Gene Name	UCSC Gene Region	Present in 450K Array	FDR Adjusted P Value	Estimated Log2 Fold Change
Hypomethylated in current smokers								
cg19965693	2	163175743	II	IFIH1	TSS1500	FALSE	4.91E-15	-0.02
cg16267679	2	145278615	II	LINC01412;ZEB2	TSS1500;TSS1500	FALSE	2.75E-11	-0.02
cg00602811	2	145278564	II	ZEB2	TSS1500	TRUE	2.75E-11	-0.02
cg18826637	2	145116633	II			TRUE	1.13E-10	-0.03
cg21323642	22	31709724	II			FALSE	2.10E-09	-0.01
cg12919873	21	38929815	II			FALSE	2.92E-09	-0.02
cg03834786	7	80571772	II			FALSE	3.31E-09	-0.01
cg00573770	2	145278485	I	ZEB2	TSS1500	TRUE	9.91E-09	-0.02
cg19344626	19	16830749	I	NWD1	TSS200	TRUE	1.68E-08	-0.02
cg11649376	12	81473234	II	ACSS3	Body	TRUE	1.68E-08	-0.01
Hypermethylated in current smokers								
cg04738965	3	147127662	I	ZIC1	1stExon;5'UTR	TRUE	2.02E-09	0.02
cg15466862	13	112722333	I	SOX1	1stExon	TRUE	2.55E-08	0.02
cg26422458	1	79472452	I	ELTD1	5'UTR;1stExon	TRUE	3.53E-08	0.02
cg10906284	12	63544430	I	AVPR1A	1stExon	TRUE	3.53E-08	0.02
cg23621097	17	1962236	I	HIC1	3'UTR	TRUE	7.00E-08	0.02
cg19942495	17	32484027	II	ACCN1	TSS1500	TRUE	9.86E-08	0.01
cg02926165	5	3595963	II	IRX1	TSS1500	FALSE	9.96E-08	0.02
cg17953764	4	48492845	I	ZAR1	1stExon	TRUE	1.04E-07	0.01
cg24125828	6	32117049	I	PRRT1	Body	TRUE	2.02E-07	0.02
cg16001722	6	127836159	II	C6orf174	Body	TRUE	2.09E-07	0.02

When comparing age-associated and duration-associated DMPs, 353 loci that were related to time spent smoking were not significantly differentially methylated with age at an FDR threshold of 0.05. The most

commonly occurring genes within these probes included four in the well characterized smoking-associated gene of *AHRR* and another four in the less reported *SLC24A3* gene on chromosome 20. *SLC24A3* is a potassium-dependent sodium/calcium exchanger on the plasma membrane. Excess salt intake has been implicated in the pathology of hypertension and one study showed epistatic interactions of SNPs in *SLC24A3* with pressure-natriuresis (Citterio et al., 2011). Given that smoking increases blood pressure, the hypomethylation of these sites with increasing duration may provide a link between this disease and calcium homeostasis.

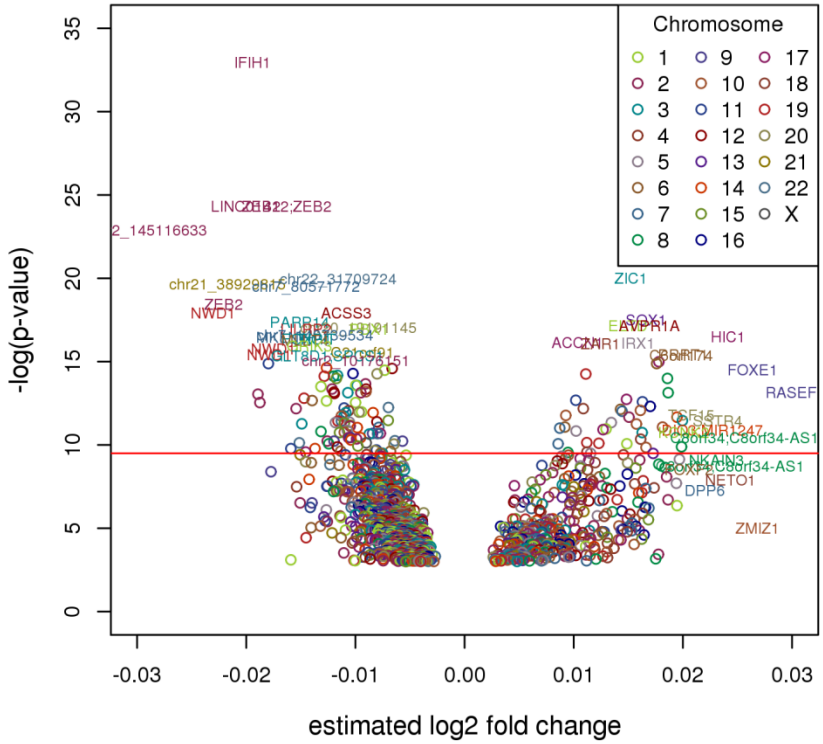

Figure 4.3.1.: Summary of 1331 significant CpG sites differentially methylated with smoking duration at false discovery rate P < 0.05. Red line indicates Bonferroni threshold. Each CpG site is represented by significance as shown by their –log(P-value) values (y-axis) and effect size and direction (x-axis), the mean β value difference between groups. Associations are colour-coded in reference to the chromosome the CpG site is located on.

The 353 CpG "duration-only" sites with available annotation were located in roughly 236 genes. To better understand the mechanisms these genes are involved in, they were inputted into the STRING database to visualize any interactions between their protein products. At the highest confidence, with a minimum required interaction score of 0.9, 53 interactions were found as shown in Figure 4.3.2.

These interactions are involved in a number of KEGG pathways. Firstly, NRP1, EPHA7, SEMA7A, NCK2, RAC1 and PAK4 are all important in axon guidance. Some of these, especially members of the semaphorin family like SEMA7A, make up molecules that guide the outgrowth of axons, called axonal guidance cues. These in turn promote normal alveolar growth and many lung disease have been characterised by damage to the alveolar, linking this gene function with disease phenotypes caused by smoking (Vadivel et al., 2013). One such disease is bronchopulmonary dysplasia (BPD). This suggests that changes in DNA methylation might act as a biomarker for BPD and other smoking-related diseases. Furthermore, given the association of these sites with dosage, perhaps this can provide a quantifiable risk predictor to such diseases.

Another interesting KEGG pathway is glutathione metabolism which ANPEP, MGST1, LAP3 and GGT1 are involved in. Smoking causes damage to respiratory tract tissues and it is known that glutathione and related thiols can help perturb this and is an antioxidant component of the tract lining. This occurs when

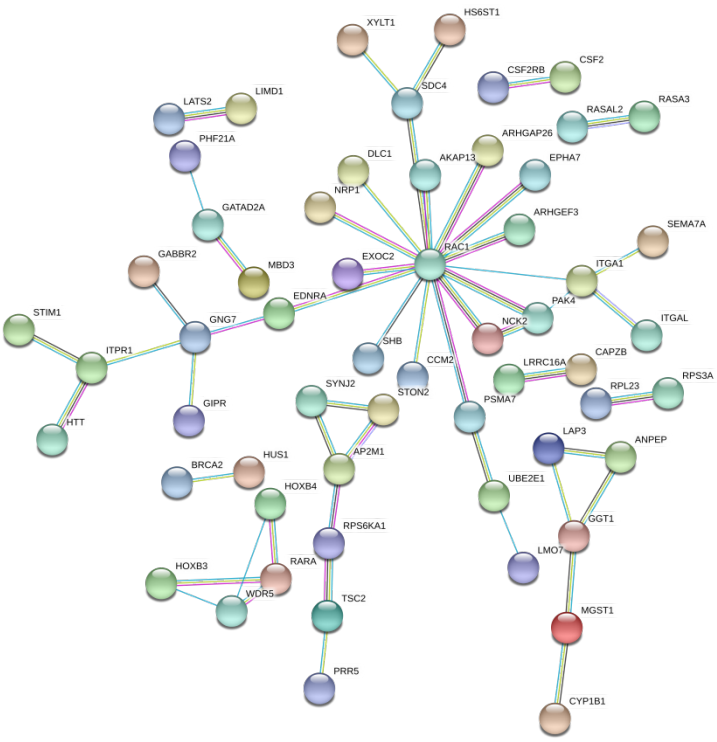

Figure 4.3.2.: STRING interaction network of the 53 interacting proteins encoded by the genes enclosing duration-associated DMPs seen at a FDR significance threshold and not associated with age. Only shows interactions that meet the highest confidence threshold, with an interaction score of 0.9. Lines are coloured by the type of evidence the interaction is based on. Red - presence of fusion, Green - neighborhood evidence, Blue - cooccurrence, Purple -experimental, Yellow - textmining evidence, Light blue - database, Black – coexpression.

cigarette smoke reacts with glutathione-aldehyde derivatives and depletes the amount of available glutathione (Toorn et al., 2007). The duration-related changes to DNA methylation seen in genes related glutathione metabolism might then be related to the gradual worsening of health seen with increasing years of tobacco use.

4.4. Intensity

Another important dosage measure is that of intensity, or how much a person smokes as opposed to how long. To see if this influences DNA methylation, two measurements of intensity were used from the same currently smoking participants used above. The first is a single time-point variable from wave 3, the same year of the nurse visit where blood samples were collected. Here participants were asked how many cigarettes smoked per day. The second measurement involved taking an average of this same variable between waves J through to R of BHPS. This gave the mean number of cigarettes smoked by participants in the nine years leading up to wave 3 in *Understanding Society.*

When the single time-point measure was used in the linear model, 33 CpG sites were significantly differentially methylated with number of cigarettes smoked. The strongest association was seen in the cg10590512 probe located in the *DIO3;MIR1247* gene region on chromosome 14. Alternatively, the average measure of number of cigarettes yielded 23 DMPs with the most significant probe, cg25992330, located in the *SPOCK2* gene on chromosome 10. The top ten loci for both measures are summarized in Table 2. It appears that the single time-point measure reflects a wider range of DNA methylation changes however the mean variable gave more significant *P*-values and thus suggests that by taking the mean a more sensitive and less noisy indication of cigarette smoke intensity can be measured. Therefore, this was used for the remaining analyses.

The *SEPT9* CpG site, cg11328665, was observed in both models suggesting that methylation at this site may be highly sensitive to intensity. 11 probes within this gene were also differentially methylated between current and never smokers, strengthening its association with smoking. *SEPT9* codes for the Septin-9

protein. The v2 region of the gene's promotor has been shown to become hypermethylated compared to healthy individuals in both the tissue of those with colorectal cancer as well as in the blood. This altered methylation pattern may indicate aberrant activation or repression of this gene and may have downstream consequences on pseudopod protrusion, tumor cell migration and invasion, processes known to rely on the function of *SEPT9* (Tetzner et al., 2009). Smokers are known to have a significantly higher risk of developing colorectal cancer than those who do not smoke. Additionally, a dose-relationship between colorectal cancer risk with increasing number of cigarettes has also been reported, but only after 30 years of smoking (Botteri et al., 2008). Thus, this may provide a link between changes in DNA methylation at *SEPT9* with intensity of smoking and in turn cancer.

Table 4.4.1.: Top 10 hypomethylated and hypermethylated DMPs associated with smoking intensity

Illumina Probe ID	Chr Number	Chromosome position (bp)	Design Type	UCSC Gene Name	UCSC Gene Region	Present in 450K Array	FDR Adjusted P Value	Estimated Log2 Fold Change
Single time-point								
cg10590512	14	102026939	I	DIO3;MIR1247	TSS1500;TSS200	FALSE	3.15E-03	0.02
cg03440944	7	45023329	II	C7orf40	Body	TRUE	1.90E-02	-0.01
cg11328665	17	75446304	II	SEPT9	TSS1500;Body	FALSE	1.90E-02	-0.01
cg17567838	6	33167488	II	SLC39A7	TSS1500;Body	TRUE	2.36E-02	-0.01
cg09945032	3	38871019	II			FALSE	2.80E-02	-0.01
cg06532880	5	176731545	II	PRELID1;RAB24	Body;TSS1500	TRUE	2.80E-02	-0.01
cg05575921	5	373378	I	AHRR	Body	TRUE	2.80E-02	-0.02
cg03220447	11	19745293	II	NAV2	Body	TRUE	2.80E-02	-0.01
cg11320225	1	161709963	II			FALSE	2.80E-02	-0.01
cg17731696	8	66734530	II	PDE7A	Body	FALSE	3.83E-02	-0.01
Mean of multiple time points								
cg25992330	10	73830099	II	SPOCK2	Body	FALSE	1.23E-03	0.01
cg05575921	5	373378	I	AHRR	Body	TRUE	1.29E-02	-0.03
cg26341457	11	72523885	II			FALSE	1.29E-02	-0.01
cg10590512	14	102026939	I	DIO3;MIR1247	TSS1500;TSS200	FALSE	1.29E-02	0.02
cg03707168	19	49379127	II	PPP1R15A	Body	TRUE	1.37E-02	-0.01
cg11328665	17	75446304	II	SEPT9	TSS1500;Body	FALSE	1.37E-02	-0.01
cg03440944	7	45023329	II	C7orf40	Body	TRUE	1.37E-02	-0.01
cg09945032	3	38871019	II			FALSE	1.37E-02	-0.01
cg07954423	9	130741881	II	FAM102A	Body	TRUE	1.37E-02	-0.01
cg00475490	11	86517110	II	PRSS23	5'UTR;Body	FALSE	1.37E-02	-0.02

However, by far the site with the strongest association with mean number of cigarettes smoked, with an FDR adjusted *P*-value of 0.001, is cg25992330 located in the *SPOCK2* gene. This encodes a member of a calcium binding proteoglycan family, important components of the extracellular matrix. It is also a known susceptibility gene for bronchopulmonary dysplasia, formerly known as chronic lung disease of infancy (Hadchouel et al., 2011) and this disease was also elucidated to in the enriched axon guidance KEGG pathway with duration DMPs.

Clearly increasing number of cigarettes causes changes in DNA methylation at important disease-related genes associated with smoking. However, all other sites were very close to the 0.05 cutoff and thus raises concerns about the reality of their associations with smoking intensity. In fact, this suggests that intensity all together may not have much influence on DNA methylation and instead the drastically larger number of associations seen with smoking duration suggests that years spent smoking may have a far larger impact on the methylome. Nevertheless, this shows that different CpG sites, even those located in the same region, respond differently to duration and intensity. This may provide an insight into the different biological mechanisms at play when studying how environmental factors and dosage influence DNA methylation.

4.5. Duration vs Intensity

As there is such a large difference between duration and intensity in both the number and location of dosage-related DMPs, the two were compared further. The aim of doing this was to better classify these differences and hopefully identify which dosage measure might be more important in DNA methylation. Firstly, Figure 4.5.1. compares the distribution of significance values and log2-fold change estimates between duration and intensity for all 5198 smoking-associated CpG sites. These plots show far more significant changes in DNA methylation with duration compared to intensity as expressed before in the two models. However, the estimates of the log2-fold-change for each probe, which in a way gives a measure of effect size with increasing dosage, do not show such large differences. In fact, both dosage measures saw estimates that were very low in comparison to the more than 1.0 log2-fold changes seen when comparing current and never smokers but this may be a consequence of using continuous variables instead of categorical factors.

These log2-fold change estimates do show that the duration-associated DMPs had slightly larger estimates than intensity. It also shows that positive coefficients also tended to have larger estimates of log2-fold-change and this seemed to be independent of their significance. Furthermore, both measures of dosage show a greater representation of negative over positive coefficients, but to a smaller extent when using intensity data. Given that hypermethylated sites in smokers tended to have modest significance values compared to the hypomethylated probes, this could suggest that the mechanism driving this change causes larger differences in DNA methylation and may, in a small way, explain why number of cigarettes smoked per day makes less of an impression on DNA methylation. Either way, by looking at these funnel plots it is clear that the number of significant DMPs does not fully reflect the actual, measurable changes to the methylome caused by increased duration or intensity. Instead it simply states how many probes met a certain threshold and this number is also highly sensitive to the covariates included in the model.

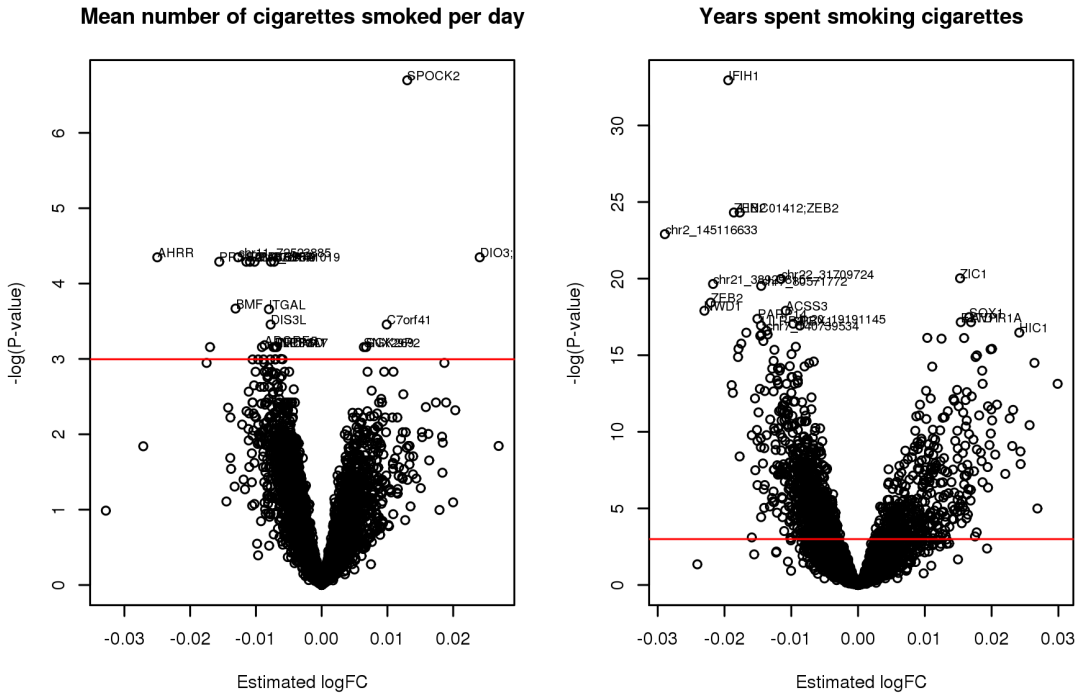

Figure 4.5.1.: Funnel plots summarizing the associations of DNA methylation with smoking intensity and duration for the 5198 smoking-associated probes. Each CpG site is represented by significance as shown by their –log(P-value) values (y-axis) and the direction and change in DNA methylation, represented by their estimated log2-fold changes (x-axis), a measure of effect for the dosage coefficient.

To overcome this and further test the effects of duration and intensity on DNA methylation, the current smokers were split by quantiles into four equal-sized groups of 40 based on number of cigarettes smoked

per day, and another four based on years spent smoking. Then, the difference in DNA methylation between each smoker group and an equal-sized sample of 40 never smokers, termed here as the absolute effect size, was calculated. By doing so it is possible to measure the extent of DNA methylation differences between non-smokers and smokers with varying dosage exposures. These differences were calculated using Beta-values for the top 10 probes associated with years spent smoking in the duration groups and the top 10 probes associated with mean number of cigarettes for the intensity groups. The total mean DNA methylation difference between the eight smoker groups and never smokers are outlined in Figure 4.5.2.

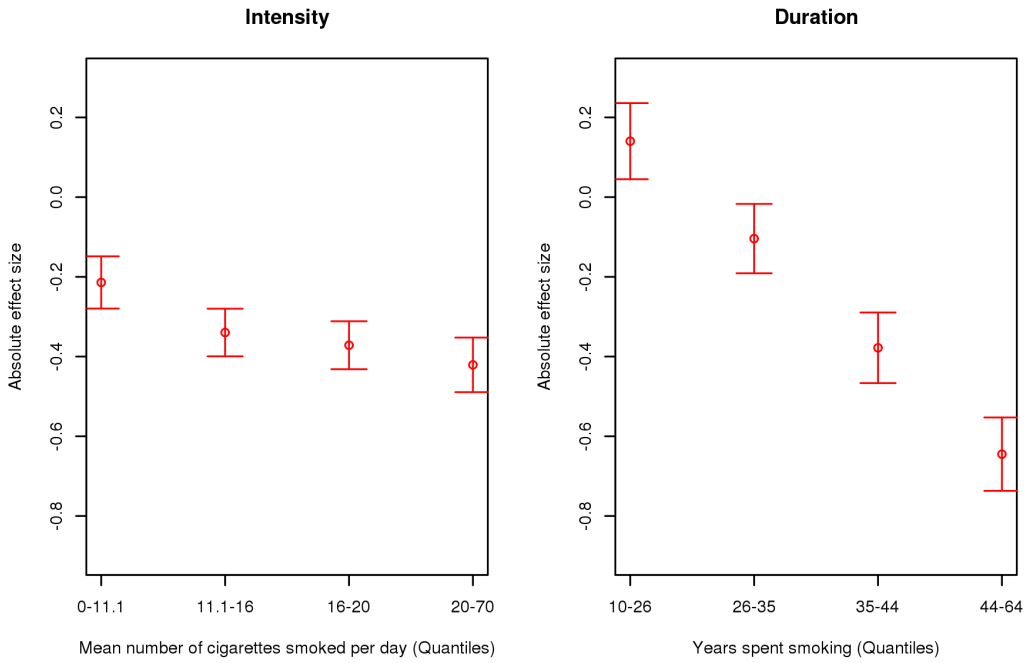

Figure 4.5.2.: Mean absolute effect size between smokers and non-smokers per dosage quantile. Shows mean difference between the 40 smokers and 40 non-smokers in each group, calculated using DNA methylation values for the top 10 DMPs associated with duration (right) and intensity (left).

This showed a relationship between increases in both duration and intensity and increased DNA hypomethylation in smokers compared to non-smokers, at least for the top 10 sites associated with each measure. However, duration was associated with a much greater change in DNA methylation compared to intensity. Here, absolute effect sizes across the duration quantiles ranged from 0.14 to -0.65 showing a huge decrease in DNA methylation, of almost 0.79, between those who have smoked less than 26 years and those having smoked more than 44 years. On the other hand, the absolute effect sizes between the intensity groups and never smokers ranged from -0.42 to -0.21. This shows a much smaller decrease in DNA methylation,

around 0.21, between those who smoke less than 11.1 cigarettes a day and those who smoke more than 20. Furthermore, the steepest methylation difference for intensity was observed between the first and second quantile groups. Thereafter, smaller decreases in DNA methylation are observed which suggests that the methylome of those who smoke less than 11.1 cigarettes a day is less effected than those who smoke more but not by much. As for duration, changes in DNA methylation between smokers and non-smokers were more gradual, with similar decreases in absolute effect sizes with each increasing dosage group. This suggests DNA methylation is more highly correlated with years spent smoking than number of cigarettes smoked. Taken together, this analysis suggests that duration rather than intensity is more important in driving changes to the methylome and thus may be more influential in smoking-related illness in the cases where DNA methylation influences the expression of important health-associated genes.

To look at the interplay between duration and intensity the analyses above were repeated, splitting the four duration quantiles into eight groups based on intensity, where four consisted of participants smoking less than 16 cigarettes per day and another four consisted of those who smoke more than 16. Equal sized samples of 15 participants were then used in each analysis and the same was carried out for intensity where the groups were split by who reported to have smoked for less or more than 35 years. The absolute effect size between these 16 groups and another subset of 15 non-smokers are outlined in Figure 4.5.3.

In all four scenarios increasing dosage led to DNA hypomethylation in smokers compared to non-smokers. However, duration again showed the biggest changes in DNA methylation with a steady decline in absolute effect size across years spent smoking. As for intensity, a steep decline in DNA methylation was again seen between the first and second quantiles but thereafter DNA methylation difference varied greatly and at some points even seemed to increase. This is likely caused by the small sample size used to calculate the average effect size between the intensity groups and never smokers. Here, increasing number of cigarettes seemed to have little impact on DNA methylation change and thus was not able to overcome this shortfall.

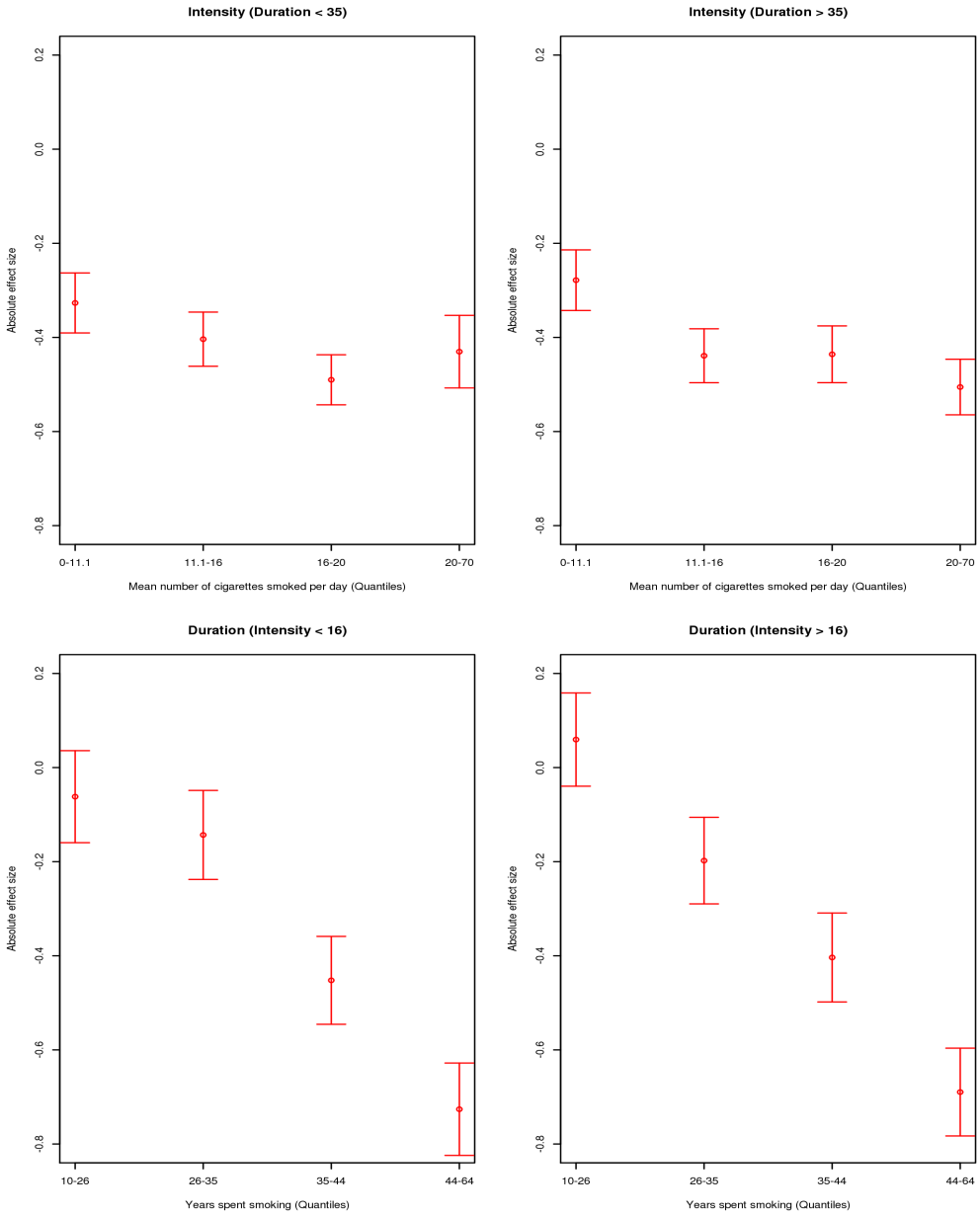

Figure 4.5.3.: Mean absolute effect size across between smokers and non-smokers per dosage quantile split by median duration (top) and intensity (bottom). Shows mean difference between the 15 smokers and 15 non-smokers in each group, calculated using DNA methylation values for the top 10 DMPs associated with duration (bottom) and intensity (top).

Increasing duration on the other hand still showed a strong negative trend with DNA methylation change despite the small group sizes.

The two dosage measures seemed to make little impact on DNA methylation differences. The one exception to this being in those who had smoked for up to 26 years within the first quantile. Beta values for participants who fell into this group and also smoked more than 16 cigarettes a day were higher by around 0.11 than other participants who smoked less but for similar durations. This suggests than intensity only impacts DNA methylation in the early years of smoking and thus for the majority of cases it appears than intensity and duration are independent of one another.

4.6. Predicting Smoking Duration

Throughout this chapter it is clear that years spent smoking is highly correlated with DNA methylation levels at a number of CpG sites. This might then allow for smoking duration to be estimated using measures of DNA methylation at duration-associated sites. To attempt this, a similar procedure was used to that developed by Horvath for the well-known epigenetic age clock (Horvath, 2013). The general principle to this is to form a weighted average of DNA methylation at a number of duration-related CpG dinucleotides and then transform this into years spent smoking. This is done using a penalized regression model which uses an elastic net regularization approach to regress self-reported duration years from a sample of 85 current smokers, termed as the training set, onto the 1331 EPIC array probes associated with duration. Elastic net regularization makes use of the "elastic net" that combines the L1 and L2 penalties of the lasso and ridge method and hence overcomes the limitations of many modelling techniques. The elastic net then automatically selects the CpG sites most associated with duration and these are then used in the prediction.

The regression model was run using the *glmnet* package from R (Friedman et al., 2010). To do so, first a 10 fold cross validation was carried out, using the **cv.glmnet** function, to find an estimate of the lambda parameter to be used. Then, a generalised linear model, with elastic net regularization, was fitted to the training set data using the **glmnet** function. This function uses a penalized maximum likelihood approach

and requires specification of the alpha parameter, which was set to 0.5 as the elastic net predictor was used, and the lambda value, which was set to 1.120921 as estimated by the cross-validation. This model then enabled the calculation of estimates for years spent smoking by using the usual **predict** function. This last step was carried out on methylation Beta-values from the other half of current smokers, the test set. The predicted and actual years since quitting are shown in Figure 4.6.1.

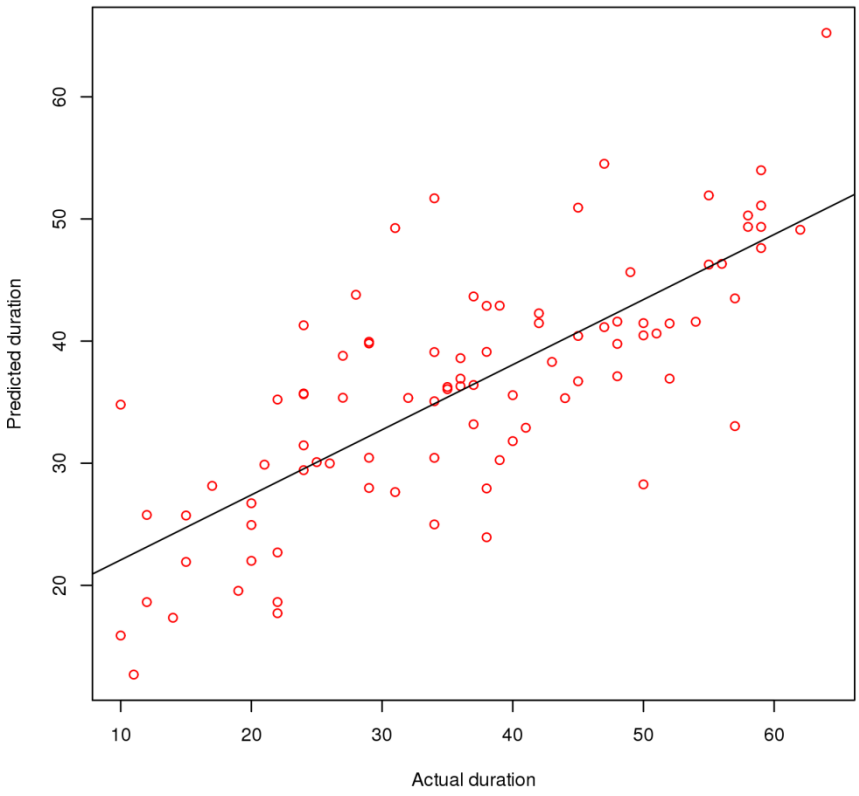

Figure 4.6.1.: Goodness of fit for predictor of years spent smoking. Shows actual years spent smoking (x-axis) against predicted years spent smoking (y-xis). Pearson correlation coefficient = 0.76.

This shows a correlation of 0.76 between the predicted and actual years spent smoking in former smokers. The mean difference between the actual and predicated duration values in this group was 0.31 with a standard deviation of 9.30. This variation is likely caused by the small sample size used when creating the model and thus DNA methylation data from a larger number of active smoker participants would likely aid in improving this prediction. Nevertheless, this predictor also showed some correlation with duration in former smokers, with a Pearson of coefficient of 0.43. This smaller correlation is likely a reflection of the reversal of DNA methylation changes that occurs when a person stops smoking. Together this suggests that this prediction has value as a biomarker of smoking duration and can be used to estimate the extent of

changes to the methylome caused by duration. Furthermore, for disease-associated genes whose expression is influenced by DNA methylation, this model may even help predict risk to health caused by tobacco use.

5. Cessation

The best tactic for preventing smoking-related illness is clearly to not smoke in the first place but given that there are still appropriately 1 billion smokers worldwide, the second best option is to quit smoking as soon as possible. By doing so the impact on the methylome associated with years of tobacco use may be minimized. In the third chapter, it was apparent that DMPs still exist when comparing former and never smokers and thus cessation may not completely recover DNA methylation to that of someone who has never smoked. It does however reduce the degree of differences between ex-smokers and non-smokers drastically, including changes to epigenetic loci located in important disease-associated genes. DNA methylation then is very sensitive to smoking status and thus may have potential as a biomarker for years of cessation. To further study this a number of analyses were carried out using the 356 participants of *Understanding Society* who stated in wave 3 that they had previously smoked but do not do so anymore. These were also selected on the availability of their cessation data which included self-reported variables on age of quitting. The number of years since cessation was thus calculated by taking the age of quitting and subtracting this from their age at the time of blood sample collection in wave 3 of the study.

5.1. Cessation-related DMPs

Years since quitting ranged from 1 to 66 years and this data had some correlation with both age (0.57) and smoking duration (-0.45). Despite this, age was kept in the model alongside sex, blood process day, batch and white blood distribution estimates to avoid any fictitious associations with cessation. First, a linear model was used to look at the relationship between DNA methylation levels and years since quitting smoking. This was restricted to the 5198 smoking-associated loci identified when comparing current smokers and participants who had never smoked. The model yielded 192 significantly differentially methylated sites with years since quitting, and the significance and log2-fold-change values of these sites are summarized in Figure 5.1.1. 171 of the sites became hypermethylated with years since quitting leaving 21 hypomethylated sites. This makes sense as a reversal of the huge surplus of hypomethylated probes seen in smokers compared to non-smokers. It also suggests that these sites gradually recover their

methylation states with increasing cessation years. Of the 192 DMPs, 82 were also associated with duration. This suggests that the process by which the reversal of DNA methylation changes occur may be related but is not simply the reverse process driving such changes with increasing duration and instead there are more factors at play here.

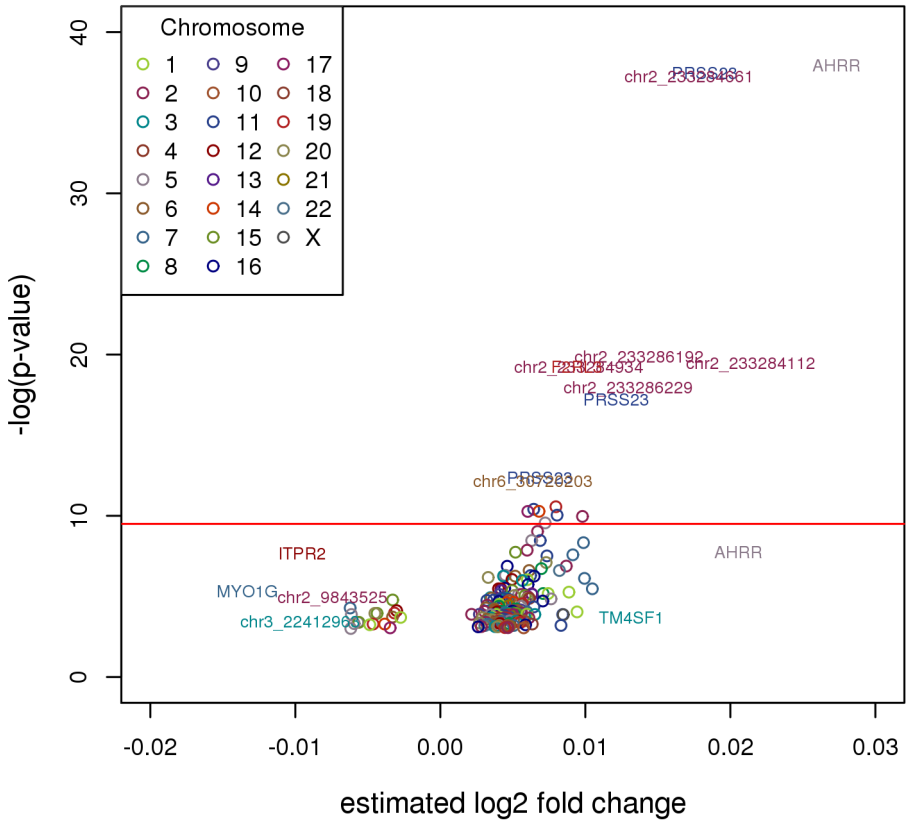

Figure 5.1.1.: Summary of 192 significant CpG sites differentially methylated with smoking cessation at false discovery rate P < 0.05. Red line indicates Bonferroni threshold. Each CpG site is represented by significance as shown by their –log(P-value) values (y-axis) and effect size and direction (x-axis), the mean β value difference between groups. Associations are colour-coded in reference to the chromosome the CpG site is located on.

The top 10 hypermethylated and 10 hypomethylated cessation-associated DMPs are shown in Table 1. The most strongly associated probe, cg05575921 within the *AHRR* gene, is also the most strongly associated site with smoking in general. Sites within this gene had weak associations with duration but had the second strongest association with intensity. Its presence as a cessation-related DMP may then be caused by the now zero number of cigarettes smoked by former smokers. This site also becomes hypermethylated very quickly after cessation for most participants, as show by a steep curve between Beta-values for this site and

years since quitting, strengthening the idea that DNA methylation at this site is influenced by number of cigarettes.

Table 5.1.1. Top 10 hypomethylated and hypermethylated DMPs associated with smoking cessation

Illumina Probe ID	Chr Number	Chromosome position (bp)	Design Type	UCSC Gene Name	UCSC Gene Region	Present in 450K Array	FDR Adjusted P Value	Estimated Log2 Fold Change
Hypermethylated in current smokers								
cg05575921	5	373378	I	AHRR	Body	TRUE	3.35E-17	0.027
cg14391737	11	86513429	II	PRSS23	5'UTR;Body	FALSE	5.27E-17	0.018
cg21566642	2	233284661	I			TRUE	6.60E-17	0.017
cg16841366	2	233286192	II			FALSE	2.33E-09	0.014
cg06644428	2	233284112	I			TRUE	3.52E-09	0.021
cg01940273	2	233284934	II			TRUE	4.42E-09	0.010
cg03636183	19	17000585	II	F2RL3	Body	TRUE	4.42E-09	0.009
cg22812571	2	233286229	II			FALSE	1.58E-08	0.013
cg00475490	11	86517110	II	PRSS23	5'UTR;Body	FALSE	3.32E-08	0.012
cg11660018	11	86510915	II	PRSS23	TSS1500	TRUE	4.35E-06	0.007
Hypomethylated in current smokers								
cg09375092	12	26576047	II	ITPR2	Body	FALSE	4.63E-04	-0.010
cg12803068	7	45002919	II	MYO1G	Body	TRUE	4.86E-03	-0.013
cg08035323	2	9843525	II			TRUE	6.96E-03	-0.007
cg10819708	15	90735422	II	SEMA4B	5'UTR	TRUE	8.41E-03	-0.003
cg05009104	7	45002980	II	MYO1G	Body	FALSE	1.39E-02	-0.006
cg02327909	12	53009774	II	KRT73	Body	FALSE	1.63E-02	-0.003
cg07790294	19	928287	II	ARID3A	5'UTR	FALSE	1.89E-02	-0.003
cg19459791	15	65363022	II			TRUE	1.93E-02	-0.004
cg01756827	20	47923790	II			FALSE	1.93E-02	-0.005
cg10874644	5	83898708	II			TRUE	2.07E-02	-0.006

Interestingly, sites located in the *PRSS23* gene were also strongly associated with years of quitting and were also present in the DMPs between former and never smokers. The DNA methylation of sites in this gene may then be reversed very slowly after cessation. Furthermore, *PRSS23* encodes the serine protease 23 enzyme, and is responsible for a huge range of biological processes, including blood clotting, food digestion, infection fighting and fertilization (Neitzel, 2010). This then hints that the changes in DNA methylation observed at this gene may be linked to the increased good health seen in ex-smokers, especially after many years of cessation. Other strongly associated sites lied in the 2q37.1 intergenic region. The closest genetic feature, found using the **nearest.gene** function from the *humarray* CRAN package, was

ALPPL2, an alkaline phosphatase. Levels of this enzymes have been show to increase up to tenfold in cigarette smokers and cancer patients (Koshida et al., 1990) and thus with increased length of cessation these levels may fall and this again shows a possible link between the reversion of DNA methylation changes and the improved health, and reduced cancer risk, seen in former smokers compared to current smokers.

Interestingly, other sites in genes strongly associated with smoking, such as *ZMIZ1* and *ELMSAN1*, were not associated with cessation. In fact, DNA methylation of only a handful of smoking-related CpG sites seemed to be influenced by smoking cessation, leaving well over 5000 sites that were not related with years since quitting. One line of reasoning for this is that for these sites DNA methylation is permanently altered by tobacco use. This then suggests that the site-specific manner by which DNA methylation changes happen may actually be split into two distinct groups of reversible and permanent smoking-related DMPs. However, this would then suggest a huge surplus of permanently altered CpG sites but only 15 probes were found to be significantly differentially methylated between former and never smokers so this seems unlikely. A more plausible explanation relates to the fact that this this model looks into changes in DNA methylation across all years of smoking cessation, starting at 1 year. Perhaps some sites are reversed very early after quitting, within the first year where no such data is available. Thus, these sites would quite rightly not be associated with cessation in this analysis and instead the 192 identified DMPs show sites that have a longer time-span for reversal, taking some decades to be restored to anywhere near non-smoker levels.

The 192 significantly associated probes with available annotation were located in 104 genes. To look at the relationship between these gene,s and more specifically their protein products, interactions were searched using the STRING database. This showed 21 of the 104 proteins interacted with at least one other inputted protein using the highest confidence threshold, where all interactions had a score of at least 0.9. This network is shown in Figure 5.1.2. and consists of 5 distinct clusters of proteins. Five of these proteins, including AKT3, HTRA2, ITPR2, TNF and BIRC3, were implicated in the apoptosis pathway from KEGG. In fact, one apoptotic pathway is directly mediated by the death receptor of the tumor necrosis factor (TNF)

receptor family. Lack of functional apoptotic pathways can lead to harmful cell proliferation and survival, a distinctive feature of cancer. Previous studies have seen aberrant methylation at important, apoptosis-related CpG sites in cancer cell lines including that of the lung. These were located in the promotor regions of O-6-methylguanine-DNA-methyltransferase (*MGMT*), a DNA repair enzyme, and RAS-associated domain family protein 1A (*RASSF1A*), a tumor suppressor gene. Methylation of promotors is suggested to be a mechanism of gene silencing and tumorigenesis and in this case caused decreased DNA repair capacity and decreased cell cycle control in cells, leading to reduced apoptosis (Koutsimpelas et al., 2012). This shows that a strong link between differential DNA methylation and apoptosis exists. Furthermore, the fact that these genes are associated with cessation, and become hypermethylated with years since quitting, may mean effective apoptotic processes can be regained by stopping smoking and this in turn would lead to better health outcomes. The implication of pathways in cancer was also seen for these genes from KEGG, strengthening their associations with health.

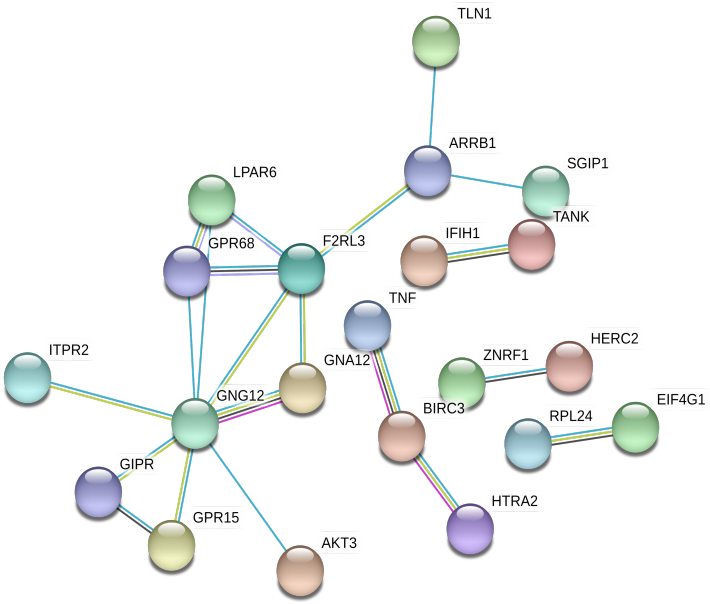

Figure 5.1.2.: STRING interaction network of the 21 interacting proteins encoded by the genes enclosing cessation-associated DMPs seen at FDR genome-wide significance threshold. Only shows interactions that meet the highest confidence threshold, with an interaction score of 0.9. Lines are coloured by the type of evidence the interaction is based on. Red - presence of fusion, Green - neighborhood evidence, Blue - cooccurrence, Purple -experimental, Yellow - textmining evidence, Light blue - database, Black – coexpression.

5.2. DNA Methylation with Years Since Quitting

As stated, the linear model that compared DNA methylation and years of quitting can only reveal sites whose DNA methylation levels change with increasing time since quitting within former smokers. It does not show which of these loci are still significantly differentially methylated compared to never smokers nor at which year of cessation these loci might lose this significance. To counteract this the 356 participants used in these analyses were split by quantiles into five equal-sized groups, each containing 60 participants.

After, linear models were repeated as before but this time used to compare each of these groups of former smokers, each with similar years since quitting smoking, to an equal sized reference sample of never smokers. This analysis was limited to the 192 cessation-associated probes identified before to ensure assumptions are only made about DNA methylation decay related to time since quitting. Figure 5.2.1. shows the number of DMPs observed between each cessation quantile and the non-smoker sample.

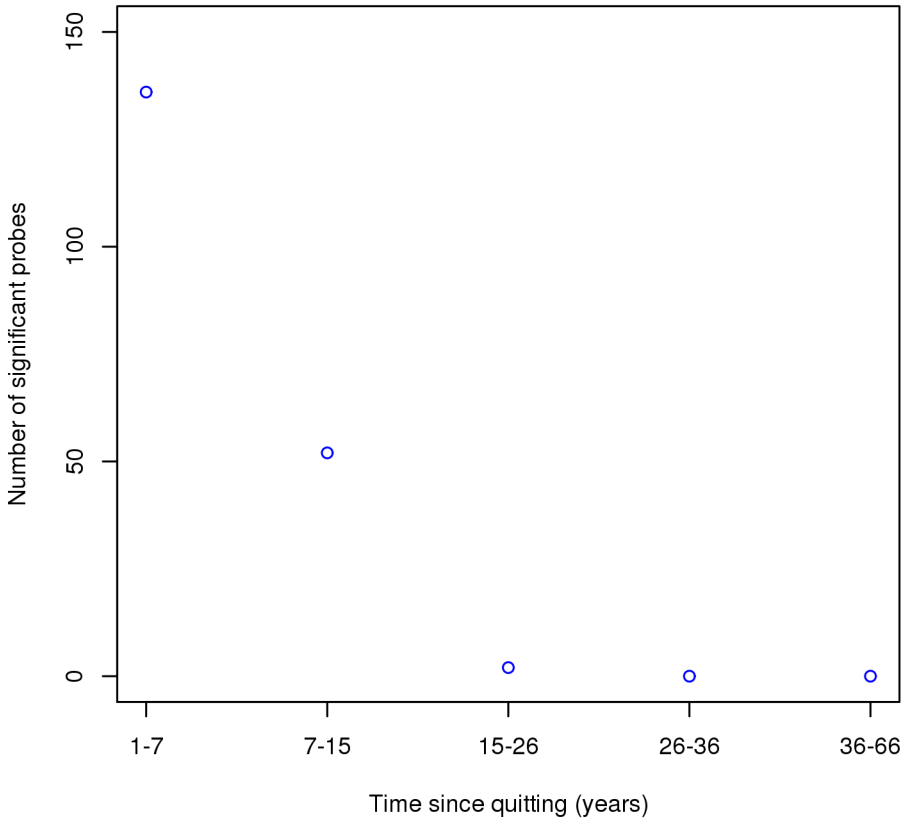

Figure 5.2.1.: Number of significantly differentially methylated probes seen between former smokers and non-smokers per cessation quantile. Each group of 60 former smokers were seperated based on years since quitting and were compared to an equal sized group of non-smokers.

142 of the 192 cessation-related DMPs were differentially methylated with smoking status in at least one group. The largest number of significant DMPs observed were the 136 seen in former smokers who had quit between 1 and 7 years ago. This number then decreased to 52 in participants who quit between 7 and 15 years and further still to just 2 DMPs in those quitting between 15 and 26 years before. Thereafter, no significantly differentially methylated CpG sites were found. This suggests that DNA methylation changes in participants with a cessation period longer than 26 years are almost completely reversed. Furthermore, the steepest drop in the number of significant DMPs occurs in the first 15 years of quitting. This shows that the differences in DNA methylation between ex-smokers and non-smokers are most apparent in those quitting for fewer years. It also suggests that the methylome may make an almost full recovery back to non-smoker levels and that this largely occurs in the first two or three decades after cessation. However, it is important to note the impact of age on DNA methylation as the mean age gradually increased in each quantile group, from 55.01 to 66.48. There was also a correlation of 0.57 between age and years of cessation as stated before. Given that many age-related differences in DNA methylation have been observed there is still some chance that the differences observed with cessation are actually caused by the complex relationship between age and DNA methylation. However, with the inclusion of age in the models, and the fact that more than half of the cessation-assocatied DMPs were not seen in the age-only model, it is likely that the decrease in DMP number represents real cessation effects on smoking-related changes to the methylome.

5.3. Quantifying Methylome Change Across Years Since Quitting

It is hard to quantify the differences in DNA methylation with years since smoking cessation by simply stating the number of significant probes that met an FDR threshold. Significance measures can be highly sensitive to the linear model in question and make it difficult to study the real differences in DNA methylation at play. To overcome the absolute effect size was calculated between each group of former smokers, split by years since quitting, and non-smokers. For each of the five groups, DNA methylation levels were averaged across the 60 former smokers used and the absolute effect size was measured between these groups and the average DNA methylation of the never smoker reference sample. Absolute effect size

is the sum of these differences for the top 10 cessation-associated probes and these are shown in Figure 5.3.1. along with the standard deviations for each quantile. For all five groups, the 10 probes were hypomethylated in former smokers compared to non-smokers. However, the degree of this hypomethylation decreases with increasing years of cessation, at least up to 36 years. In fact, there is a quite a strong linear relationship between DNA methylation differences at these 10 CpG sites and time since quitting, as shown in the first four quantiles. This suggests a strong association between DNA methylation and cessation length where absolute effect sizes return closer to 0 with increasing cessation, but only up to 36 years. The fifth quantile, containing participants who stopped smoking from 36 to 66 years ago, shows a slight increase in hypomethylation compared to the fourth quantile of participants quitting 26 to 36 years ago. This may be a consequence of the large range of cessation years within this group as few participants in *Understanding Society* had quit more than 36 years ago. This measurement may also be hindered by the higher mean age within this group.

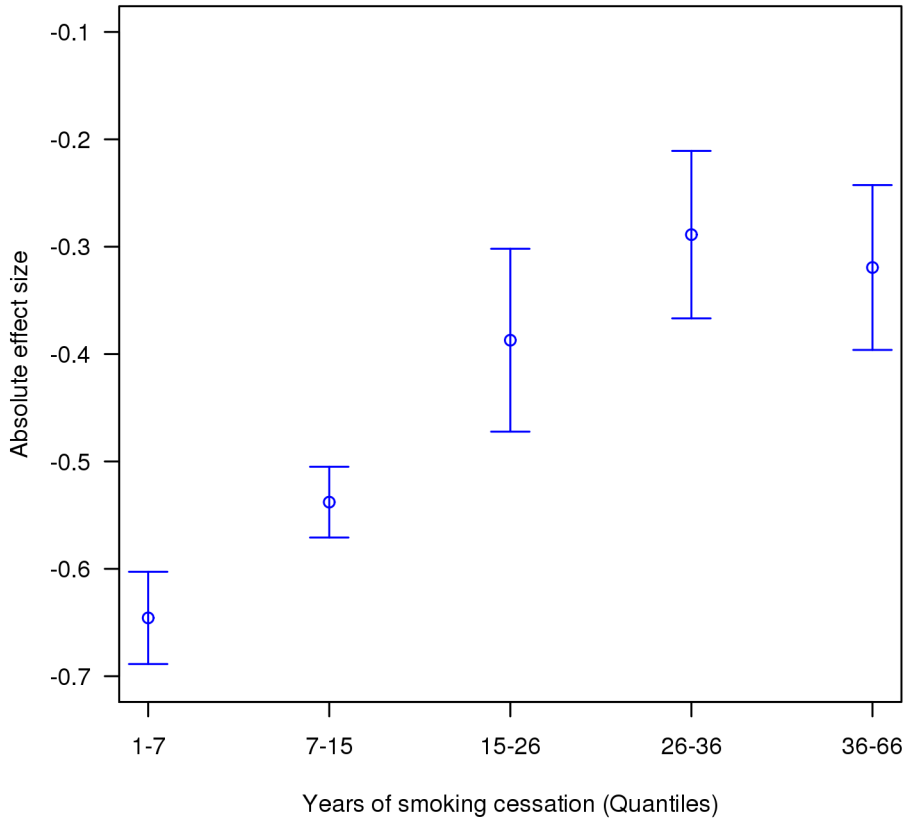

Figure 5.3.1.: Mean absolute effect size between former smokers and non-smokers. Shows mean difference between the 60 former smokers and 60 non-smokers in each group, calculated using DNA methylation values for the top 10 DMPs associated with cessation.

Clearly measuring absolute effect size provides a good insight into the real effects of cessation on DNA methylation. However, this is also sensitive to outlying data and other factors. A better way to quantify the total methylation difference of former smokers is to use these 10 sites in the creation of a smoking index (SI), detailed in the following equation from Teschendorff's work in 2015;

$$\mathrm{SI}(s) = \frac{1}{n}\sum_{c}^{n} W_c \frac{\beta_{cs} - \mu_c}{\sigma_c}$$

This gives a measure of deviation from a normal reference, standardized by the deviation of DNA methylation, and takes into account the directionality of such changes. In this case the reference group consisted of the 60 non-smokers used in the above analyses, and was created using the DNAm Beta-values for the top *n* (10) number of cessation-related probes. Here, μc is the mean beta-value, and σc the standard deviation, for each probe across the reference samples. For any given *s* sample, W_c is +1 if the probe in question is hypermethylated and -1 if hypomethylated in former smokers across years since quitting. B_{cs} is the beta value of the CpG site *c* in sample *s*. This summation is over all 10 most strongly cessation-associated DMPs and was calculated for all 356 ex-smoker participants. This is shown in Figure 5.3.2.

The computed SI scores for the majority of former smokers was negative, showing the strong influence of previous smoking on DNA methylation. In general, there is more variation in smoking index scores in the early years of smoking cessation than those who have quit for longer periods of time. This suggests that smoking may cause increased variation in DNA methylation and this has been commented on before (Ambatipudi et al., 2016). As time since quitting increases, the SI scores of former smokers tends to get closer to a score of 0 meaning DNA methylation gets closer to non-smoker levels. It also supports the finding that sites become hypermethylated with years since quitting. This is shown in the small but real correlation of 0.25 between SI scores and length of cessation. There was also an even stronger correlation of -0.55 between duration dosage and the calculated smoking index scores. This supports the associations

between DNA hypomethylation and years spent smoking and again shows the huge impact of smoking on DNA methylation as well as the complexities in reversing duration-induced changes to the methylome. It also hints at an interaction between duration and cessation in their impact on DNA methylation. Furthermore, the mean SI score for each cessation quantile was calculated and again DNA methylation change decreases across the first four quantiles, consisting of participants with 1 to 36 years of smoking cessation. Clearly this is an effective biomarker in showing the decay of changes to the methylome upon quitting smoking and this is even sensitive enough to be observed in only 10 CpG sites. It also suggests that this reversal may not fully reach non-smoker levels, even after 36 years of cessation and supports previous findings showing similar results (Guida et al., 2015).

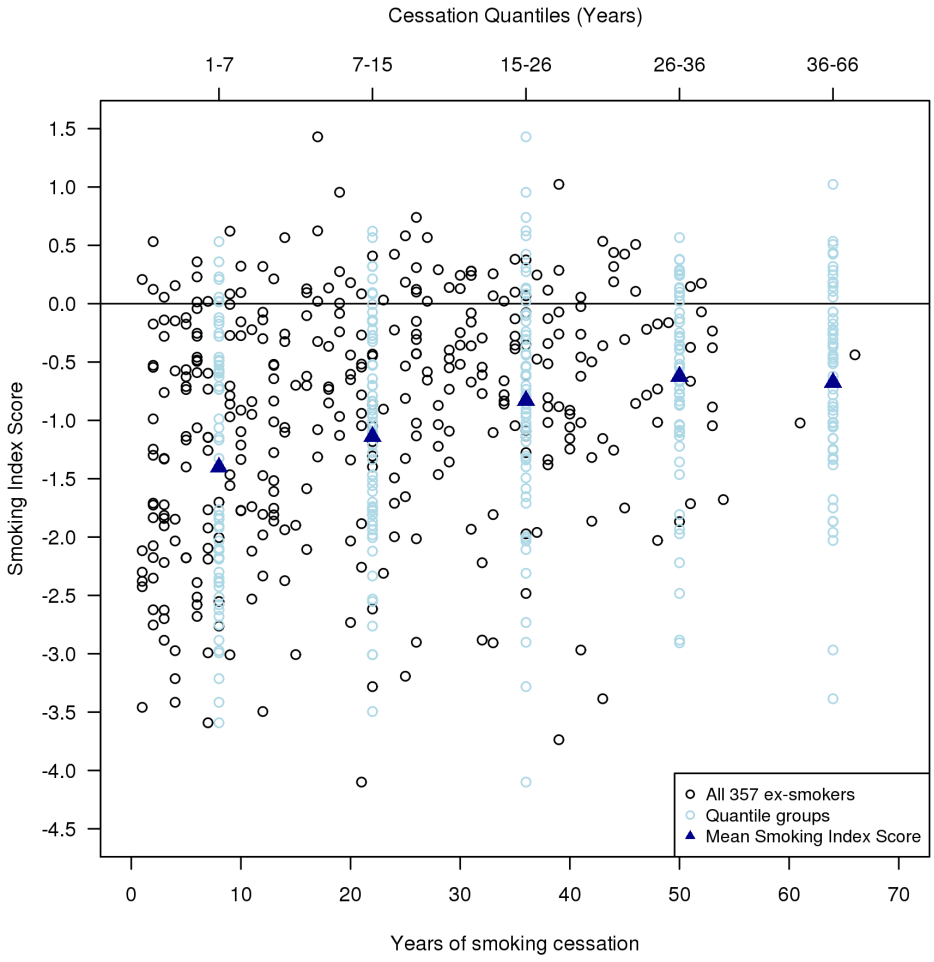

Figure 5.3.2.: Distribution of smoking index (SI) scores for all 356 former smoker partipants.. Light blue dots represent distribution of SI scores within each of the five cessation quantile groups (top axis). The mean SI score for each group was calculated and shown as a dark blue triangle.

5.4. Role of Duration in Cessation-related Decay

Multiple sites associated with cessation were also associated with duration. For these sites, most are hypomethylated with increasing duration and hypermethylation with increasing years since quitting. Furthermore, previous studies have showed a relationship at some sites, such as those in the *F2RL3* and *GPR15* genes, where methylation increases with cumulative exposure, i.e. pack years, and decreases with time since quitting (Wan et al., 2012). To try and replicate these findings in this study, the 356 former smokers were split into equal-sized groups based on quartiles for years of smoking and years since quitting. This created 16 groups with varying duration and cessation histories. Duration measures were used instead of cumulative exposure due to the weak associations of DNA methylation change with intensity.

The sum of average DNA methylation was then measured for the three smoking-associated probes located in the *F2RL3* gene, identified when comparing current and never smokers, in an attempt replicate the findings of Wan. Unfortunately, there were no participants that fell into both the fourth duration and fourth cessation quantiles and thus the average methylation could not be calculated for this group. However, for the other 15 groups there is a clear increase in DNA methylation with more years since quitting across all four duration quantile groups. The

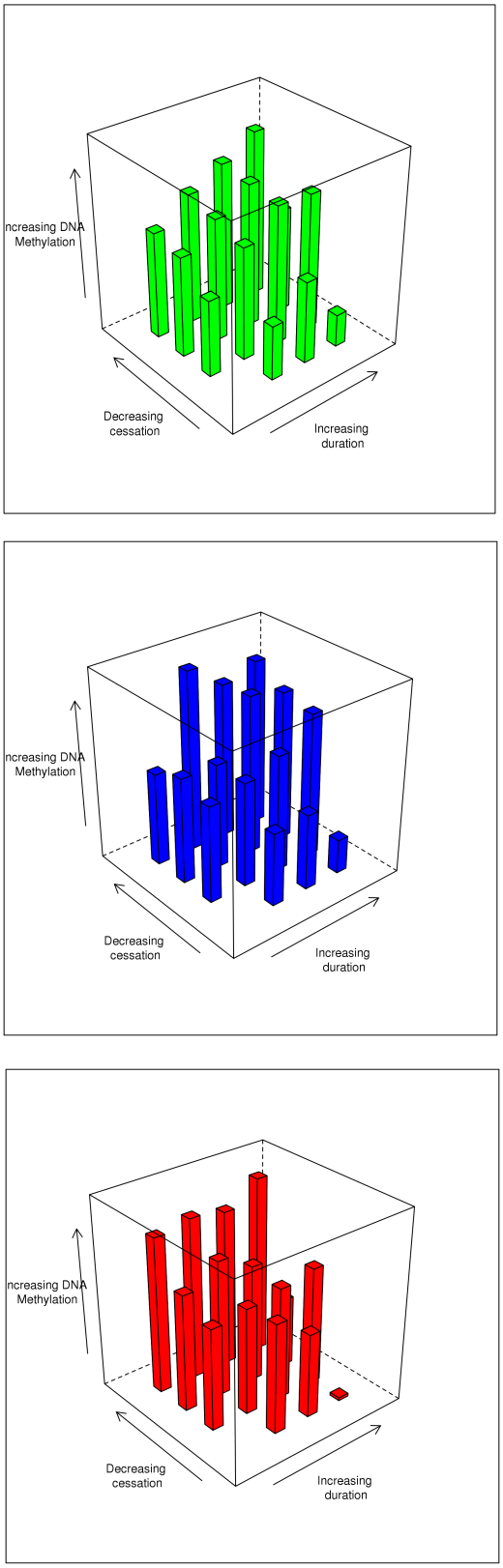

Figure 5.4.1.: Relationship between duration and cessation on DNA methylation at three loci. Average DNA methylation for each of the 16 groups, split based on duration and cessation, for smoking-associated sites located in F2RL3 (green, top), PRSS23 (blue, middle) and AHRR (red, bottom).

opposite is also true where DNA methylation is reduced with increasing years of smoking across all four cessation quantile groups. Therefore, average DNA methylation with increasing duration was greater for those who had the longest time since quitting and vice versa for increasing cessation. This shows that participants who had smoked the longest and quit for the least number of years ago had the lowest average methylation at this *F2RL3* locus. The same was also observed for the 8 smoking DMPs located on chromosome 11 in the *PRSS23* gene and the top 10 DMPs located on chromosome 5 in the *AHRR* gene. The average methylation of the 16 groups at these three loci are displayed in Figure 5.4.1. Average DNA methylation had more variation for the *PRSS23* and *AHRR* loci and thus the trend seen observed in *F2RL3* was not as clear. This may be caused by the small sample sizes within each group or perhaps different directional changes in DNA methylation on probes located in the same gene. Regardless, the obvious relationship between duration and cessation seen in the DNA methylation averages og the *F2RL3* locus show that this gene is highly sensitive to both dosage and decay. It also hints at an interaction whereby the effect of one variable on DNA methylation is different at different values of other variables. Therefore, it is important to bear this in mind when creating a biomarker of cessation as this finding suggests that the rate and degree of DNA methylation change decay following smoking cessation is dependent on the years the individual had spent smoking before quitting, at least for some sites. This also strengthens the view that the reversal of methylome changes is more effective in those who have smoked for fewer years and, given the genes impacted by smoking, also argues that the best health outcomes can be proposed for those who quit as soon as possible after smoking initiation, regardless of the intensity of smoking during those years of tobacco use.

5.5. Predicting Years Since Quitting

A clear link between DNA methylation levels and smoking cessation has been established in this chapter. This offers an opportunity to use this quantification of DNA methylation in the creation of another predictor but this time for years since smoking cessation. This was carried out using the same protocol that Horvath developed when producing his epigenetic clock of age (Horvath, 2013) and as detailed in the previous chapter when predicting smoking duration.

The penalized regression model, with elastic net regularization, consisted of regressing a transformed version of cessation years from the training set, a sample of 178 former smokers, onto the 192 EPIC array probes associated with cessation. The elastic net predictor then automatically selects the CpG sites most associated with cessation and these were then used in the prediction. Before carrying out the regression, the cessation data was transformed by taking the cube root of each value. This altered the distribution shape of the data to make it perform better when carrying out the regression and also corrected for the slight right skewness of the data. The regression model was then run using the *glmnet* package from R as before (Friedman et al., 2010). Here the alpha parameter was again set to 0.5 and the lambda value set to 0.03116628 as elucidated in a 10 fold cross validation using the same training set. The prediction of cessation was carried out using a test set consisting of the remaining former smokers and these values are compared to self-reported cessation lengths in Figure 5.5.1.

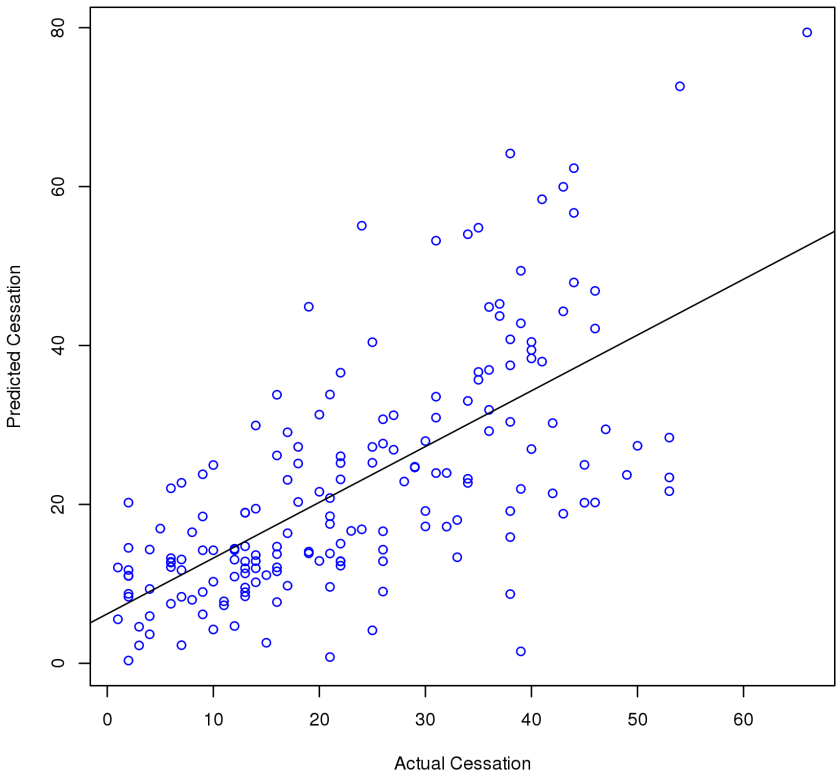

Figure 5.5.1.: Goodness of fit for predictor of years since quitting smoking. Shows actual years since quitting (x-axis) against predicted years since quitting (y-xis). Pearson correlation coefficient = 0.69.

This shows a correlation of 0.69 between the predicted and actual years since cessation, with a mean difference between the two of just 0.71 and a large standard deviation of 11.85. This large variation may

be explained by the varying duration these former smokers had spent smoking as it was observed that this can have an impact of the rate of DNA methylation change decay. It may also be the case that the reversal of DNA methylation is a multifaceted process and is more complicated than the accumulation of DNA methylation changes seen with increasing duration in current smokers. This might also explain the stronger correlation seen between predicted and actual years spent smoking despite the smaller sample size of current smokers available when creating it. Another limitation of the model comes from the lack of participants with a cessation period over 36 years. This may also explain why the model generally fits better in those with less than 36 years since quitting. With this said, the cessation predictor is still able to distinguish between those quitting for shorter or longer periods of time and this aspect may be beneficial as a biomarker of cessation. Furthermore, given the plethora of disease-related loci associated with smoking, such a biomarker may also give some indication of health outcomes for former smokers and relative risk of disease and cancer.

6. Summary

This dissertation set out to consider the impact of smoking on DNA methylation in almost 1200 participants of the *Understanding Society* household study. Firstly, a simply qualitative comparison was made using linear models in the R package *limma* (Ritchie et al., 2015) which including the covariates of age, sex, blood process day, batch and estimates of blood leukocyte proportions (Houseman et al., 2012). This was run between beta methylation values, measured in whole blood using the Infinium MethylationEPIC BeadChip (Illumina, 2016) from participants with differing smoking statuses as reported in the wave 3 questionnaires in the study. This revealed 5198 differentially methylated probes (DMPs) between current and never smokers, 826 between current and former smokers and 17 between former and never smokers after FDR adjustment for multiple testing. The majority of these loci were hypomethylated in smokers and ex-smokers compared to non-smokers and for the sites seen in all models, a reduction in their significance value and effect size was observed in each comparison. Within these DMPs came some novel probes located in genes whose DNA methylation state had not yet been linked to smoking. This included *GNMT*, related to the methyl donor for most cellular methylation reactions (Yen et al., 2013), and *SLAMF7*, a self-ligand receptor in the signalling lymphocytic activation molecule family that is important in immunity (Guo et al., 2015). What this analysis could not show is any smoking-related variation in DNA methylation within active smoker and ex-smoker participants. To overcome this, a second set of analyses used quantitative data that detailed information on dosage, including the years spent smoking and the mean number of cigarettes smoked per day by participants. This was limited to participants who currently smoke given the reversal of DNA methylation changes observed in the qualitative analysis which may have hindered the linear models. Here a strong correlation between age and smoking duration became apparent and thus had to be removed from the model which then showed 1331 DMPs associated with years spent smoking. The intensity measure only yielded 23 DMPs. Additionally, effect sizes rose with increasing measures of both dosages however this was more strongly influenced by duration. Taken together this suggested that years of smoking is more closely linked to DNA methylation changes than number of cigarettes smoked. It also hints that duration rather than intensity is important in smoking-related disease as several genes associated with years spent smoking are crucial for good health. This suggested that DNA methylation levels might

offer potential as a biomarker for smoking years. Such a predictor was then created using a similar penalized regression approach, with elastic net regularization, to Horvath in his epigenetic clock of age. This was carried out in the *glmnet* R package (Friedman et al., 2010) where the predicted lengths of duration had a good correlation of 0.76 with the actual years spent smoking. The third and final analysis considered cessation and how it is involved in the decay of differential methylation between current and former smokers when compared to never smokers. A linear model yielded 192 DMPs associated with years since quitting in former smokers, most of which were hypermethylated. Looking at effect sizes, this hypermethylation of probes increased with increasing cessation, at least up to 36 years. The degree of reversed differential methylation was then quantified for each participant who had previously smoked by creating a smoking index (SI) using an equation created by Teschendorff et al. (2015). This measured the deviation of DNA methylation of the top 10 cessation-related probes from a non-smoker reference sample. In general, this showed that former smokers with the longest years since quitting had SI scores closer to zero than those with shorter cessation periods. As DNA methylation was clearly shown here to reflect the length of cessation of the participant, another predictor was created, using the same procedure as before, but this time for years since quitting. This showed a slightly weaker correlation of 0.69 between predicted and actual values for cessation but was still able to distinguish between long-term and short-term ex-smokers.

The work outlined in this dissertation has not only considered but perhaps advanced understandings of the relationship between DNA methylation and smoking. This has been done by identifying novel associations between DNA methylation at many CpG sites and tobacco use and further still through the creation of predictors capable of estimating years of duration in current smokers and years of cessation in former smokers. Thus, the aims of this project were met and has reinforced previous findings in this field of study.

References

1. Abboud, R.T. and Vimalanathan, S. (2008) Pathogenesis of COPD. Part I. The role of protease-antiprotease imbalance in emphysema [State of the Art Series. Chronic obstructive pulmonary disease in high-and low-income countries. Edited by G. Marks and M. Chan-Yeung. Number 3 in the series]. *The international journal of tuberculosis and lung disease*, **12**(4), 361-367.
2. Albrecht, W., Santis, M.D. and Dossenbach-Glaninger, A. (2004) Testicular tumor markers: Corner-stones in the management of malignant germ cell tumors/Hoden-Tumor-marker: Eckpfeiler in der Behandlung maligner Keimzelltumoren. *LaboratoriumsMedizin*, **28**(2), 109-115.
3. Ambatipudi, S., Cuenin, C., Hernandez-Vargas, H., Ghantous, A., Calvez-Kelm, L., Kaaks, R., Barrdahl, M., Boeing, H., Aleksandrova, K., Trichopoulou, A., Lagiou, P., Naska, A., Palli, D., Krogh, V., Polidoro, S., Tumino, R., Panico, S., Bueno- de-Mesquita, B., Peeters, P., Quirós, J., Navarro, C., Ardanaz, E., Dorronsoro, M., Key, T., Vineis, P., Murphy, N., Riboli, E., Romieu, I. and Herceg, Z. (2016) Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics*, **8**(5), 599–618.
4. Bantscheff, M., Hopf, C., Savitski, M.M., Dittmann, A., Grandi, P., Michon, A.M., Schlegl, J., Abraham, Y., Becher, I., Bergamini, G. and Boesche, M. (2011) Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. *Nature biotechnology*, **29**(3), 255-265.
5. Barua, R.S. and Ambrose, J. A. (2013) Mechanisms of Coronary Thrombosis in Cigarette Smoke Exposure. *Arteriosclerosis, Thrombosis, and Vascular Biology*, **33**(7), 1460-1467.
6. Bauer, M., Fink, B., Thürmann, L., Eszlinger, M., Herberth, G. and Lehmann, I. (2016) Tobacco smoking differently influences cell types of the innate and adaptive immune system—indications from CpG site methylation. *Clinical Epigenetics*. **8**(1), 83.
7. Benjamini, Y. and 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), 289-300.
8. Besingi, W. and Johansson, A. (2013) Smoke-related DNA methylation changes in the etiology of human disease. *Human molecular genetics*. **23**(9), 2290–7.
9. Bestor, T. (2000) The DNA methyltransferases of mammals. *Human Molecular Genetics*, **9**(16), 2395-2402.
10. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L. and Fan, J.B. (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**(4), 288-295.
11. Bochtler, M., Kolano, A. and Xu, G. (2016) DNA demethylation pathways: Additional players and regulators. *BioEssays*. **39**(1), 1600178.
12. Botteri, E., Iodice, S., Bagnardi, V., Raimondi, S., Lowenfels, A.B. and Maisonneuve, P. (2008) Smoking and colorectal cancer: a meta-analysis. *Jama*, **300**(23), 2765-2778.
13. Breitling, L., Yang, R., Korn, B., Burwinkel, B. and Brenner, H. (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *American journal of human genetics*. **88**(4): 450–7.
14. Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A., Socci, N. and Scandura, J. (2011) DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing. *PLoS ONE*. **6**(1), e14524.
15. Buck, N. and McFall, S. (2011) Understanding Society: design overview. *Longitudinal and Life Course Studies*, **3**(1), 5-17.
16. Charlesworth, J.C., Curran, J.E., Johnson, M.P., Göring, H.H., Dyer, T.D., Diego, V.P., Kent, J.W., Mahaney, M.C., Almasy, L., MacCluer, J.W. and Moses, E.K. (2010) Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. *BMC medical genomics*, **3**(1), 29.
17. Chiappinelli, K.B., Strissel, P.L., Desrichard, A., Li, H., Henke, C., Akman, B., Hein, A., Rote, N.S., Cope, L.M., Snyder, A. and Makarov, V. (2015) Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell*, **162**(5), 974-986.
18. Citterio L, Simonini M, Zagato L, et al. (2011) Genes involved in vasoconstriction and vasodilation system affect salt sensitive hypertension. *PLoS One*, **6**:e19620.
19. Cooper, N. (2017). humarray: Simplify Analysis and Annotation of Human Microarray Datasets. R package version 1.1. https://CRAN.R-project.org/package=humarray
20. Cuozzo, C., Porcellini, A., Angrisano, T., Morano, A., Lee, B., Di Pardo, A., Messina, S., Iuliano, R., Fusco, A., Santillo, M.R. and Muller, M.T. (2007) DNA damage, homology-directed repair, and DNA methylation. *PLoS genetics*, **3**(7), e110.
21. Dogan, M., Shields, B., Cutrona, C., Gao, L., Gibbons, F., Simons, R., Monick, M., Brody, G., Tan, K., Beach, S. and Philibert, R. (2014) The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. *BMC genomics*, **15**(1), 151.
22. Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2004) Mortality in relation to smoking: 50 years' observations on male British doctors. *British Medical Journal*, **328**(7455), 1519.

23. Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L. and Lin, S.M. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**(1), 587.
24. Dunn, O. J. (1961) Multiple comparisons among means. *Journal of the American Statistical Association*, **56**(293), 52-64.
25. Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., Haefliger, C., Horton, R., Howe, K., Jackson, D.K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C, West, T., Rogers, J., Olek, A., Berlin, K., Beck, S. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, **38**, 1378–1385.
26. Eeden, S. V., & Hogg, J. (2000) The response of human bone marrow to chronic cigarette smoking. *European Respiratory Journal*, **15**(5), 915-921.
27. Ehrlich, M. (2009) DNA hypomethylation in cancer cells. *Epigenomics*, **1**(2), 239-259.
28. Elliott, H., Tillin, T., McArdle, W., Ho, K., Duggirala, A., Frayling, T., Smith, D., Hughes, A., Chaturvedi, N. and Relton, C. (2014) Differences in smoking associated DNA methylation patterns in south Asians and Europeans, *Clinical epigenetics*. **6**(1), 4.
29. Evans, B.R., Karchner, S.I., Allan, L.L., Pollenz, R.S., Tanguay, R.L., Jenny, M.J., Sherr, D.H. and Hahn, M.E. (2008) Repression of aryl hydrocarbon receptor (AHR) signaling by AHR repressor: role of DNA binding and competition for AHR nuclear translocator. *Molecular pharmacology*, **73**(2), 387-398.
30. Filzmoser, P., Maronna, R. and Werner, M. (2008) Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, **52**(3): 1694-1711.
31. Fiorentino, D., Chung, L., Zwerner, J., Rosen, A. and Casciola-Rosen, L. (2011) The mucocutaneous and systemic phenotype of dermatomyositis patients with antibodies to MDA5 (CADM-140): a retrospective study. *Journal of the American Academy of Dermatology*, **65**(1), 25-34.
32. Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1), 1-22. URL: http://www.jstatsoft.org/v33/i01/.
33. Gorrie-Stone TJ, Saffari A, Malki K and Schalkwyk LC (2017). bigmelon: Illumina methylation array analysis for large experiments. R package version 1.2.0.
34. Greco A, Rizzo MI, De Virgilio A, Gallo A, Fusconi M and Pagliuca G. (2015) Goodpasture's syndrome: a clinical update. *Autoimmunity Reviews*, **14**:246–53.
35. Guida, F., Sandanger, T., Castagné, R., Campanella, G., Polidoro, S., Palli, D., Krogh, V., Tumino, R., Sacerdote, C., Panico, S., Severi, G., Kyrtopoulos, S., Georgiadis, P., Vermeulen, R., Lund, E., Vineis, P. and Chadeau-Hyam, M. (2015) Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Human molecular genetics*. **24**(8): 2349–59.
36. Guo, H., Cruz-Munoz, M.E., Wu, N., Robbins, M. and Veillette, A. (2015) Immune cell inhibition by SLAMF7 is mediated by a mechanism requiring src kinases, CD45, and SHIP-1 that is defective in multiple myeloma cells. *Molecular and cellular biology*, **35**(1), 41-51.
37. Guth, S. I. E., & Wegner, M. (2008) Having it both ways: Sox protein function between conservation and innovation. *Cellular and molecular life sciences*, **65**(19), 3000-3018.
38. Hadchouel A, Durrmeyer X, Bouzigon E, Incitti R, Huusko J, Jarreau P-H. (2011) Identification of SPOCK2 as a susceptibility gene for bronchopulmonary dysplasia. *American Journal of Respiratory and Critical Care Medicine*, 184, 1164–70.
39. Han, L., Lin, I.G. and Hsieh, C.L. (2001) Protein binding protects sites on stable episomes and in the chromosome from de novo methylation. *Molecular and cellular biology*, **21**(10), 3416-3424.
40. Horvath, S. (2013) DNA methylation age of human tissues and cell types. *Genome biology*, **14**(10), 3156.
41. Houseman, E., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K. and Kelsey, K. T. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, **13**(1), 86.
42. Hällfors, J. (2017) Nicotine Dependence–Identifying the Contribution of Specific Genes. Academic Dissertation. University of Helsinki, Institute for Molecular Medicine Finland, Finland.
43. Ingebrigtsen, T.S., Thomsen, S.F., van der Sluis, S., Miller, M., Christensen, K., Sigsgaard, T. and Backer, V. (2011) Genetic influences on pulmonary function: a large sample twin study. *Lung*, **189**(4), 323-330.
44. Isik, B., Ceylan, A. and Isik, R. (2007) Oxidative stress in smokers and non-smokers. *Inhalation toxicology*, **19**(9), 767-769.
45. Issa, J.P. (1999) Aging, DNA methylation and cancer. *Critical reviews in oncology/hematology*, **32**(1), 31-43.
46. Itoh, T., Fairall, L., Muskett, F.W., Milano, C.P., Watson, P.J., Arnaudo, N., Saleh, A., Millard, C.J., El-Mezgueldi, M., Martino, F. and Schwabe, J.W. (2015) Structural and functional characterization of a cell cycle associated HDAC1/2 complex reveals the structural basis for complex assembly and nucleosome targeting. *Nucleic acids research*, **43**(4), 2033-2044.
47. Joehanes, R., Just, A., Marioni, R., Pilling, L., Reynolds, L., Mandaviya, P., Guan, W., Xu, T., Elks, C., Aslibekyan, S., Moreno-Macias, H., Smith, J., Brody, J., Dhingra, R., Yousefi, P., Pankow, J., Kunze, S., Shah,

S., McRae, A., Lohman, K., Sha, J., Absher, D., Ferrucci, L., Zhao, W., Demerath, E., Bressler, J., Grove, M., Huan, T., Liu, C., Mendelson, M., Yao, C., Kiel, D., Peters, A., Wang-Sattler, R., Visscher, P., Wray, N., Starr, J., Ding, J., Rodriguez, C., Wareham, N., Irvin, Zhi, D., Barrdahl, M., Vineis, P., Ambatipudi, S., Uitterlinden, A., Hofman, A., Schwartz, J., Colicino, E., Hou, L., Vokonas, P., Hernandez, D., Singleton, A., Bandinelli, S., Turner, S., Ware, E., Smith, A., Klengel, T., Binder, E., Psaty, B., Taylor, K., Gharib, S., Swenson, B., Liang, L., DeMeo, D., O'Connor, G., Herceg, Z., Ressler, K., Conneely, K., Sotoodehnia, N., Kardia, S., Melzer, D., Baccarelli, A., Meurs, van, Romieu, I., Arnett, D., Ong, K., Waldenberger, M., Deary, I., Fornage, M., Levy, D. and London, S. (2016) Epigenetic signatures of cigarette smoking. *Cardiovascular genetics*, **9** (5), 436–447.

48. Kastan M.B. (2008) DNA damage responses: mechanisms and roles in human disease: 2007 G.H.A. Clowes Memorial Award Lecture. *Molecular Cancer Research*, **6**(4), 517–24.
49. Kim, D.-H., Kim, J. S., Ji, Y.-I., Shim, Y. M., Kim, H., Han, J. and Park, J. (2003) Hypermethylation of RASSF1A promoter is associated with the age at starting smoking and a poor prognosis in primary non-small cell lung cancer. *Molecular Biology and Genetics*, **63**(13), 3743–3746.
50. Koks G, Uudelepp ML, Limbach M, Peterson P, Reimann E and Koks S. (2015) Smoking- Induced Expression of the GPR15 Gene Indicates Its Potential Role in the Chronic Inflammatory Pathologies. *American Journal of Pathology*, **185**(11), 2898-2906.
51. Koshida, K., Stigbrand, T., Munck-Wikland, E., Hisazumi, H. and Wahren, B. (1990) Analysis of serum placental alkaline phosphatase activity in testicular cancer and cigarette smokers. *Urological research*, **18**(3), 169-173.
52. Koutsimpelas, D., Pongsapich, W., Heinrich, U., Mann, S., Mann, W. J., and Brieger, J. (2012) Promoter methylation of MGMT, MLH1 and RASSF1A tumor suppressor genes in head and neck squamous cell carcinoma: Pharmacological genome demethylation reduces proliferation of head and neck squamous carcinoma cells. *Oncology Reports*, **27**(4), 1135-1141.
53. Lankisch PG, Apte M, Banks PA, (2015) Acute pancreatitis. *Lancet*. **386**, 85–96.
54. Lee, J., Taneja, V. and Vassallo, R. (2012) Cigarette smoking and inflammation: cellular and molecular mechanisms. *Journal of dental research*, **91**(2), 142-149.
55. Lee, M., Hong, Y., Kim, S., London, S. and Kim, W. (2016) DNA methylation and smoking in Korean adults: Epigenome-wide association study. *Clinical Epigenetics*, **8**(1):103.
56. Leenen, F. A., Muller, C. P., & Turner, J. D. (2016). DNA methylation: conducting the orchestra from exposure to phenotype?. *Clinical epigenetics*, **8**(1), 92.
57. Levin, H.L. and Moran, J.V. (2011) Dynamic interactions between transposable elements and their hosts. *Nature Reviews Genetics*, **12**(9), 615-627.
58. Ligthart, S., Steenaard, R.V., Peters, M.J., van Meurs, J.B., Sijbrands, E.J., Uitterlinden, A.G. and Bonder, M.J. (2016) Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes. *Diabetologia*, **59**, 998.
59. Lisanti, S., Omar, W. A., Tomaszewski, B., De Prins, S., Jacobs, G., Koppen, G., and Langie, S. A. (2013) Comparison of methods for quantification of global DNA methylation in human cells and tissues. *PLoS One*, **8**(11), e79044.
60. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M. and Edsall, L. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**(7271), 315-322.
61. Liu, Q., Liu, L., Zhao, Y., Zhang, J., Wang, D., Chen, J., He, Y., Wu, J., Zhang, Z. and Liu, Z. (2011) Hypoxia induces genomic DNA demethylation through the activation of HIF-1α and transcriptional upregulation of MAT2A in hepatoma cells. *Molecular cancer therapeutics*.
62. Lohse, M.J., Engelhardt, S. and Eschenhagen, T. (2003) What is the role of β-adrenergic signaling in heart failure?. *Circulation research*, **93**(10), 896-906.
63. Luka, Z., Mudd, S.H. and Wagner, C. (2009) Glycine N-methyltransferase and regulation of S-adenosylmethionine levels. *Journal of Biological Chemistry*, **284**(34), 22507-22511.
64. Mak, C. H., Li, Z., Allen, C. E., Liu, Y., & Wu, L. C. (1998) KRC transcripts: identification of an unusual alternative splicing event. *Immunogenetics*, **48**(1), 32-39.
65. McLean, C., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010) GREAT improves functional interpretation of cis-regulatory regions. Nature *Biotechnology*, **28**(5), 495-501.
66. Mehta, H., Nazzal, K. and Sadikot, R.T. (2008) Cigarette smoking and innate immunity. *Inflammation Research*, **57**(11), 497-503.
67. Mitra, A., Chakraborty, B., Mukhopadhay, D., Pal, M., Mukherjee, S., Banerjee, S. and Chaudhuri, K. (2012) Effect of smoking on semen quality, FSH, testosterone level, and CAG repeat length in androgen receptor gene of infertile men in an Indian city. *Systems biology in reproductive medicine*, **58**(5), 255-262.

68. Monge, M., Colas, E., Doll, A., Gil-Moreno, A., Castellvi, J., Diaz, B., Gonzalez, M., Lopez-Lopez, R., Xercavins, J., Carreras, R. and Alameda, F. (2009) Proteomic approach to ETV5 during endometrial carcinoma invasion reveals a link to oxidative stress. *Carcinogenesis*, **30**(8), 1288-1297.
69. Moore, L.D., Le, T. and Fan, G. (2013) DNA methylation and its basic function. *Neuropsychopharmacology*, **38**(1), 23-38.
70. Needham, M. (2004) 1-Antitrypsin deficiency * 3: Clinical manifestations and natural history. *Thorax*, **59**(5), 441-445.
71. Neitzel, J.J. (2010) Enzyme catalysis: the serine proteases. *Nature Education*, **3**(9), 21.
72. NIH (1964) Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service.
73. Nishihara, R., Morikawa, T. and Kuchiba, A. (2013) A prospective study of duration of smoking cessation and Colorectal cancer risk by Epigenetics-related tumor classification. *American journal of epidemiology*, **178**(1), 84-100.
74. Ohno, R., Nakayama, M., Naruse, C., Okashita, N., Takano, O., Tachibana, M., Asano, M., Saitou, M. and Seki, Y. (2013) A replication-dependent passive mechanism modulates DNA demethylation in mouse primordial germ cells. *Development*, **140**(14), 2892-2903.
75. Olety, B., Wälte, M., Honnert, U., Schillers, H. and Bähler, M. (2010) Myosin 1G (Myo1G) is a haematopoietic specific myosin that localises to the plasma membrane and regulates cell elasticity. *FEBS letters*, **584**(3), 493-499.
76. Peters, T.J., Buckley, M.J., Statham, A.L., Pidsley, R., Samaras, K., Lord, R.V., Clark, S.J. and Molloy, P.L. (2015) De novo identification of differentially methylated regions in the human genome. *Epigenetics & chromatin*, **8**(1): 6.
77. Peto, J. (2011) That lung cancer incidence falls in ex-smokers: misconceptions 2. *British journal of cancer*, **104**(3): 389.
78. Peto, J. (2012) That the effects of smoking should be measured in pack-years: misconceptions 4.
79. Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E. and Doll, R. (2000) Smoking, smoking cessation, and lung cancer in the UK since 1950: Combination of national statistics with two case-control studies. *British Medical Journal*, **321**(7257), 323–329.
80. Philibert, R., Beach, S. and Brody, G. (2012) Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. *Epigenetics*, **7**(11), 1331–8.
81. Philibert, R., Hollenbeck, N., Andersen, E., McElroy, S., Wilson, S., Vercande, K., Beach, S.R., Osborn, T., Gerrard, M., Gibbons, F.X. and Wang, K. (2016) Reversion of AHRR Demethylation is a quantitative biomarker of smoking cessation. *Frontiers in psychiatry*, **7**.
82. Phipson, B., Lee, S., Majewski, I.J., Alexander, W.S. and Smyth, G.K. (2016) Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The Annals of Applied Statistics*, **10**(2), 946-963.
83. Pidsley, R., Wong, C. C., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics*, **14**(1), 293.
84. Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C. and Clark, S.J. (2016) Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, **17**(1), 208.
85. Proctor, R. N. (1996) The anti-tobacco campaign of the Nazis: a little known aspect of public health in Germany, 1933-45. *British Medical Journal*, **313**(7070), 1450–1453.
86. R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
87. Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., & Jaenisch, R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(10), 5237–5242.
88. Reynolds, L., Magid, H., Chi, G., Lohman, K., Barr, R., Kaufman, J., Hoeschele, I., Blaha, M., Navas-Acien, A. and Liu, Y. (2016) Secondhand tobacco smoke exposure associations with DNA Methylation of the aryl hydrocarbon receptor Repressor. *Nicotine & tobacco research:official journal of the Society for Research on Nicotine and Tobacco*, **ntw**219.
89. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, **p.gkv**007.
90. Robertson, K.D. (2005) DNA methylation and human disease. *Nature reviews. Genetics*, **6**(8), 597.
91. Russell, R.E., Thorley, A., Culpitt, S.V., Dodd, S., Donnelly, L.E., Demattos, C., Fitzgerald, M. and Barnes, P.J. (2002) Alveolar macrophage-mediated elastolysis: roles of matrix metalloproteinases, cysteine, and serine proteases. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, **283**(4), L867-L873.

92. Satta, R., Maloku, E., Zhubi, A., Pibiri, F., Hajos, M., Costa, E. and Guidotti, A. (2008) Nicotine decreases DNA methyltransferase 1 expression and glutamic acid decarboxylase 67 promoter methylation in GABAergic interneurons. *Proceedings of the National Academy of Sciences*, **105**(42), 16356-16361.
93. Satterthwaite, F.E. (1946) An approximate distribution of estimates of variance components. *Biometrics bulletin*, **2**(6), 110-114.
94. Schmoll, H.J., Souchon, R., Krege, S., Albers, P., Beyer, J., Kollmannsberger, C., Fossa, S.D., Skakkebaek, N.E., De Wit, R., Fizazi, K. and Droz, J.P. (2004) European consensus on diagnosis and treatment of germ cell cancer: a report of the European Germ Cell Cancer Consensus Group (EGCCCG). *Annals of Oncology*, **15**(9), 1377-1399.
95. Sharma, M., Li, X., Wang, Y., Zarnegar, M., Huang, C.Y., Palvimo, J.J., Lim, B. and Sun, Z. (2003) hZimp10 is an androgen receptor co-activator and forms a complex with SUMO-1 at replication foci. *The EMBO journal*, **22**(22), 6101-6114.
96. Shenker, N., Ueland, P., Polidoro, S., Veldhoven, van, Ricceri, F., Brown, R., Flanagan, J. and Vineis, P. (2013) DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology (Cambridge, Mass.),* **24**(5), 712–6.
97. Slotkin, T.A. (2004) Cholinergic systems in brain development and disruption by neurotoxicants: nicotine, environmental tobacco smoke, organophosphates. *Toxicology and applied pharmacology*, **198**(2), 132-151.
98. Smyth, G. (2005) Limma: linear models for microarray data. *Bioinformatics and computational biology solutions using R and Bioconductor*, 397-420.
99. Staaf, J., Vallon-Christersson, J., Lindgren, D., Juliusson, G., Rosenquist, R., Höglund, M., Ringnér, M. (2008) Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics*, **9**(1), 409.
100. Steemers, F., Chang, W., Lee, G., Barker, D., Shen, R. and Gunderson, K. (2006) Whole-genome genotyping with the single-base extension assay. *Nature Methods*, **3**(1): 31-33.
101. Stephens, J.W. and Humphries, S.E. (2003) The molecular genetics of cardiovascular disease: clinical implications. *Journal of internal medicine*, **253**(2): 120-127.
102. Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, **9**(6), 465-476.
103. Takeuchi, O. and Akira, S. (2008) MDA5/RIG-I and virus recognition. *Current opinion in immunology*, **20**(1), 17-22.
104. Tefferi, A. (2014) Primary myelofibrosis: 2014 update on diagnosis, risk-stratification, and management. *American Journal of Hematology*, **89**(9), 915-925.
105. Teschendorff, A., Yang, Z., Wong, A., Pipinikas, C., Jiao, Y., Jones, A., Anjum, S., Hardy, R., Salvesen, H., Thirlwell, C., Janes, S., Kuh, D. and 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.
106. Tetzner, R., Model, F., Weiss, G., Schuster, M., Distler, J., Steiger, K.V., Grützmann, R., Pilarsky, C., Habermann, J.K., Fleshner, P.R. and Oubre, B.M. (2009) Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. *Clinical chemistry*, **55**(7), 1337-1346.
107. Toorn, M. V., Vries, M. P., Slebos, D., Bruin, H. G., Abello, N., Oosterhout, A. J., Kauffman, H. F. (2007) Cigarette smoke irreversibly modifies glutathione in airway epithelial cells. *AJP: Lung Cellular and Molecular Physiology*, **293**(5).
108. Tsaprouni, L., Yang, T., Bell, J., Dick, K., Kanoni, S., Nisbet, J., Viñuela, A., Grundberg, E., Nelson, C., Meduri, E., Buil, A., Cambien, F., Hengstenberg, C., Erdmann, J., Schunkert, H., Goodall, A., Ouwehand, W., Dermitzakis, E., Spector, T., Samani, N. and Deloukas, P. (2014) Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics*, **9**(10), 1382–96.
109. Vadivel, A., Alphonse, R. S., Collins, J. J., Haaften, T. V., O'Reilly, M., Eaton, F., & Thébaud, B. (2013) The Axonal Guidance Cue Semaphorin 3C Contributes to Alveolar Growth and Repair. *PLoS ONE*, **8**(6).
110. Vitoux, D., Nasr, R. and de The, H. (2007) Acute promyelocytic leukemia: New issues on pathogenesis and treatment response. *The international journal of biochemistry & cell biology*, **39**(6), 1063-1070.
111. Vlaanderen, J., Portengen, L., Schüz, J., Olsson, A., Pesch, B., Kendzia, B., Stücker, I., Guida, F., Brüske, I., Wichmann, H.-E., Consonni, D., Landi, M. T., Caporaso, N., Siemiatycki, J., Merletti, F., Mirabelli, D., Richiardi, L., Gustavsson, P., Plato, N., Jöckel, K.-H., Ahrens, W., Pohlabeln, H., Tardón, A., Zaridze, D., Field, J. K., 't Mannetje, A., Pearce, N., McLaughlin, J., Demers, P., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Dumitru, R. S., Bencko, V., Foretova, L., Janout, V., Boffetta, P., Forastiere, F., Bueno-de-Mesquita, B., Peters, S., Brüning, T., Kromhout, H., Straif, K. and Vermeulen, R. (2014) Effect modification of the association of cumulative exposure and cancer risk by intensity of exposure and time since exposure cessation: A flexible method applied to cigarette smoking and lung cancer in the SYNERGY study. *American journal of epidemiology*, **179**(3), 290-298.

112. Wan, E., Qiu, W., Baccarelli, A., Carey, V., Bacherman, H., Rennard, S., Agusti, A., Anderson, W., Lomas, D. and Demeo, D. (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Human molecular genetics*, **21**(13), 3073–82.
113. Ware, J.J., van den Bree, M. and Munafò, M.R. (2012) From men to mice: CHRNA5/CHRNA3, smoking behavior and disease. *Nicotine & Tobacco Research*, **14**(11), 1291-1299.
114. Weisenberger, DJ. et al. (2008) Comprehensive DNA Methylation Analysis on the Illumina Infinium Assay Platform.
115. Xie, X.T., Liu, Q., Wu, J. and Wakui, M. (2009) Impact of cigarette smoking in type 2 diabetes development. *Acta Pharmacologica Sinica*, **30**(6), 784-787.
116. Yen, C.H., Lin, Y.T., Chen, H.L., Chen, S.Y. and Chen, Y.M.A. (2013) The multi-functional roles of GNMT in toxicology and cancer. *Toxicology and applied pharmacology*, **266**(1), 67-75.
117. Zaghlool, S., Al-Shafai, M., Muftah, A., Kumar, P., Falchi, M. and Suhre, K. (2015) Association of DNA methylation with age, gender, and smoking in an Arab population. *Clinical epigenetics*, **7**(1), 6.
118. Zhang, Y., Florath, I., Saum, K. and Brenner, H. (2016) Self-reported smoking, serum cotinine, and blood DNA methylation. *Environmental research*, **146**, 395–403.
119. Zhang, Y., Yang, R., Burwinkel, B., Breitling, L.P. and Brenner, H. (2015) F2RL3 methylation as a biomarker of current and lifetime smoking exposures. *Environmental Health Perspectives (Online)*, **122**(2), 131.
120. Zhu, X., Li, J., Deng, S., Yu, K., Liu, X., Deng, Q., Sun, H., Zhang, X., He, M., Guo, H., Chen, W., Yuan, J., Zhang, B., Kuang, D., He, X., Bai, Y., Han, X., Liu, B., Li, X., Yang, L., Jiang, H., Zhang, Y., Hu, J., Cheng, L., Luo, X., Mei, W., Zhou, Z., Sun, S., Zhang, L., Liu, C., Guo, Y., Zhang, Z., Hu, F. B., Liang, L. and Wu, T. (2016) Genome-wide analysis of DNA Methylation and cigarette smoking in a Chinese population. *Environmental health perspectives*, **124**(7), 966.