Blood DNA Methylation Biomarkers for Breast Cancer Risk

Kevin Brennan

A thesis submitted for the degree of Doctor of Philosophy

Division of Cancer, Department of Surgery and Cancer, Imperial College London

September 2013
Abstract

Breast cancer is the most common malignancy affecting women worldwide with an average lifetime risk of 12%. Risk is affected by age, family history, genetics, reproductive factors and environmental exposures, however many unknown risk factors may exist. Regular screening, lifestyle advice and preventative therapy may be offered to women at highest risk; however in the absence of high-penetrance mutations, personal breast cancer risk cannot be accurately estimated. Risk biomarkers are therefore required to help improve current risk prediction models. Epigenetic mechanisms control gene expression and genome function, and are influenced by both heritable and environmental factors. DNA methylation, the most widely studied epigenetic mark, is widely deregulated in cancer and cancer precursor lesions; however the contribution to disease risk of DNA methylation variability in normal tissue prior to disease is poorly understood. Several blood DNA methylation markers associated with cancer have been reported, including genome-wide hypomethylation and hypermethylation of the ATM gene associated with breast cancer, however, validation of these associations in samples collected prior to diagnosis (prospectively collected) are required to determine association with breast cancer risk.

The aims of this thesis were to 1) validate ATM methylation as a breast cancer risk marker in three nested case-control studies from prospective cohorts; 2) To investigate hypomethylation of LINE1 in the same prospective cohorts and compare this in a meta-analysis with all other published LINE1 data; 3) To investigate potential mechanisms or modifiers of ATM methylation; 4) To perform discovery microarray studies to identify novel DNA methylation markers of breast cancer risk. Herein, we show that ATM hypermethylation showed a 1.9 fold increased risk of breast cancer limited to women in the highest quintile of methylation (OR =1.89 (1.36-2.64), p= 1.64x10^{-4}). There was no evidence of LINE1 methylation associated with cancer risk in the prospective cohort studies. The meta-analysis
of LINE1 and other global methylation markers showed little evidence of association with cancer risk for surrogate assays of repetitive elements, but relatively consistent association with cancer risk using HPLC based total methyl-cytosine levels. Investigation of potential modifiers of ATM methylation revealed that methylation was independent of genetic haplotype, but independently associated with age, genotype of the one-carbon metabolism enzyme MTHFR, and serum levels of serum kynurenic acid levels in controls (p=0.02). Surprisingly, ATM methylation was reduced in controls (p= 5.707e-06) and cases (p= 0.008) that had fasted compared to those that had not. The effect of fasting on ATM methylation could be recapitulated by glucose restriction in ex-vivo PBMCs (p=0.046), independent of cell proliferation. Discovery studies to identify novel DNA methylation risk markers were conducted using differential methylation hybridisation and Illumina Infinium HumanMethylation450 BeadChip microarrays; however, significant associations were not reproducible in validation sample sets. Discussed are prospects and caveats for epigenetics association studies, including the implications of temporal alteration of DNA methylation by environmental exposures, biases associated with genetic influences on DNA methylation, and the potential for investigation of these interactions to better understand the contribution of epigenetics to gene expression and cancer risk.
CHAPTER 4 ........................................................................................................................ 107
Investigation of Factors Influencing ATM Methylation ....................................................... 107
4.1 Introduction .................................................................................................................. 108
4.2 Materials and Methods ............................................................................................... 110
   4.2.1 Subjects ............................................................................................................... 110
   4.2.2 Laboratory methods ............................................................................................ 114
   4.2.3 Statistical analysis .............................................................................................. 119
4.3 Results ....................................................................................................................... 122
   4.3.1 ATM Haplotype .................................................................................................. 122
   4.3.2 MTHFR c77t genotype ...................................................................................... 124
   4.3.5 Fasting Status .................................................................................................... 132
   4.3.6 Age .................................................................................................................... 136
   4.3.7 Ex-vivo Glucose Restriction ............................................................................. 139
   4.3.8 Genome-wide Analysis of Effect of Fasting on DNA Methylation ...................... 144
     4.3.8.2 Analysis Method 2 ....................................................................................... 149
4.4 Discussion .................................................................................................................. 157

CHAPTER 5 ........................................................................................................................ 167
Microarray-based discovery studies to identify novel breast cancer risk markers ............. 167
5.1 Introduction ................................................................................................................ 168
5.2 Materials and methods ............................................................................................... 172
   5.2.1 Discovery study 1 - Differential Methylation Hybridisation (DMH) .................... 172
   5.2.2 Discovery study 2 (Illumina 450k beadchip array) ............................................ 180
5.3 Results ....................................................................................................................... 183
   5.3.1 Differential Methylation Hybridization ............................................................. 183
Statistical tests ............................................................................................................... 270
plotting ........................................................................................................................... 273
Meta-analyses ............................................................................................................... 273
450k array analysis ....................................................................................................... 274
Table of Figures

**Figure 2.1.** ATMv2a methylation distribution in breast cancer cases and controls 55

**Figure 2.2.** Effect estimate of the logistic model of splined ATMv2a methylation as it relates to case-control status 59

**Figure 2.3.** Association between ATMv2a methylation levels and breast cancer risk in combined analysis stratified by age 63

**Figure 2.4.** ATMv2a DNA methylation in blood, measured by pyrosequencing plotted against age at blood draw 64

**Figure 2.5.** Difference in ATMv2a and LINE1 DNA methylation between time-points in serially collected blood samples 69

**Figure 2.6.** Meta-analysis of ATM methylation in four case-control studies 71

**Figure 2.7.** Single nucleotide resolution methylation of the entire ATM gene in PBMCs 77

**Figure 3.1.** Meta Analysis of Studies investigating Genome-wide DNA methylation in peripheral blood DNA for cancer risk (studies published between 01/01/2000 and 08/06/2012) 92

**Figure 3.2.** Funnel plot for potential publication bias in LINE1 reports 94

**Figure 3.3.** Meta Analysis of Studies investigating LINE1 DNA methylation in peripheral blood DNA for cancer risk 96

**Figure 3.4.** Update meta Analysis of Studies investigating Genome-wide DNA methylation in peripheral blood DNA for cancer risk (studies published between 08/06/2012 and 05/09/2013) 98

**Figure 4.1.** ATMv2a DNA methylation stratified by ATM haplotype 122

**Figure 4.2.** ATMv2a DNA methylation stratified by MTHFRc677t genotype 124

**Figure 4.3.** ATMv2a DNA methylation measured by pyrosequencing, stratified by quartiles of kynurenic acid (KA) 128

**Figure 4.4.** DNA methylation analysis (pyrosequencing) of (A) ATMv2a and (B) ATMv2b in blood cell fractions 130
Figure 4.5. Association of ATMmvp2a DNA methylation with Fasting Status.

Figure 4.6. Meta-analysis of association between ATMmvp2a methylation and fasting status stratified by case-control status

Figure 4.7. ATMmvp2a DNA methylation in blood, measured by pyrosequencing plotted against age at blood draw

Figure 4.8. LINE1 and ATMmvp2b methylation stratified by fasting status

Figure 4.9. Reduction of ATM methylation upon glucose restriction

Figure 4.10. Cell proliferation in PBMCs in response to glucose restriction

Figure 4.11. ATMmvp2a methylation analysis in glucose restricted cell lines

Figure 4.12. Top 8378 loci significantly associated with fasting status

Figure 4.13. Discovery of additional candidate fasting-association loci in Illumina 450k array data

Figure 4.14. Top 106 loci significantly associated with fasting status in analysis 2

Figure 4.15. Selecting of loci significantly associated with fasting

Figure 4.16. Strip-charts showing DNA methylation distributions

Figure 5.1. Genomic loci of PCR amplicons and pyrosequencing assays for validation of PARP1

Figure 5.2. Heat-map of Differential methylation Hybridisation (DMH) microarray top hits

Figure 5.3. Methylation variability at all genes analysed by DMH

Figure 5.4. Validation of PARP1 hypermethylation by bisulphite pyrosequencing

Figure 5.5. Bisulphite pyrosequencing of three regions of PARP1

Figure 5.6. Association of PARP1 DNA methylation with time to diagnosis

Figure 5.7. Manhattan plot showing statistical significance of associations between DNA
methyltransferase and breast cancer risk

**Figure 5.8.** Significant blood hypermethylation of neighbouring CpG sites in TFDP3

**Figure 5.9.** Association of loci within SLFN12L and HEATR2 with breast cancer case-control status

**Figure 6.1.** Trimodal DNA methylation of CpG sites within LGR6
List of Tables

Table 2.1. Characteristics of study populations 46

Table 2.2. ATM methylation primers 51

Table 2.3. Distribution of methylation levels in ATMmvp2a, ATMmvp2b and LINE1 in breast cancer cases and controls 54

Table 2.4: Association between methylation levels in ATMmvp2a and breast cancer risk in BGS, EPIC, KConFab and Combined analysis 57

Table 2.5. Association between methylation levels in ATMmvp2a and breast cancer risk in combined analysis stratified by age at blood draw 61

Table 2.6. Association between methylation levels in ATMmvp2a and breast cancer risk in combined analysis stratified by menopausal status at blood drawn 65

Table 2.7. Association between ATMmvp2a methylation levels and breast cancer risk by time from blood collection to diagnosis in combined analysis 67

Table 2.8. Association between methylation levels at the ATM locus and breast cancer risk in the British Bilateral Breast Cancer Study (BBCS) (Flanagan et al., 2009b) 72

Table 4.1 Characteristics of EPIC study subgroups used in chapter 4 112

Table 4.2 Primers designed for research within chapter 4 116

Table.4.3 General linear models for investigation of association between ATM methylation and serum metabolites in test and validation sets 126

Table 4.4. Mean and standard deviation of ATMmvp2a methylation in PBMCs of 6 participants (A-F) after culture for 24 hours in media of different glucose concentrations 142

Table 4.5. 450k array loci associated with fasting in analysis 1 and 2 149

Table 5.1. KConFab study samples used for investigation of PARP1 DNA methylation 171

Table 5.2 Primers for chapter 5 176

Table 5.3. Discovery and validation sample sets for EWAS study using the 450k beadchip array 179
Table 5.4. Ten DMH fragments most significantly associated with breast cancer case-control status

Table 5.5. Assessment of association of blood PARP1 DNA methylation with tumour pathological features in the KConFab study

Table 5.6. DNA methylation of controls and cases for HEATR2 and SLFN12L

Appendix table 1. Studies included in meta-analyses of genome-wide DNA methylation and cancer risk (chapter 3)

Appendix table 2. Assessment of association with case-control status of 450k array loci (EPIC4) within the regions identified as significantly differentially methylated in the differential methylation hybridisation array discovery study (chapter 5)
**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5meC</td>
<td>5-methylcytosine</td>
</tr>
<tr>
<td>5hmC</td>
<td>5-hydroxymethylation</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>Bp</td>
<td>base-pairs</td>
</tr>
<tr>
<td>DDR</td>
<td>DNA damage repair</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association study</td>
</tr>
<tr>
<td>EWAS</td>
<td>Epigenome-wide association study</td>
</tr>
<tr>
<td>PBMC</td>
<td>Peripheral blood mononuclear cell</td>
</tr>
<tr>
<td>MVP</td>
<td>Methylation variable position</td>
</tr>
<tr>
<td>DMR</td>
<td>Differentially methylated region</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcriptional start site</td>
</tr>
<tr>
<td>TTD</td>
<td>Time to diagnosis (from blood-draw)</td>
</tr>
<tr>
<td>DCIS</td>
<td>Ductal carcinoma <em>in situ</em></td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>OR</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>RR</td>
<td>Relative risk</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>BMI</td>
<td>Body mass index</td>
</tr>
<tr>
<td>CRA</td>
<td>Colorectal adenoma</td>
</tr>
<tr>
<td>CIN</td>
<td>Cervical intraepithelial neoplasia</td>
</tr>
<tr>
<td>HPLC</td>
<td>High performance liquid Chromatography</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------</td>
</tr>
<tr>
<td>OCM</td>
<td>One carbon metabolism</td>
</tr>
<tr>
<td>KA</td>
<td>Kynurenic acid</td>
</tr>
<tr>
<td>TE</td>
<td>Transposable element</td>
</tr>
</tbody>
</table>
Publications from Thesis:

ORIGINAL PAPERS:

1: Kevin Brennan\(^1\), Kirsty Flower\(^1\), Charlotte Wilhelm-Benartzi\(^1\), Karin van Veldhoven\(^2,3\), Silvia Polidoro\(^3\), Carlotta Sacerdote\(^3\), Alessio Nacarrati\(^3\), Fulvio Ricceri\(^3\), Heather Thorne on behalf of KConFab Investigators\(^4\), Elio Riboli\(^5\), María-José Sánchez\(^6\), Francoise Clavel-Chapelon\(^7\), Salvatore Panico\(^8\), H. Bas Bueno-de-Mesquita\(^9,2\), Dimitrios Trichopoulos\(^10\), Rudolf Kaaks\(^11\), Kay-Tee Khaw\(^12\), Paolo Vineis\(^2,3\), Robert Brown\(^1,13\) and James M. Flanagan*Metabolic dietary factors associated with peripheral blood based DNA methylation in the ATM gene: implications for Epigenome-wide association studies, submitted, under review.


biomarker of breast cancer risk. Cancer Res. 2012 May 1;72(9):2304-13. doi:

Acknowledgments

I wish to thank Dr. James M. Flanagan for his constant guidance and encouragement throughout my PhD, for giving me independence and freedom, and for pushing me towards completing this piece of work. To Professor Robert Brown for his advice, support and inspiration, and for building a wonderful environment in which to work. To everyone within the epigenetics unit for camaraderie and team-spirit, for sharing thoughts and idea with me, and for helping me to prepare for and celebrate all the milestones along the way. I couldn’t have wished for a more dynamic and likeable group of people to work with. Special Thanks to Dr. Kirsty Flower, Dr. Charlotte Wilhelm-Benartzi, Dr. Wei Dai, Dr. Natalie Morris, and Ms. Michelle Metcalf for training and collaboration, and to Dr. Jane Borley for solidarity during thesis writing. To teachers and mentors past, through whose guidance I found biology and epigenetics. I wish to thank Professor Paolo Vineis and his group, and Professor Montserrat Garcia-Closas for invaluable advice and collaboration. To my family, for encouragement and reassurance over the last few years. Special thanks to my parents for their endless generosity and support, and for encouraging me to pursue my interest in science and the natural world from an early age. To those wonderful friends, on whose generosity I have been far too reliant, for always being there to listen, console and entertain, and to clothe, feed and house me during desperate times; you know who you are. I would like to thank Breast Cancer Campaign for providing funding and inspiration for my research. Most importantly, I wish to the thousands of women who have generously participated in this research by providing samples.
I would like to thank the following people for contributing to this research project as follows:

To Dr. James M. Flanagan, Professor Montserrat Garcia-Closas and Dr Charlotte Wilhelm-Benartzi for conducting statistical analyses reported in chapter 2, and for our published paper reporting these results (Brennan et al., 2012b) (chapter 2).

To Dr. James Flanagan et al (Flanagan et al., 2009b) for design of the pyrosequencing assays used for methylation analysis of the ATM gene (chapter 2).

To Dr. Silvia Polidoro and Dr. Alessio Naccarati for preparation of samples from EPIC-Turin and running of Illumina 450k arrays (chapters 4 and 5).

To Mr. Gianluca Campanella, Professor Paolo Vineis et al for provision of the Ch3 data analysis package used for processing of Illumina 450k array data (chapters 4 and 5).

To Dr. James M. Flanagan and Ms. Karin van Veldhoven: For conducting quality control and data pre-processing of all Illumina 450k array data (chapters 4 and 5).

To Francoise Clavel-Chapelon, Professor Salvatore Panico, and Dr. N. Charlotte Onland-Moret, Professor. Rudolph Kaaks, Professor Elio Riboli, Miren Dorronsoro, Professor Dimitrios Trichopoulos, Professor Kay-Tee Khaw, Dr Carlotta Sacerdote and Fulvio Ricceri, for provision of prepared blood DNA samples from the EPIC cohort centres.

Prof Montse Garcia-Closas, Dr. Olivia Fletcher, Dr. Nick Orr, Prof Michael Jones , Professor Alan Ashworth, Professor Anthony Swerdlow for providing prepared blood DNA samples from the Breakthrough Generations Study.

To Ms. Heather Thorne and the KConFab investigators for provision of extracted blood DNA samples from the KConFab study (chapters 2, 4 and 5).
To Dr. Jane Borley and Mr. Jay Chatterjee for collection of blood samples for \textit{ex vivo} glucose restriction (chapter 4)

To Dr. Sadaf Ghaem-Maghami for provision of materials for blood collection and extraction of PBMCs for the \textit{ex vivo} glucose restriction cell culture experiments (chapter 4).

To Dr. Kirsty Flower: For extracting PBMCs from whole blood and performing \textit{ex vivo} glucose restriction cell culture (chapter 4).

To Dr. Charlotte Wilhelm-Benartzi for assistance with statistical analyses carried out in chapter 4.
Declaration of Originality

I Kevin Brennan declare that this thesis is solely my own work. All references have been consulted by me in the preparation of this manuscript. Except where stated, the work presented in this thesis was performed by me.

Copyright Declaration

The copyright if this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives license. Researchers are free to copy distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work.
Chapter 1

GENERAL INTRODUCTION
1.1 Breast Cancer

Breast cancer is the most common malignancy affecting women worldwide with an average lifetime risk of 12% (Amir et al., 2010; McPherson et al., 2000). Incidence is highest in western countries and is increasing worldwide, especially in developing countries, and is predicted to increase in the coming decades (Cuzick et al., 2011). However, mortality rates have decreased substantially in recent years, due to earlier diagnosis and improved treatment (Collins and Politopoulos, 2011; Colston, 2008; Cuzick et al., 2011; Umar et al., 2012). Most breast cancers are initially responsive to treatment, however 20-30% of breast cancer patients eventually develop incurable metastatic diseases (Eroles et al., 2012).

Breast cancers are mainly classified based on expression of the estrogen receptor (ER), as ER+ and ER- breast cancers are considered distinct diseases with different etiologies, gene expression signatures and prognosis (Keller et al., 2012), and may have different risk factors (Tamimi et al., 2012). Further treatment stratification is based on expression of the progesterone receptor (PR) and amplification of the epidermal growth factor receptor 2 (HER2) gene (Dawson et al., 2013). Breast cancers may also be classified by histological features, mutation profiles and mRNA expression profiles, and a recent classification based on genomic and gene expression data identified ten reproducible subtypes with distinct pathological features (Curtis et al., 2012).

Ductal carcinomas make up ~80% of breast cancers (Keller et al., 2012), and the majority are thought to derive from luminal epithelial cells, though rare breast cancer subtypes may derive from basal cells (Keller et al., 2012).

Common features of breast cancer include impaired DNA damage repair (DDR) resulting in genomic instability, mutations and chromosome rearrangements (Eroles et al., 2012; O'Donovan and Livingston, 2010), angiogenesis (Vona-Davis and Rose, 2009), and up-
regulation of signalling pathways that promote cellular growth and proliferation, including the RAS/RAF/MAPK, JAK/STAT or PI3K/AKT/mTOR pathways (Eroles et al., 2012).

1.1.1 Breast cancer risk

Breast cancer mortality and disease burden are reduced with diagnosis at earlier disease stage (Nickson et al., 2012). Frequent breast cancer screening with magnetic resonance imaging (MRI) is offered to women at known high risk including women carrying mutations in the BRCA1 and BRCA2 breast cancer susceptibility genes, and women with a personal history or strong family history of cancer (Pashayan and Pharoah, 2012). Less sensitive mammographic screening is offered less frequently to women over 50 who are not known to be at high risk (Autier et al., 2010). Whereas Individual studies and meta-analysis reveal that mammographic breast screening is associated with decreased mortality from breast cancer (2012; Nickson et al., 2012), the current screening approach is limited by our current inability to accurately predict breast cancer risk.

Breast cancer is considered one of the most preventable cancers, and a significant proportion of breast cancers could be prevented though weight loss, physical exercise and chemoprevention (Colditz and Wei, 2012), therefore, lifestyle advice and preventative therapy may be offered to women at known high risk (Colditz and Wei, 2012; Cuzick et al., 2011; Umar et al., 2012). Preventative chemotherapy may benefit women at highest risk, as epidemiological studies and assessment of breast cancer relapse prevention has identified several potential chemopreventative agents including anti-estrogens such as Tamoxifen and Raloxifene, and Aromatase inhibitors for ER+ disease, and bisphonates, metformin, statins and anti-inflammatory drugs such as aspirin for all breast cancer prevention (Cuzick et al., 2011). Conversely, as regular mammographic screening is thought to also result in over-
Several breast cancer risk prediction models have been developed to predict breast cancer incidence (reviewed in (Evans and Howell, 2007) and (Amir et al., 2010). The current models predict cancer risk based on age, family history, reproductive factors and genetic data. Such models are useful for predicting likelihood of having a breast cancer susceptibility mutation based on family history, and are used for selection of candidates for genetic screening for BRCA1/BRCA2 genetic screening (Parmigiani et al., 2007). The most accurate risk models, such as the Cuzick-Tyrer model, incorporate multiple risk factors, including genetics and non-genetic factors such as parity, and may be improved by incorporating additional genetic loci and breast density (Amir et al., 2010). However, many non-genetic factors such as body mass index, alcohol consumption and circulating hormone levels are not included. Furthermore, the current models have very limited ability to predict individual risk in the absence of high penetrance mutations or strong family history of cancer (Parmigiani et al., 2007). Given that most breast cancers are sporadic, novel breast cancer risk markers are required to improve risk model performance for the general population (Cuzick et al., 2011; Prado et al., 2010).

1.1.2 Breast cancer genetics

Breast cancer displays familial clustering, and having relatives with early onset breast cancer is a strong risk factor for disease. Ethnicity also affects breast cancer risk in a subtype-specific manner, independent of geographic location and socioeconomic factors (Carey et al., 2006). These factors are indicative of a strong genetic component to disease risk (Mavaddat et al., 2010). However, only 5-10% of breast cancers are associated with high-
penetrance mutations in one of the known BRCA1 and BRCA2 breast cancer susceptibility genes (Armes et al., 1997; Collins and Politopoulos, 2011; Stadler et al., 2010), and many families displaying familial clustering of breast cancer lack known susceptibility mutations. Furthermore, linkage studies suggest that no more high-penetrance breast cancer susceptibility genes are likely to exist (Smith et al., 2006). It has been estimated that 20-25% of familial breast cancer risk is accounted for by mutations in breast cancer susceptibility genes (Mavaddat et al., 2010). These include mutations in TP53, PTEN and STK11, associated with the familial breast cancer susceptibility syndromes Li Fraumeni syndrome, Cowden syndrome and Peutz-Jegher syndrome respectively (Mavaddat et al., 2010). Germline mutations in E-cadherin (CDH1) are associated with susceptibility to breast, gastric and colorectal cancers (Pharoah et al., 2001). Moderate-penetrance mutations occur in the CHEK2, ATM, BRIP1 and PALB2 genes, and are associated with roughly 2-fold greater breast cancer incidence (Easton et al., 2007; van der Groep et al., 2011).

Whereas mutations in known susceptibility genes are important risk factors for affected families, they are rare, explain only a minor proportion of breast cancer heritability, and offer little risk prediction value for the general population (Mavaddat et al., 2010; Salhab et al., 2010). Genome-wide association studies (GWAS) studies have attempted to identify additional loci associated with breast cancer risk. ~68 low penetrance (OR 1.1-1.3) significant loci have been found to date, bringing up to ~30%, the percentage of breast cancer heritability explained by all genetic loci identified so far (Michailidou et al., 2013; Southey et al., 2013).

GWAS studies support the theory that common complex diseases such as sporadic breast cancer are ‘polygenic’ in nature; are caused by multiple common risk factors rather than single rare lesions (Pharoah et al., 2002). The low penetrance of GWAS SNPs limits their suitability as biomarkers (Manolio et al., 2009; Politopoulos et al., 2011), however, combining
multiple SNPs into risk prediction models may help to improve model performance (Dite et al., 2013). This also presents a paradigm for including other information, such as epigenetic markers, into these prediction models.

1.1.3 Other breast cancer risk factors

Many well-established breast cancer risk factors, in addition to known genetic loci, exist. Age is the strongest breast cancer risk factor, with doubling of risk with every decade until menopause, after which risk increases more gradually with age (McPherson et al., 2000). Age may contribute to cancer risk through increased rate or accumulation of somatic mutations, or through association with other risk factors such as menopausal status and obesity. Having a prior personal history of benign breast disease is associated with increased breast cancer risk (Hartmann et al., 2005). Anthropometric measures such as size at birth (Lof et al., 2007), stature (Kabat et al., 2013), body mass index (Cheraghi et al., 2012) and breast density (Boyd et al., 2007) are all associated with breast cancer risk. While, anthropometric factors have both genetic and environmental underpinnings, the role of these factors in cancer risk is poorly understood (Kabat et al., 2013).

Levels of circulating sex hormones are associated with increased breast cancer risk in postmenopausal (Key et al., 2003) and premenopausal women (2013). Reproductive factors which modulate endogenous estrogen levels, including parity, age at menarche, age at first birth, age at menopause, and breast-feeding (He et al., 2012; Li et al., 2003), all affect breast cancer risk. Exposures to exogenous estrogens (Li et al., 2003), including oral contraceptives, hormone replacement therapy, and in utero exposure to synthetic estrogens also increase breast cancer risk (Parkin, 2011; Yager and Davidson, 2006).
Other lifestyle-associated breast cancer risk factors include alcohol consumption (Key et al., 2006), physical exercise (Lynch et al., 2011), smoking (Luo et al., 2011), night-shift work (Megdal et al., 2005) and socioeconomic status (Webster et al., 2008). Many lifestyle risk factors for breast cancer relate to diet and metabolism (Thomson, 2012). Levels of glucose (Sieri et al., 2012), folate (Maruti et al., 2009) and vitamin D (Colston, 2008) have been reported to be associated with breast cancer risk.

Lastly, interactions between genetic and environmental breast cancer risk factors may be frequent, as a recent consortium effort has successfully identified two examples, including an interaction between \textit{LSP1}-rs3817198 and parity and \textit{CASP8}-rs17468277 and alcohol consumption (Nickels et al., 2013), and risk associated with dietary factors may be modified by genetic variability in genes encoding enzymes in metabolism (Kaklamani et al., 2011; Liu et al., 2013).

Less well studied to date is the possibility that epigenetic factors contribute to cancer susceptibility, whether as an innate epigenetic variability between individuals, akin to SNPs, or as a mediator of other genetic, environmental or lifestyle factors.

\subsection{1.2 Epigenetics}

Epigenetics refers to the cellular mechanisms that control gene expression and genome function independent of alterations in genome sequence (Feinberg and Tycko, 2004; Richards, 2006), and accounts for the phenotypic variability between genetically identical cell types that is introduced through cellular differentiation (Goldberg et al., 2007).

Within the nucleus of mammalian cells, DNA is wrapped around octamers of histone proteins to form nucleosomes, the building blocks of chromatin. Gene expression is regulated by
chromatin structure, where open euchromatin is associated with gene transcription and access to gene regulatory elements of transcription factors and DNA binding factors, whereas dense heterochromatin is associated with gene repression (Hon et al., 2009). Chromatin structure and gene expression are regulated in part through changes in nucleosome occupancy (Struhl and Segal, 2013), incorporation of histone variants (Ooi et al., 2009), and post-translational modifications of N-terminal histone tails, including histone acetylation, methylation, phosphorylation, ubiquitylation, sumoylation and ribosylation (Kouzarides, 2007). Complex chromatin signatures are associated with gene expression across the genome, for example, active gene promoters are characterised by enrichment of histone 3 lysine 4 trimethylation (H3K4me3) whereas histone 3 lysine 27 trimethylation (H3K27me3) is enriched at inactive gene promoters (Hon et al., 2009). Histone modifications are catalysed by enzymes such as histone methyltransferases (HMTs) and histone acetyltransferases (HATs), and removed by enzymes such as histone demethylase (HDMs) and histone-deacetylases (HDACs).

MicroRNAs (miRNA) are short RNA molecules that bind to messenger RNA molecules and regulate their stability and translation, and are an example of post-transcriptional epigenetic regulation of gene expression. (Baylin and Jones, 2011; You and Jones, 2012). Many other types of non-coding RNA regulate gene expression, including long intergenic non-coding RNAs (lincRNAs), that are implicated in regulation of chromatin structure, PIWI-interacting RNAs (piRNAs) that regulate transposable element transcripts, and small nucleolar RNAs (snoRNAs) that regulate ribosomal RNA function (Esteller, 2011).
1.2.1 DNA methylation

DNA methylation is the addition of a methyl group to cytosine nucleotides at the 5-carbon position of the cytosine pyrimidine ring to form 5-methylcytosine (5meC), occurring primarily at cytosines within CpG dinucleotides in mammals (Chiacchiera et al., 2013).

Among epigenetic marks, DNA methylation is the most frequently studied and well understood, largely due to its stability and the ability to measure methylation quantitatively and at high resolution using PCR-based techniques, enabling methylation analysis of stored DNA and tissue samples (Rakyan et al., 2011; Robertson, 2005). Whereas DNA methylation is frequently studied in isolation, DNA methylation and chromatin modifications function cooperatively to regulate transcriptional regulation (Cameron et al., 1999; Jones, 2012; Nan et al., 1998), and emerging evidence suggests that DNA methylation patterns are determined by pre-existing chromatin marks (Ooi et al., 2009; Ooi et al., 2007).

CpG dinucleotides occur at just 25% of their expected frequency within the human genome, due to high propensity for spontaneous deamination to thymine of methylated CpG site (Deaton and Bird, 2011; You and Jones, 2012). CpG sites are depleted at intergenic regions, and concentrated at CpG islands, regions of high CpG density found at the 5’-regions of ~60% of human genes (Deaton and Bird, 2011). CpG islands are usually protected from DNA methylation by an unknown mechanism, therefore, only ~2% of human DNA methylation occurs at CpG islands compared with 34% found at gene bodies (Maunakea et al., 2010).
1.2.1.1 Establishment and loss of DNA methylation

DNA methylation is catalysed by the DNA methyltransferase (DNMT) enzymes, of which there are five in mammals (Jurkowska et al., 2011). DNMT3a and DNMT3b, the ‘de novo’ DNMTs, have high affinity for unmethylated DNA, and are responsible for establishing the first patterns of DNA methylation during early development (Jurkowska et al., 2011). DNMT1 has highest affinity for hemi-methylated (methylated on one DNA strand) DNA, and is described as the ‘maintenance’ DNA methyltransferase responsible for methylation of newly synthesised daughter DNA strands following DNA replication (Song et al., 2012). However, there is overlap between the roles of DNMT1 and de novo DNA methyltransferases (Ooi et al., 2009). The function of DNMT2 is poorly understood, however it appears to methylate tRNA rather than DNA (Goll et al., 2006). DNMT3L is required for regulation of de novo methylation of maternally imprinted regions during oocyte development (Bourc'his et al., 2001).

DNA methylation is erased and re-established within the mammalian germline, and within the preimplantation embryo, allowing epigenetic reprogramming within each generation (Chen et al., 2003). There is some evidence suggesting that some epigenetic patterns can escape erasure during reprogramming and undergo transgenerational inheritance (de Assis et al., 2012; Hitchins et al., 2007a), however, this idea remains controversial (Chong et al., 2007; Daxinger and Whitelaw, 2012).

Relatively little is known about the mechanism by which DNA methylation is removed. Demethylation may simply occur passively, through the failure or prevention of maintenance DNA methylation during mitosis. Active demethylation, independent of cell division, occurs rapidly across the entire paternal genome paternal genome during embryonic preimplantation, and is thought to occur though oxidation of 5meC to 5-
hydroxymethylcytosine (5hmC) by the ten-eleven-translocase (TET) enzymes followed by a base-excision repair mechanism (Bhutani et al., 2011; Chiacchiera et al., 2013).

1.2.1.2 Function of DNA methylation

The complete function of DNA methylation across the genome remains unclear, with different functions ascribed to different genomic contexts. DNA methylation at gene promoters is well established as having a transcriptionally repressive role (Appanah et al., 2007; Jones, 2012). DNA methylation is indispensable for genomic imprinting, the phenomenon by which gene expression occurs on either the paternal or maternal allele, whilst the other allele is epigenetically silenced (Ferguson-Smith, 2011). Silencing of the inactive X chromosome in female mammals is associated with antisense RNA-induced DNA hypermethylation and heterochromatin formation (Wutz, 2011). Promoter hypermethylation plays a role in cellular differentiation, and is required for silencing of developmentally expressed genes in adult tissues, and genes that display tissue-specific expression (Pujadas and Feinberg, 2012). Curiously, intragenic DNA methylation appears to be associated with increased transcriptional activity, especially at intragenic CpG islands (Deaton et al., 2011; Maunakea et al., 2010). Recent research has implicated DNA methylation in alternative splicing, as DNA methylation is enriched at genes that undergo alternative splicing (Flores et al., 2012; Maunakea et al., 2013; Wan et al., 2013). Lastly, transposable elements are usually heavily methylated, consistent with an evolutionary role of DNA methylation in preventing transcription of endogenous parasitic DNA and maintaining genomic stability (Cordaux and Batzer, 2009; Ito et al., 2011; Munoz-Lopez et al., 2011).

Also elusive is the mechanism by which DNA methylation induces transcriptional change. Methylation of some transcription factor binding sites, such as the MYC binding site, directly
prevents binding (Jones, 2012; Prendergast and Ziff, 1991). Recruitment of methyl-CpG binding domain proteins MBD1, MBD2, MBD4 and MECP2 may also contribute to transcriptional silencing (Bogdanovic and Veenstra, 2009; Deaton and Bird, 2011). Whereas it is unclear whether DNA methylation is the causative mechanism of transcriptional silencing of hypermethylated transcription start sites, methylation may stabilise silencing induced by formation of a repressive chromatin state (Jones, 2012). DNA methylation patterns vary greatly between tissues (Christensen et al., 2009; Koestler et al., 2012; Meissner et al., 2008). Tissue-specific differential methylation is enriched at gene bodies, and within development genes and transcription factors (Eckhardt et al., 2006; Flanagan et al., 2009b; Illingworth et al., 2008; Kaminsky et al., 2009; Rakyan et al., 2008; Reinius et al., 2012).

1.2.1.3 Factors affecting DNA methylation variability

Considerable research efforts have been made to classify the factors that define DNA methylation patterns that may contribute to phenotype and health outcomes. Genetic factors play a strong role, as the epigenomes of monozygotic twin-pairs are more similar than those of dizygotic twin-pairs (Gordon et al., 2012; Kaminsky et al., 2009). In vitro studies have further helped to classify the genetic motifs that define DNA methylation pattern (Lienert et al., 2011). Twin studies also demonstrate the profound influence of age and environmental exposures on the epigenome, as epigenetic signatures within monozygotic twin pairs diverge with age and exposure to different environments (Fraga et al., 2005; Gordon et al., 2012; Kaminsky et al., 2009).

Age-associated changes in DNA methylation include sporadic hypermethylation of tumour suppressor genes, sometimes associated with transcriptional silencing, and genome wide reduction of 5meC (genome-wide hypomethylation) associated with genomic instability and
increased mutation rate (Chen et al., 1998; Cho et al., 2010; Eden et al., 2003; Wilson et al., 2007). Age-associated DNA methylated loci are also enriched at bivalent chromatin domains, i.e. regions at which both active and repressive chromatin marks are co-exist, implying a novel mechanism for age-associated epigenetic dysregulation in tumorigenesis, as bivalent chromatin domains are frequently abnormally methylated in cancer (Rakyan et al., 2010). Therefore age-associated DNA methylation changes may contribute to ageing and etiology of age-related diseases (Calvanese et al., 2009; Chen et al., 1998; Gonzalo, 2010).

The epigenome is thought to be especially sensitive to alteration by environmental exposures and stressors during specific developmental windows such as the periconceptual period during which epigenetic programming occurs (Feinberg, 2007; Heijmans et al., 2009). The ‘foetal origin of human disease’ hypothesis postulates that stable alterations of epigenetic states during prenatal development, caused by environmental, endogenous, and stochastic factors, may contribute to risk of adult onset disease (Feinberg, 2004). Evidence for this ‘epigenetic progenitor origin’ of human disease includes the identification of abnormal DNA methylation patterns with periconceptual exposure to famine, detectable in blood of exposed adults several decades later (Tobi et al., 2009), and in vitro epigenetic alteration by treatment with diethylstilbestrol (Hsu et al., 2009), a teratogenic synthetic estrogen associated with increased risk of breast and reproductive organ cancers in prenatally exposed individuals (Reed and Fenton, 2013). Epidemiological and animal studies support a role of other in utero exposures, such as maternal diet, alcohol consumption and environmental toxin exposure (Burdge et al., 2009; Dolinoy et al., 2006; Fei et al., 2013; Haggarty, 2013; Hilakivi-Clarke and de Assis, 2006; Pilsner et al., 2009; Waterland et al., 2006; Xu and Du, 2010). DNA methylation patterns in adult tissues are also susceptible to
alteration by diet (Li et al., 2011; Milagro et al., 2011; Weisenberger et al., 2005; Zhang et al., 2011b) and smoking (Shenker et al., 2012; Shenker et al., 2013).

An emerging field of study, termed ‘epigenetic epidemiology’ explores the potential role of epigenetic mechanisms in mediating the effects of environmental exposures on disease risk. Associations of DNA methylation with exposure to environmental carcinogens such as lead (Wright et al., 2010), arsenic (Reichard et al., 2007) and benzene (Bollati et al., 2007) have been reported. Recently, Shenker et al identified an intergenic CpG site independently associated with both smoking and breast cancer risk (Shenker et al., 2012).

DNA methylation dysregulation plays an etiological role in several disorders, such as fragile-X syndrome, through hypermethylation-mediated silencing of the \textit{FMR1} gene (Dobrovic and Kristensen, 2009), and imprinting disorders, in which loss of allele-specific epigenetic regulation leads to gene-dosage alterations, resulting in pleiotropic phenotypes including increased susceptibility to several cancers (Robertson, 2005).

In summary, emerging research indicates that DNA methylation is susceptible to alteration by multiple developmental, innate and environmental cancer risk factors, suggesting that DNA methylation variability in normal tissues may represent an ‘archive’ of lifetime exposures that may be exploited for development of cancer risk biomarkers. Furthermore, the possibility that epigenetic regulation may partially mediate the carcinogenic effects of some cancer risk factors warrants further attention.

1.2.2 Epigenetics in tumours

Epigenetic dysregulation is a hallmark of cancer, and is thought to contribute to tumorigenesis and disease progression at least as much as genetic abnormalities (Baylin...
and Jones, 2011; You and Jones, 2012). Several epigenetic events, including loss of acetylation and methylation of histone 4, and loss and gain of expression of tumour suppressor miRNAs and oncogenic miRNAs, respectively, are characteristic of cancer cells. The gain or loss in cancer of conserved transcriptional enhancers represents a novel epigenetic mechanism of cancer initiation (Akhtar-Zaidi et al., 2012). The most widely studies epigenetic event in cancer is the widespread abnormal DNA methylation patterns displayed in virtually all cancers (Feinberg and Tycko, 2004).

Genes encoding epigenetic modifying enzymes are some of the most frequently mutated in cancers, often resulting in epigenetic dysregulation (You and Jones, 2012). For example, gain-of-functions mutations in Isocitrate dehydrogenase1 and 2, found in acute myeloid leukemia are associated with genome-wide and locus specific hypermethylation (Figueroa et al., 2010). Conversely, DNA methylation represents an endogenous mutagen, as methylated cytosines are prone to deamination to thymine, therefore a high proportion of SNPs, including 30% of disease associated germline SNPs, occur at CpG sites, and methylated gene bodies are highly prone to somatic mutation (Rideout et al., 1990; You and Jones, 2012).

DNA methylation variability is increased in cancers compared with their respective normal tissues, indicating a general loss of control of DNA methylation (Hansen et al., 2011). However, breast cancer associated methylation patterns display spatial clustering, indicating a non-random process (Novak et al., 2008). Furthermore, transcriptional inactivation by promoter hypermethylation of specific genes represents a key pathogenic mechanism analogous to somatic mutation. For example, hypermethylation of $BRCA1$ and $MLH1$ represents an alternative inactivating mechanism in tumours lacking somatic mutations in these genes, whereas other genes that are rarely mutated, such as $MGMT$, $CDKN2B$ and $RASFF1A$, are frequently silenced by promoter hypermethylation in many cancers (You and
Jones, 2012). Over 100 genes have been reported to be hypermethylated in breast cancer, including genes involved in DNA repair, cell cycle regulation, apoptosis, cellular invasion and cellular homeostasis (Jovanovic et al., 2010; Widschwendter et al., 2008). Abnormal DNA methylation patterns, including hypermethylation of some genes, including \textit{FOXC1} and \textit{RARB1}, is apparent in breast cancer precursor lesions and low grade disease (Fackler et al., 2003; Muggerud et al., 2010) indicating that epigenetic inactivation represents an early, and potentially causative lesion in tumorigenesis (Faryna et al., 2012; Feinberg et al., 2006; Tycko, 2003).

Promoter hypomethylation of some oncogenes, including \textit{CDH3} and \textit{NAT1}, occurs in breast cancer, potentially leading to spurious activation (Jovanovic et al., 2010). Interestingly, many loci are differentially methylated between breast cancers of different hormone receptor expression status (Li et al., 2010a).

Most solid tumour types, including breast cancer, display genome-wide hypomethylation compared with normal adjacent tissue and benign lesions (Soares et al., 1999). Genome-wide hypomethylation is an early tumorigenic event, detectable in precursor lesion and normal adjacent tissues associated with some cancers (Flatley et al., 2009; Suter et al., 2004). However, hypomethylation is correlated with breast cancer stage, suggesting that hypomethylation may also be a consequence of disease progression (Wild and Flanagan, 2010).

\subsection{1.2.3 DNA methylation and cancer risk}

Whereas abnormal epigenetic dysregulation in cancer and precursor lesions are well studied, little is known about the potential contribution to disease risk of epigenetic variability in normal tissue prior to disease. Several studies have identified DNA methylation marks in blood that are associated with cancer, suggesting systemic epigenetic dysregulation that
may precede disease and contribute to susceptibility, representing an epigenetic disease precursor state.

Blood is frequently used as a ‘surrogate tissue’ for investigation of epigenetic patterns in relation to cancer risk, as many cancer risk factors, such as age, genetics, and environmental exposures are known to affect blood DNA methylation (Milagro et al., 2011; Shenker et al., 2012), and as blood is relevant to cancer-related physiological factors such as inflammation and metabolism (Koestler et al., 2012; Xu et al., 2013a). Furthermore, blood is easily sampled, and the most widely collected tissue sample in population study cohorts. Lastly, blood DNA methylation would be ideal for biomarker development and population screening, due to ease of sampling and measurement.

The only blood epigenetic markers with a known causative role in disease susceptibility are rare constitutional (soma-wide) promoter hypermethylation events, known as ‘epimutations’ associated with silencing of the DNA mismatch repair genes \textit{MLH1} and \textit{MSH2} in Lynch syndrome, a familial colorectal cancer syndrome (Chan et al., 2006; Hesson et al., 2010). Underlying genetic mutations associated with the \textit{MSH2} epimutations, and some \textit{MLH1} epimutations, have been identified, explaining their heritable nature (Hitchins et al., 2011; Ligtenberg et al., 2012). No genetic lesions has yet been found to explain the non-mendelian heritability of some \textit{MLH1} epimutation cases (Hesson et al., 2010; Hitchins and Ward, 2009; Hitchins et al., 2007b), however \textit{MLH1} methylation is subject to germline erasure and re-establishment, at least within the male germline (Hitchins and Ward, 2007b), indicating that the DNA methylation pattern is not heritable \textit{per se} (Hitchins, 2013).

Genome-wide hypomethylation is by far the most widely studied epigenetic marker in blood in studies relating to cancer (see chapter 3). The experimental question being addressed, and the outcomes, vary widely between studies, depending on the assay used to measure
genome-wide DNA methylation, and various aspects of the study design (Brennan and Flanagan, 2012b; Nelson et al., 2011). Therefore, it remains unclear whether blood genome-wide hypomethylation contributes to cancer risk.

Four studies have reported the occurrence of subtle \textit{BRCA1} promoter hypermethylation in blood of breast cancer patients (Hansmann et al., 2012; Iwamoto et al., 2010; Snell et al., 2008; Wong et al., 2010). \textit{BRCA1} hypermethylation in blood was significantly more frequent in breast cancer patients displaying tumour \textit{BRCA1} hypermethylation, than in healthy controls in two studies (Iwamoto et al., 2010; Wong et al., 2010), perhaps suggesting that increased propensity for \textit{BRCA1} hypermethylation in normal tissue increases breast cancer risk. Another study, however, failed to confirm this finding (Kontorovich et al., 2009). A recent study (Hansmann et al., 2012) confirmed the presence of allele-specific \textit{BRCA1} promoter hypermethylation in a small proportion (0.015\%) of familial breast and ovarian cancer patients, suggesting that \textit{BRCA1} hypermethylation may represent a rare epimutation, however, given that only 10 healthy controls were included in the study (compared with 641 cases) it is not clear whether blood \textit{BRCA1} hypermethylation is associated with cancer risk. Whereas it appears that \textit{BRCA1} hypermethylation does occur in some normal cells in blood, prospective studies are required to further investigate the implications for cancer risk.

Previous research from our laboratory (Flanagan et al., 2009b) identified a region within the Ataxia telangiectasia mutated (\textit{ATM}) gene that displaying high DNA methylation variability (methylation variable position (MVP)), and that displayed significantly higher DNA methylation in blood of bilateral breast cancer patients (n=190) compared with healthy controls (n=190). As bilaterality is considered an indicator of constitutional breast cancer susceptibility (Weitzel et al., 2005), it was hypothesized that ATM hypermethylation may have contributed to disease risk prior to cancer incidence, however, a prospective
Investigation of DNA methylation at this locus was required to exclude the possibility of reverse causality.

Analogous to GWAS studies, the first epigenome-wide association studies (EWAS) have been conducted in order to identify epigenetic loci associated with various different cancers including bladder (Marsit et al., 2011), head and neck (Langevin et al., 2012) and nerve sheath tumours (Feber et al., 2011). A smaller study (Widschwendter et al., 2008), using a candidate gene approach, identified 5 loci that were differentially methylated between blood samples of breast cancer cases and controls. Methylation at the loci identified may reflect the presence of active disease, and may be suitable for diagnostic biomarker development (Teschendorff et al., 2009), or may be associated with cancer risk, confirmation of which would require prospective investigation.

EWAS studies will be greatly assisted by availability of reference ‘normal’ epigenotypes for peripheral blood mononuclear cells (PBMCs) and individual blood cell types, which have been, and are being generated by large international collaborations including the International Human Epigenome Consortium (IHEC), BLUEPRINT (http://www.blueprint-epigenome.eu/), ENCODE (http://genome.ucsc.edu/ENCODE/) and Road-map epigenenomics project (http://www.roadmapepigenomics.org/) (Langevin and Kelsey, 2013; Reinius et al., 2012). A full peripheral blood mononuclear cell (PBMC) DNA methylome from a healthy male individual is already publicly available (Li et al., 2010c).

1.2.4 Epigenetic Biomarkers in Clinical Use

There has been great interest in development of DNA methylation-based non-invasive biomarkers for various different cancer outcomes, including diagnosis (Warren et al., 2011), prognosis, survival, treatment response, and treatment toxicity, recurrence (Brock et al., 2008; Heyn and Esteller, 2012). The most well established biomarkers to date include
hypermethylation of MGMT in serum, which has been validated for use for diagnosis, prognosis-prediction and treatment stratification in glioblastoma (Esteller et al., 2000; Heyn and Esteller, 2012), and hypermethylation of GSTP1, detectable in serum and urine, which may be used for prostate cancer diagnosis (Esteller et al., 1998; Van Neste et al., 2012).

For several cancer types, tissues and bodily fluids that derive from the cancer affected tissue, and that may be easily and non-invasively collected, may provide the most relevant source of nucleic acids for development of cancer biomarkers. For example, DNA derived from buccal cells may be utilised for development of biomarkers for head and neck cancer, hair follicles may be useful for melanoma, sputum for lung cancers, urine for bladder cancer, semen and urine for prostate cancer, and stool samples may be for development of colorectal cancers biomarkers (Rakyan et al., 2011). For other cancers however, including breast, ovarian and pancreatic cancers, blood represents the most relevant, non–invasively collectable tissue for investigation (Hanash et al., 2011). Future studies of breast cancer biomarkers will likely benefit from availability of breast epithelial DNA samples, as biobanking of breast epithelium biopsies is increasing, and other non-invasive sources of breast epithelial cells such as nipple aspirate fluid and ductal lavage fluid, and breast milk are being developed (Ljung et al., 2004; Twelves et al., 2013). In the meantime, blood represents the most relevant tissue for breast cancer, for which large cohorts of prospectively collected tissue samples are available (Hanash et al., 2011).

1.3 Hypothesis and Aims

We hypothesised that breast cancer susceptibility may be influenced by systemic and gene-specific DNA methylation patterns that may be detectable in blood prior to breast cancer diagnosis. Such cancer-associated DNA methylation patterns may be determined by genetic
variability, age, environmental exposures, or other factors, therefore epigenetic patterns may mediate the effects of risk factors on disease susceptibility, likely through gene-expression changes. Blood DNA methylation may therefore provide molecular biomarkers for breast cancer risk prediction.

We aimed to investigate the relationship between blood DNA methylation and breast cancer risk in prospective studies, and to assess the suitability of blood DNA methylation for breast cancer risk biomarker development. Candidate risk markers of interest included the \( ATM \) and \( LINE1 \) markers previously associated with disease in retrospective studies. Next we aimed to investigate potential mechanisms defining methylation variability at the \( ATM \) gene locus. Lastly, we aimed to discover novel epigenetic breast cancer risk markers using microarray approaches.
CHAPTER 2

INTRAGENIC ATM METHYLATION AS A BREAST CANCER RISK FACTOR
2.1 Introduction

In a previous report (Flanagan et al., 2009b), Flanagan et al identified regions within the ATM gene that displayed increased DNA methylation variability in blood compared with other gene regions, which they termed methylation variable position (MVPs). Furthermore they showed, using bisulphite-pyrosequencing, that one of these MVPs, labelled ATMvmp2b, was significantly hypermethylated in blood samples of bilateral breast cancer cases (n=190), taken at diagnosis, compared with healthy controls (n=190). Furthermore, Inter-quartile analysis revealed that individuals with ATMvmp2b methylation within the highest methylation quartile were at three-fold increased breast cancer risk compared with women in the lowest quartile. ATMvmp2b is located within an intragenic repetitive element 4kb downstream of the ATM transcriptional start site (TSS), consistent with a significant enrichment of methylation variability at intragenic repetitive elements among 55 genes analysed by differential methylation hybridization in the same report. As bilaterality is proposed to be an indicator of constitutional, or systemic, breast cancer susceptibility (Weitzel et al., 2005), it was hypothesized that ATMvmp2b methylation may have represented a constitutional susceptibility factor for breast cancer, detectable prior to disease onset. However, prospective investigation of the association was required to exclude the possibility of reverse causality, or the possibility that hypermethylation represented an early tumorigenic event. (Marsit et al., 2011; Teschendorff et al., 2009).

The possibility that ATM DNA methylation is related to breast cancer risk is particularly interesting as ATM is a well known breast cancer susceptibility gene. Ataxia telangiectasia, the rare recessive disorder of pleiotropic phenotype caused by ATM insufficiency, is associated with strong susceptibility to several cancer types, and non-affected female mutation carriers are at increased breast cancer risk (Goldgar et al., 2011; Lavin, 2008; Lu et al., 2011). Furthermore, SNPs within ATM confer increased breast cancer risk, though their
penetrance is not well established (Lee et al., 2005a). ATM is a highly conserved serine/threonine kinase with several critical cellular functions, including sensing of double strand breaks and oxidative stress (Goldgar et al., 2011; Okuno et al., 2012; Shiloh, 2001) and regulation of cell cycle, DNA damage response (DDR) and cell survival (Shiloh, 2001; Shiloh, 2003). Furthermore, ATM is a regulator of important kinases including breast cancer tumour-suppressor genes BRCA1 and P53 (Shiloh, 2003).

Interestingly, Differential ATMmvp2b methylation was thought to be independent of overall changes in genome-wide DNA methylation, as DNA methylation of LINE1, a commonly used surrogate marker of genome-wide methylation, was not different between cases and controls. This was inconsistent with previous reports of LINE1 hypomethylation in blood of cancer patients compared with controls (Cash et al., 2011; Cho et al., 2010; Hsiung et al., 2007; Moore et al., 2008; Wilhelm et al., 2010).

In order to assess the relationship between ATM DNA methylation and breast cancer risk, we measured methylation of ATMmvp2b, and a neighbouring MVP termed ATMmvp2a, lying 381bp upstream of ATMmvp2b, in pre-diagnostic blood samples from three breast cancer case-control studies. LINE1 DNA methylation was also assessed in the same studies.
2.2 Materials and Methods

2.2.1 Study Populations

Study participants were drawn from three large studies with blood samples collected prior to breast cancer diagnosis (Table 2.1). All contributing studies have appropriate ethical approval for sample and data collection.

2.2.1.1 Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer (KConFab)

The study population has been described previously (Mann et al., 2006). Between 1997 and May 2011, KConFab collected peripheral blood samples from 12,240 members of 1,395 families (~8.8 samples per family) in Australia and New Zealand. Included families had on average three verified (5.4 unverified) breast cancers per family. Case status was confirmed by clinical pathology report, doctor’s notes, cancer council registry verification or death certificate. Incident cases of invasive breast cancer were selected for this study from all individuals with a breast cancer diagnosis >1 month after blood sample collection (n=171). Five cases of non-white ethnicity were excluded from the analyses resulting in 166 invasive cases that were compared to 225 healthy unrelated controls without a family history of breast cancer drawn from “best friends” of subjects enrolled in KConFab. The average time to diagnosis (TTD, the duration between blood sample collection and subsequent breast cancer diagnosis) for incident cases was 45 months (range 1 to 140 months). Information on breast cancer risk factors including hormonal and reproductive factors, cigarette smoking and alcohol drinking was collected from questionnaires at enrolment. Pathology data (histology, grade, nodal status, ER, PR, HER2 status) and BRCA1/2 mutation status was available for breast cancer cases. Genetic polymorphism data for 15 common breast cancer susceptibility SNPs (rs1011970, rs10995190, rs2380205, rs2981582, rs614367, rs704010,
rs11249433, rs13387042, rs2046210, rs4973768, rs6504950, rs999737, rs3817198, rs889312, rs13281615) were available for 12 cases and 190 controls.

### 2.2.1.2 Breakthrough generations study (BGS)

This study cohort has been described elsewhere (Swerdlow et al., 2011). BGS represents a large general population cohort consisting of ~110,000 women enrolled in the UK between 2003 and 2011. Participants were sampled from a nested case-control study of all incident cases of breast cancer diagnosed within the cohort before June 2010 and controls individually matched on recruitment source, year of enrolment, ethnicity (white only), date of birth within 12 months, availability of blood sample and the duration during which the blood sample was in the mail. Breast cancer cases were self-reported in a follow-up questionnaire completed approximately 2.5 years after enrolment, or notified by study participants by phone or letter. Self-reported diagnoses were confirmed through an electronic linkage with England/Wales/Scotland/Northern Ireland cancer registrations (or by the general practitioner for a small number of cases that could not be successfully linked). Checks against UK cancer registrations were made for participants known to have died by the time of the 2.5 year follow-up, and those who failed to provide follow-up (but had given permission for follow-up).

A random sample of 257 case-control pairs was selected for analysis. Excluded from this were 4 controls that were subsequently found to have invasive breast cancer, 3 cases whose blood was collected after diagnosis, and 1 case of ductal carcinoma in situ (DCIS) resulting in a final sample set of 253 cases and 253 controls. Average TTD for incident cases was 18 months (range 0.03-59) (Table 2.1). Information on breast cancer risk factors was collected from a baseline questionnaire at enrolment. Pathology information from was available for ~25% of cases.
A longitudinal study to evaluate DNA methylation stability was conducted using paired buffy coat samples collected at two time-points (enrolment (T0) and follow-up ~6 years later (T1)) from 92 healthy women from the BGS study. A sample and questionnaire were collected at T0 (2004), and again at T1. Inclusion criteria were: Age 35-84 years at enrolment, no history of breast cancer up to second blood collection, not known to have any relatives within the study, blood samples received at processing lab <1 day after collection, expected amount of blood received at the lab (three 10ml vials at baseline and two 10ml vials at follow up), no reported problems at collection or processing (e.g. lipemic, haemolysed, clotted samples), and duration between T0 and T1 between 5.5 and 6.5 years.

2.2.1.3 European Prospective Investigation into Cancer and Nutrition (EPIC)

The study population has been described elsewhere (Riboli and Kaaks). EPIC collected ~520,000 individuals with standardized lifestyle and personal history questionnaires, anthropometric data and blood samples collected for DNA extraction. The sample set used for this study included two subgroups, including a group of premenopausal women (127 cases and 145 controls) and post-menopausal women (121 cases and 149 controls), with menopausal status defined at the time of blood collection. Controls were individually matched on age at baseline, recruitment centre, date and time of blood collection. Precise ethnicity data was not available for these individuals, however ~80% of individuals were provided from study centres in Italy, Spain and the Netherlands and the remainder (20%) were provided from France, Germany, UK and Greece. DCIS cases were excluded from analyses (n=36). Blood samples from cases were taken on average 55 months before diagnosis (range 24-108) (Table 2.1). Extensive information on cancer risk factors, including extensive alcohol, smoking and dietary data, family history and hormonal factors were collected from a baseline questionnaire at enrolment. Pathology information includes morphology, grade, stage, ER, PR and HER2 status.
Table 2.1 Characteristics of study populations used in chapter 2

<table>
<thead>
<tr>
<th>Study</th>
<th>KConFab</th>
<th>BGS**</th>
<th>EPIC***</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-control</td>
<td>Controls</td>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td>N</td>
<td>225</td>
<td>166</td>
<td>253</td>
</tr>
<tr>
<td>Age (mean (range))</td>
<td>60 (33-83)</td>
<td>50 (21-91)</td>
<td>54 (23-82)</td>
</tr>
<tr>
<td>Family History, N (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0 (0%)</td>
<td>166 (100%)</td>
<td>51 (20%)</td>
</tr>
<tr>
<td>No</td>
<td>225 (100%)</td>
<td>0 (0%)</td>
<td>201 (80%)</td>
</tr>
<tr>
<td>Menopausal status, N (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-menopausal</td>
<td>37 (17%)</td>
<td>101 (65%)</td>
<td>135 (61%)</td>
</tr>
<tr>
<td>Post-menopausal</td>
<td>183 (83%)</td>
<td>54 (35%)</td>
<td>87 (39%)</td>
</tr>
<tr>
<td>Time to diagnosis, mean (range)</td>
<td>-</td>
<td>45 (1-140)</td>
<td>-</td>
</tr>
<tr>
<td>Age at diagnosis, mean (range)</td>
<td>-</td>
<td>52 (29-88)</td>
<td>-</td>
</tr>
<tr>
<td>Tumor Characteristics*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morphology - Ductal</td>
<td>-</td>
<td>n/a</td>
<td>-</td>
</tr>
<tr>
<td>Morphology - Lobular</td>
<td>-</td>
<td>n/a</td>
<td>-</td>
</tr>
<tr>
<td>Morphology - Other</td>
<td>-</td>
<td>n/a</td>
<td>-</td>
</tr>
<tr>
<td>Mutation - BRCA1</td>
<td>-</td>
<td>23%</td>
<td>-</td>
</tr>
<tr>
<td>Mutation - BRCA2</td>
<td>-</td>
<td>14%</td>
<td>-</td>
</tr>
<tr>
<td>Grade – I</td>
<td>-</td>
<td>12%</td>
<td>-</td>
</tr>
<tr>
<td>Grade – II</td>
<td>-</td>
<td>33%</td>
<td>-</td>
</tr>
<tr>
<td>Grade – III</td>
<td>-</td>
<td>40%</td>
<td>-</td>
</tr>
<tr>
<td>Nodes Positive</td>
<td>-</td>
<td>37%</td>
<td>-</td>
</tr>
<tr>
<td>ER+</td>
<td>-</td>
<td>63%</td>
<td>-</td>
</tr>
<tr>
<td>HER2+</td>
<td>-</td>
<td>27%</td>
<td>-</td>
</tr>
<tr>
<td>ER-</td>
<td>-</td>
<td>24%</td>
<td>-</td>
</tr>
</tbody>
</table>

*Proportions in each category are reported due to missing/incomplete data for these variables in each of the studies. n/a = data not available.

** Cases were individually matched to controls for recruitment source, year of completion of the baseline questionnaire at enrolment, ethnicity, availability of blood sample and date of birth within 12 months and duration that the blood sample was in the mail.

*** Cases and controls were selected within strata of menopausal status (pre- and post-) and individually matched on age, recruitment centre, date and time of blood collection.
2.2.2 Laboratory methods

2.2.2.1 DNA extraction

DNA samples were extracted from whole blood using Qiagen DNA blood Mini Kits in KConFab. DNA samples from BGS and EPIC were extracted from buffy coats using DNA Blood Mini Kits (Qiagen, UK), except for 29 cases and 15 controls in BGS extracted using Nucleon Genomic DNA Extraction Kit (Tepnel, Life Sciences, UK).

2.2.2.2 Bisulphite conversion

500 ng of DNA (KConFab) or 250ng (BGS and EPIC studies) from each subject was bisulphite converted using EZ-96 DNA Methylation-Gold kit according to the manufacturer’s protocol (Zymo Research, Orange, CA).

2.2.2.3 PCR and pyrosequencing

PCR and pyrosequencing assays for ATM were previously described (Flanagan et al., 2009a). ATM primers are shown in table 2.2. Nested PCR was used to amplify repetitive element regions of ATM, with the first-round primer set located outside of repetitive regions. A common tag was incorporated at the 5’ end of reverse primers of each assay and a universal biotinylated primer was included in the second PCR reaction round for each ATM assay as previously described (Royo et al., 2007). The first round of PCR was carried out in a 10µl volume consisting of 0.2ul Faststart taq DNA polymerase (Roche), 0.6ul MgCl2 (25mM), 1µl 10x PCR buffer (Roche), 0.5µl of 10mM DNTP mix (Geneamp), 0.2µl of outer forward (F) and outer reverse (R) PCR primers (10mM), 6.6ul H2O and 1µl of bisulphite treated DNA template. PCR thermocycling conditions for the 1st round of PCR included denaturation at 95°C for 4 min, followed by 10 cycles of 94°C for 15 s, touchdown from 60–50°C (~1 degree/cycle) for 15 s and 72°C for 20 s, followed by a further 30 cycles of
denaturation at 94°C for 20 s, annealing at 50°C for 20 s, and extension at 72°C for 20 s. Round 1 product were diluted 1:10 by adding 90µl H₂O. Diluted product (1µl) was added to a 45ul master mix including 0.45ul taq polymerase, 2.7µl Mgcl2, 6.3µl 10 x buffer, 1.8µl DNTP mix, 0.9µl of each of inner forward and inner reverse primers at 10mM, and 31.95µl H₂O.

PCR thermocycling conditions for the 2nd round of PCR included denaturation at 95°C for 10 min, followed by 40 cycles of 95°C for 20 s, 50°C, 72°C for 20 s, followed by a final annealing step of 72°C for 5 min. Thermocycling conditions were the same as for the first round PCR step. For long interspersed nuclear element-1 (LINE-1) methylation analysis, commercially available primers (Qiagen UK) were used as per user manual. Each PCR product was confirmed as a single product by ethidium bromide-stained agarose gel electrophoresis, along with no-template control samples. Pyrosequencing was carried out using the PyroMark Q96 MD system from Qiagen, following standard protocol as per user manual, and 10µl PCR product for each reaction. Methylation values were calculated as an average of all high quality CpG sites (determined as “passed” by the quality-control thresholds within the Pyro Q-CpG Software (Qiagen, UK)). The Pyro Q-CpG Software has inbuilt overall quality assessment for each sample which flags any sequence that deviates from the expected pattern. Any sample failing quality control was removed from analysis. The number of samples failing in each assay were ATMmp2a (55/1436 subjects), ATMmp2b (56/1436 subjects) and LINE1 (87/1436 subjects). Additionally, a commercially available fully methylated genomic DNA sample was used as a positive control (Zymo Research, Orange, CA) and in-house whole genome amplified genomic DNA (Genomiphi, GE Healthcare) used as an unmethylated negative control. The percentage of cells with methylated DNA at each MVP was calculated as the average of 3 (ATMmp2a) or 4 (LINE1 and ATMmp2b) CpG sites and was used as the measure of methylation for each subject. Based on previous experimental results the range for a typical assay is 90-98% for the positive control and 1-6% for the negative control. Further quality assurance was performed
with blinded duplicate samples (12 pairs) in the BGS and reference controls on each plate with median differences of 2.1%, 3.4%, 1.8% for ATMvmp2a, ATMvmp2b and LINE1, respectively.
Table 2.2. *ATM* methylation primers

<table>
<thead>
<tr>
<th>Universal biotinylated primer</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>biotin-GGGACACCGCTGATCGTTTA</strong></td>
<td><strong>ATMmvp2a F (outer)</strong> AGTTGAGATGGAAGTGGTAGA</td>
</tr>
<tr>
<td><strong>ATMmvp2a R (outer)</strong> CTTTTTATTACTCTAAAAACCAAAA</td>
<td><strong>ATMmvp2a F (inner)</strong> gacgggacacccgtgtctgtaCATTAACAAATAACTATT</td>
</tr>
<tr>
<td><strong>ATMmvp2a R (inner)</strong> gacgggacacccgtgtctgtaCATTAACAAATAACTATT</td>
<td><strong>ATMmvp2a sequencing</strong> GTTATTAGGTTGGAGTGTAG</td>
</tr>
<tr>
<td><strong>ATMmvp2b F (outer)</strong> GAGTGTTTAATAGTTTATGTTTAATG</td>
<td><strong>ATMmvp2b R (outer)</strong> AACACAATAATTTCTTAACATTTCC</td>
</tr>
<tr>
<td><strong>ATMmvp2b F (inner)</strong> TTGGTTTATAGTTAATGAAAGTA</td>
<td><strong>ATMmvp2b R (inner)</strong> gacgggacacccgtgtctgtaCATTAACAAATAACTATT</td>
</tr>
<tr>
<td><strong>ATMmvp2b sequencing</strong> TTTTGAATAGTTGGGATTAT</td>
<td><strong>ATMmvp2b sequencing</strong> TTTTGAATAGTTGGGATTAT</td>
</tr>
</tbody>
</table>

*F=forward, R=reverse, S=sequencing

*Lower case letters represent universal sequence tag (to which universal biotinylated primer binds)*.
2.2.2.4 Statistical Analysis

The non-parametric Wilcoxon test for matched pairs was used for BGS and EPIC and the Wilcoxon rank sum test was used for KConFab, to test for differences in mean methylation levels between cases and controls. Non-parametric tests were used as ATM DNA methylation was not normally distributed, as indicated by a Shapiro Wilks test ($p<0.05$). DNA methylation Z-scores were calculated for each sample in order to make DNA methylation data comparable between studies. Z-scores were calculated for DNA methylation data categorised in quintiles based on their distribution in the combined control population.

Logistic regression was used to estimate odds ratios (OR) and 95% confidence intervals (CI) for individuals in the 2nd, 3rd, 4th and 5th methylation quintiles, compared with individuals in the 1st (lowest (reference)) quintile. This enabled comparison of breast cancer risk between groups within DNA methylation categories at either end of the DNA methylation range (e.g highest versus lowest methylation category). Analyses of combined data from all studies were adjusted by age in 5-year categories and study. Age at blood draw, age at menarche, parity, age at menopause, alcohol consumption, body mass index (BMI), oral contraceptive and hormone replacement use and family history of breast cancer were tested as potential significant effect modifiers (i.e. variables that change the effect estimate by >10%) using multivariable linear regression. Analyses were stratified by age at blood drawn using tertiles (33%) of the combined control population (21-49, >49-59, >59-91), and by family history and time from blood collection to diagnosis to evaluate effect modification by these variables.

Estimates from conditional logistic regression models for individually matched pairs in the BGS and EPIC cohorts were similar to estimates from unconditional logistic models adjusted or unadjusted by matching factors. Only findings from the unconditional logistic analyses are presented to avoid loss of data from exclusion of pairs with one member excluded because of missing methylation data or other reasons (see study population section). Heterogeneity of estimates by study was tested by including an interaction term for the biomarker and an
indicator variable for study in the logistic model. Fixed-effect meta-analyses of estimated ORs from all studies in this report and a previously published study were performed in R using the “metafor” package. Multinomial (ordinal) logistic regression models with categories of methylation levels as the outcome variable were used to test for associations between methylation levels and the breast cancer risk factors specified above, adjusted for age. B-spline quadratic logistic regression models fitted in the “bs” R package were used to explore the relationship between continuous measures of methylation levels and breast cancer risk. All statistical tests were performed using R (v 2.12.0). Statistical analyses were performed by Dr. James M. Flanagan, Prof. Montserrat Garcia-Closas and Dr. Charlotte Wilhelm-Benartzi.

2.3 Results

For the ATMmvp2a locus there was significantly higher median methylation in cases than controls in the familial samples from KConFab (81.8% vs. 76.9%, \( p=4.87 \times 10^{-6} \); Table 2) and marginally higher median methylation in the population-based cases from BGS compared with controls (76.8% vs 76.4%, \( p=0.02 \)). There was no significant differences in mean methylation levels in cases compared with controls in the EPIC cohort (75.7% vs 76.1%, \( p=0.40 \)). There was an upward shift in the distribution of methylation in cases compared with controls in BGS and KConFab, which was not observed in EPIC (Fig 2.1). No significant differences were found for methylation levels at the ATMmvp2b locus (131bp downstream from ATMmvp2a) or LINE1 in any of the studies (Table 2.3).
Table 2.3 Distribution of methylation levels in ATMvmp2a, ATMvmp2b and LINE1 in breast cancer cases and controls

<table>
<thead>
<tr>
<th>Assay</th>
<th>Study</th>
<th>Control N*</th>
<th>Case N*</th>
<th>Control Median</th>
<th>Case Median</th>
<th>Control IQR</th>
<th>Case IQR</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATMvmp2a</td>
<td>BGS</td>
<td>248</td>
<td>249</td>
<td>76.4</td>
<td>76.8</td>
<td>70.2-80.2</td>
<td>70.9-82.7</td>
<td>0.02</td>
</tr>
<tr>
<td>ATMvmp2b</td>
<td>EPIC</td>
<td>283</td>
<td>235</td>
<td>76.1</td>
<td>75.7</td>
<td>70.5-80.6</td>
<td>70.0-80.8</td>
<td>0.4</td>
</tr>
<tr>
<td>ATMvmp2b</td>
<td>KConFab</td>
<td>210</td>
<td>156</td>
<td>76.9</td>
<td>81.8</td>
<td>71.6-81.5</td>
<td>75.8-86.5</td>
<td>4.87x10^-6</td>
</tr>
<tr>
<td>ATMvmp2b</td>
<td>BGS</td>
<td>234</td>
<td>248</td>
<td>91</td>
<td>91.4</td>
<td>87.0-94.8</td>
<td>85.6-95.0</td>
<td>0.61</td>
</tr>
<tr>
<td>ATMvmp2b</td>
<td>EPIC</td>
<td>287</td>
<td>240</td>
<td>92.2</td>
<td>92.3</td>
<td>87.3-95.2</td>
<td>88.3-95.7</td>
<td>0.36</td>
</tr>
<tr>
<td>ATMvmp2b</td>
<td>KConFab</td>
<td>208</td>
<td>162</td>
<td>92.6</td>
<td>92.3</td>
<td>87.2-96.3</td>
<td>82.4-96.5</td>
<td>0.24</td>
</tr>
<tr>
<td>LINE1</td>
<td>BGS</td>
<td>242</td>
<td>241</td>
<td>79</td>
<td>79</td>
<td>77.9-80.1</td>
<td>78.1-79.9</td>
<td>0.96</td>
</tr>
<tr>
<td>LINE1</td>
<td>EPIC</td>
<td>263</td>
<td>232</td>
<td>75.1</td>
<td>75.2</td>
<td>73.9-76.3</td>
<td>73.9-76.3</td>
<td>0.89</td>
</tr>
<tr>
<td>LINE1</td>
<td>KConFab</td>
<td>218</td>
<td>153</td>
<td>76</td>
<td>76.6</td>
<td>74.3-78.0</td>
<td>75.2-77.6</td>
<td>0.2</td>
</tr>
</tbody>
</table>

* Differences in numbers of cases and controls within each study with total numbers are due to missing data (failed QC) on methylation markers.

** Wilcoxon matched pairs test for BGS and EPIC and Wilcox-rank-sum test for KConFab
Figure 2.1. ATMmvp2a methylation distribution in breast cancer cases and controls. BGS (A), EPIC (B) and KConFab (C). DNA methylation of peripheral blood DNA displayed as a density distribution with controls in black and cases in red.
Quintile analyses for the ATMmvp2a locus, adjusted by age at blood collection in 5 year categories, showed a significantly increased risk of breast cancer for women in the highest quintile compared with the lowest quintile in the BGS and KConFab studies, but not in EPIC (Table 2.4). Further adjustment by age as continuous variable and conditional logistic analyses for paired samples individually matched in BGS and EPIC showed similar results (data not shown). Analyses of combined data from all studies adjusting by study and age at blood collection indicated that women in the highest quintile (>6.3% methylation above study mean methylation) were at 1.9-fold increased risk of breast cancer compared with women in the lowest quintile (OR=1.89 (1.36-2.64), p=1.64x10^{-4}) (Table 2.4). While the overall difference in median levels between cases and controls was small (1.1%), the difference in median methylation between the highest quintile (86%) and lowest (65%), where the association with cancer status is observed was large (21%). A quadratic B-spline regression model of continuous levels of methylation at ATMmvp2a and breast cancer risk confirmed a threshold association, rather than a linear association, between methylation levels and breast cancer risk (Fig 2.2).
Table 2.4: Association between methylation levels in ATMmvp2a and breast cancer risk in BGS, EPIC, KConFab and Combined analysis

<table>
<thead>
<tr>
<th>Study</th>
<th>Quintile*</th>
<th>Meth Range</th>
<th>N</th>
<th>Freq.</th>
<th>N</th>
<th>Freq.</th>
<th>OR**</th>
<th>95%CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Controls</td>
<td></td>
<td></td>
<td></td>
<td>Cases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BGS</td>
<td>Qi1</td>
<td>3.4 - 68.0%</td>
<td>50</td>
<td>0.2</td>
<td>35</td>
<td>0.14</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Qi2</td>
<td>68.0 - 74.1%</td>
<td>40</td>
<td>0.16</td>
<td>46</td>
<td>0.18</td>
<td>1.56</td>
<td>0.85-2.88</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Qi3</td>
<td>74.1 - 77.6%</td>
<td>54</td>
<td>0.22</td>
<td>47</td>
<td>0.19</td>
<td>1.15</td>
<td>0.66-2.16</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Qi4</td>
<td>77.6 - 81.0%</td>
<td>56</td>
<td>0.23</td>
<td>42</td>
<td>0.17</td>
<td>1</td>
<td>0.55-1.81</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Qi5</td>
<td>81.0 - 91.7%</td>
<td>48</td>
<td>0.19</td>
<td>79</td>
<td>0.32</td>
<td>2.31</td>
<td>1.31-4.06</td>
<td>3.7x10^-3</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td></td>
<td>248</td>
<td></td>
<td>249</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPIC</td>
<td>Qi1</td>
<td>53.6 - 69.7%</td>
<td>60</td>
<td>0.21</td>
<td>49</td>
<td>0.21</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Qi2</td>
<td>69.7 - 74.8%</td>
<td>61</td>
<td>0.22</td>
<td>51</td>
<td>0.22</td>
<td>1.02</td>
<td>0.60-1.74</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Qi3</td>
<td>74.8 - 78.6%</td>
<td>58</td>
<td>0.19</td>
<td>46</td>
<td>0.2</td>
<td>0.97</td>
<td>0.56-1.67</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Qi4</td>
<td>78.6 - 82.4%</td>
<td>49</td>
<td>0.17</td>
<td>38</td>
<td>0.16</td>
<td>0.95</td>
<td>0.54-1.68</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Qi5</td>
<td>82.4 - 97.5%</td>
<td>55</td>
<td>0.19</td>
<td>51</td>
<td>0.22</td>
<td>1.13</td>
<td>0.66-1.94</td>
<td>0.76</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td></td>
<td>283</td>
<td></td>
<td>235</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Cut-off values determined by quintiles of Z-scores based on the distribution in the combined control population. Z-score cut off values are -0.83, -0.24, 0.19 and 0.64.

**ORs within each study are adjusted by 5-year age categories, with further adjustment by menopausal status in EPIC to account for stratified sampling. ORs in combined analyses are adjusted by 5-year age categories and study (with EPIC defined by two categories of menopausal status to account for stratified sampling).
Table 2.4: Association between methylation levels in ATMvmp2a and breast cancer risk in BGS, EPIC, KConFab and Combined analysis (continued from previous page)

<table>
<thead>
<tr>
<th>Study</th>
<th>Quintile*</th>
<th>Meth Range</th>
<th>Controls</th>
<th>Cases</th>
<th>OR**</th>
<th>95%CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>Freq.</td>
<td>N</td>
<td>Freq.</td>
<td></td>
</tr>
<tr>
<td>KConFab</td>
<td>Qi1</td>
<td>19.0-70.2%</td>
<td>38</td>
<td>0.18</td>
<td>20</td>
<td>0.13</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Qi2</td>
<td>70.2-75.4%</td>
<td>47</td>
<td>0.22</td>
<td>16</td>
<td>0.1</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>Qi3</td>
<td>75.4-79.1%</td>
<td>36</td>
<td>0.17</td>
<td>22</td>
<td>0.14</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>Qi4</td>
<td>79.1-83.0%</td>
<td>43</td>
<td>0.2</td>
<td>31</td>
<td>0.2</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>Qi5</td>
<td>83.0-100%</td>
<td>46</td>
<td>0.22</td>
<td>67</td>
<td>0.43</td>
<td>3.06</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td></td>
<td>210</td>
<td></td>
<td>156</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Combined  | Qi1       | -71.3 – -6.5% | 148     | 0.2   | 104  | 0.16       | 1       |
|           | Qi2       | -6.5 – -1.2%  | 148     | 0.2   | 113  | 0.18       | 1.08    | 0.76-1.54 | 0.64   |
|           | Qi3       | -1.2 – 2.5%   | 148     | 0.2   | 115  | 0.18       | 1.09    | 0.77-1.54 | 0.64   |
|           | Qi4       | 2.5% – 6.3%   | 148     | 0.2   | 111  | 0.17       | 1.06    | 0.74-1.51 | 0.75   |
|           | Qi5       | 6.3% –23.3%   | 149     | 0.2   | 197  | 0.31       | 1.89    | 1.36-2.64 | 1.6x10^-4 |
| Totals    |           |            | 741      |       | 640  |           |         |

* Cut-off values determined by quintiles of Z-scores based on the distribution in the combined control population. Z-score cut off values are -0.83, -0.24, 0.19 and 0.64.

**ORs within each study are adjusted by 5-year age categories, with further adjustment by menopausal status in EPIC to account for stratified sampling. ORs in combined analyses are adjusted by 5-year age categories and study (with EPIC defined by two categories of menopausal status to account for stratified sampling).
**Figure 2.2:** Effect estimate of the logistic model of splined ATMmp2a methylation as it relates to case-control status. Combined analysis of three prospective case-control studies using Z-scores, shows the cutoff level for quintile five at +6.3% above the median methylation level in controls on the X-axis with log odds ratio of the quadratic splined ATM methylation on the Y-axis. Yellow dotted line represents 95% CI. Black vertical bars each represent a sample, indicating the frequency distribution of samples within the methylation range.
Study-adjusted analyses stratified by age at blood collection suggested a weaker ATMmvp2a risk association when methylation was measured in samples collected from women >59 years of age (Table 2.5). Similar results were obtained when stratified by age at diagnosis (data not shown). However, age-specific estimates within study revealed that the weaker association was driven by the EPIC cohort that showed no increased risk in this age subgroup (Fig 2.3). Overall analyses showed some evidence for heterogeneity of estimates between studies (Table 2.4; p-value for test for heterogeneity=0.07). This evidence was limited to women in the older age group and there was no evidence for study heterogeneity within the younger age subgroups (p-value for study heterogeneity by age subgroups 21-49 (p=0.51), 50-59 (p=0.72), 60-91 years (p=0.09)) (Table 2.5).

There was a significant association between ATMmvp2a methylation levels and increasing age at blood collection in controls (Spearman’s rho=0.15, p=0.0015), but not in cases (Spearman’s rho=-0.02, p=0.43), that was most significant in the EPIC cohort (rho=0.11, p=0.007) compared to KConFab (rho=0.06, p=0.26) and BGS (rho=0.02, p=0.40). This underlying age association may account for the apparent cross-over risk association with ATM methylation by age at blood collection seen in EPIC (Fig 2.4). Analyses by menopausal status at blood collection, adjusted by study and age, showed similar risk estimates for pre- and post-menopausal women (Table 2.6).
Table 2.5. Association between methylation levels in ATMmp2a and breast cancer risk in combined analysis stratified by age at blood draw

<table>
<thead>
<tr>
<th>Age range</th>
<th>Quintiles</th>
<th>Controls</th>
<th>Cases</th>
<th>Proportions By Study</th>
<th>OR**</th>
<th>95%CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>Freq.</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(K, B, E)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21-49 years</td>
<td>Qi1</td>
<td>45</td>
<td>0.21</td>
<td>44</td>
<td>0.17</td>
<td>0.20, 0.27, 0.52</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Qi2</td>
<td>51</td>
<td>0.23</td>
<td>39</td>
<td>0.15</td>
<td>0.28, 0.26, 0.46</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Qi3</td>
<td>44</td>
<td>0.2</td>
<td>41</td>
<td>0.16</td>
<td>0.27, 0.32, 0.41</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Qi4</td>
<td>38</td>
<td>0.18</td>
<td>44</td>
<td>0.17</td>
<td>0.43, 0.18, 0.39</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Qi5</td>
<td>39</td>
<td>0.18</td>
<td>90</td>
<td>0.35</td>
<td>0.40, 0.33, 0.27</td>
<td>2.07***</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>217</td>
<td>258</td>
<td></td>
<td></td>
<td>0.33, 0.28, 0.38</td>
<td></td>
</tr>
<tr>
<td>&gt;49-59 years</td>
<td>Qi1</td>
<td>54</td>
<td>0.2</td>
<td>27</td>
<td>0.13</td>
<td>0.19, 0.41, 0.41</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Qi2</td>
<td>55</td>
<td>0.2</td>
<td>43</td>
<td>0.21</td>
<td>0.05, 0.49, 0.47</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>Qi3</td>
<td>62</td>
<td>0.23</td>
<td>48</td>
<td>0.24</td>
<td>0.19, 0.38, 0.44</td>
<td>1.56</td>
</tr>
<tr>
<td></td>
<td>Qi4</td>
<td>50</td>
<td>0.19</td>
<td>33</td>
<td>0.16</td>
<td>0.09, 0.48, 0.42</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>Qi5</td>
<td>49</td>
<td>0.18</td>
<td>53</td>
<td>0.26</td>
<td>0.19, 0.49, 0.32</td>
<td>2.25***</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>270</td>
<td>204</td>
<td></td>
<td></td>
<td>0.14, 0.45, 0.41</td>
<td></td>
</tr>
<tr>
<td>&gt;59-91 years</td>
<td>Qi1</td>
<td>49</td>
<td>0.19</td>
<td>33</td>
<td>0.19</td>
<td>0.18, 0.36, 0.45</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Qi2</td>
<td>42</td>
<td>0.17</td>
<td>31</td>
<td>0.17</td>
<td>0.10, 0.48, 0.42</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Qi3</td>
<td>42</td>
<td>0.17</td>
<td>26</td>
<td>0.15</td>
<td>0.08, 0.62, 0.31</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Qi4</td>
<td>60</td>
<td>0.24</td>
<td>34</td>
<td>0.19</td>
<td>0.26, 0.53, 0.21</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>Qi5</td>
<td>61</td>
<td>0.24</td>
<td>54</td>
<td>0.31</td>
<td>0.39, 0.43, 0.19</td>
<td>1.39**</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>254</td>
<td>178</td>
<td></td>
<td></td>
<td>0.23, 0.47, 0.30</td>
<td></td>
</tr>
</tbody>
</table>

* Study proportions for KConFab, BGS and EPIC contributing to each quintile in each age group.

**ORs within each age-group are adjusted by study. Test for Heterogeneity of effects by age group in combined analysis p= 0.3109 for Qi5 vs Qi1

***P-values for study heterogeneity for ORs comparing Qi5 vs Qi1 within age subgroups: 21-49 years (p=0.51), 50-59 years (p>0.73), 60-91 years (p=0.09)
Figure 2.3. Association between ATMmvp2a methylation levels and breast cancer risk in combined analysis stratified by age. Study specific estimates (unadjusted) are shown. Study heterogeneity by age subgroups was 21-49 (p=0.51), 50-59 (p=0.72), 60-91 years (p=0.09).
Figure 2.4. ATMvmp2a DNA methylation in blood, measured by pyrosequencing plotted against age at blood draw. (A) combined control samples from the EPIC, BGS and KConFab studies. Linear regression lines for association between ATMvmp2a methylation and age are shown for EPIC (black), BGS (red) and KConFab (blue), with corresponding p-values, and Spearman correlation rho-values shown. (B) Combined cases and controls from all three studies. Linear regression lines for association between ATMvmp2a methylation and age are shown for controls (black), and cases (red) with corresponding p-values, and Spearman correlation rho-values.
Table 2.6. Association between methylation levels in ATMvmp2a and breast cancer risk in combined analysis stratified by menopausal status at blood drawn

<table>
<thead>
<tr>
<th>Menopausal status</th>
<th>Quintiles*</th>
<th>N</th>
<th>Freq.</th>
<th>N</th>
<th>Freq.</th>
<th>OR**</th>
<th>95%CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-menopausal</td>
<td>Qi1</td>
<td>51</td>
<td>0.17</td>
<td>50</td>
<td>0.15</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Qi2</td>
<td>69</td>
<td>0.23</td>
<td>60</td>
<td>0.18</td>
<td>0.82</td>
<td>0.48-1.39</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>Qi3</td>
<td>69</td>
<td>0.23</td>
<td>61</td>
<td>0.18</td>
<td>0.89</td>
<td>0.53-1.51</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Qi4</td>
<td>62</td>
<td>0.2</td>
<td>65</td>
<td>0.19</td>
<td>1.01</td>
<td>0.59-1.71</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Qi5</td>
<td>53</td>
<td>0.17</td>
<td>103</td>
<td>0.3</td>
<td>1.76</td>
<td>1.04-2.96</td>
<td>0.03</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>304</td>
<td></td>
<td>339</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-menopausal</td>
<td>Qi1</td>
<td>90</td>
<td>0.22</td>
<td>49</td>
<td>0.18</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Qi2</td>
<td>76</td>
<td>0.19</td>
<td>48</td>
<td>0.18</td>
<td>1.16</td>
<td>0.7-1.92</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Qi3</td>
<td>73</td>
<td>0.18</td>
<td>46</td>
<td>0.17</td>
<td>1.19</td>
<td>0.71-1.98</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Qi4</td>
<td>77</td>
<td>0.19</td>
<td>44</td>
<td>0.17</td>
<td>1.13</td>
<td>0.67-1.89</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Qi5</td>
<td>87</td>
<td>0.22</td>
<td>79</td>
<td>0.3</td>
<td>1.8</td>
<td>1.12-2.88</td>
<td>0.02</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>403</td>
<td></td>
<td>266</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test for Heterogeneity of effects by menopausal status for combined analysis p= 0.62 for Qi5 vs Qi1

* Cut-off values determined by quintiles of Z-scores based on the distribution in the combined control population. Z-score cut off values are -0.83, -0.24, 0.19 and 0.64.

**ORs adjusted by study and age in 5-year categories.
Adjustment by breast cancer risk factors (age at menarche, menopausal status at blood drawn, parity, age at menopause, alcohol consumption, body mass index, oral contraceptive and hormone replacement use and family history of breast cancer) did not result in appreciable changes in relative risk estimates for any of the markers across the three studies in this report. Consistently, these risk factors were not significantly associated with ATMmvp2a methylation levels in any of the three control populations (data not shown). In the familial cases from KConFab, ATMmvp2a methylation was not associated with BRCA1/2 mutation status, tumour pathology (morphology, grade, node status, ER, PR and HER2 status) (data not shown). Similarly in EPIC and BGS samples, ATMmvp2a methylation was not associated with tumour pathological features (data not shown).

Analyses stratified by length of time between blood collection and diagnosis, known as time to diagnosis (TTD) (<=1 year vs. >1 year) showed no significant differences in effect estimates using the combined data (Table 2.7). Consistently, there was no association between ATMmvp2a methylation and TTD (0-11 years) using linear regression (KConFab p=0.97; BGS p=0.10, EPIC p=0.28).
Table 2.7. Association between ATMMvp2a methylation levels and breast cancer risk by time from blood collection to diagnosis in combined analysis

<table>
<thead>
<tr>
<th>Study</th>
<th>Quintile*</th>
<th>Controls N</th>
<th>Controls Freq.</th>
<th>Cases N</th>
<th>Cases Freq.</th>
<th>OR** 95%CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1yr from diagnosis</td>
<td>Qi1 148</td>
<td>0.2</td>
<td>18</td>
<td>0.16</td>
<td>1</td>
<td>0.59-2.43</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Qi2 148</td>
<td>0.2</td>
<td>20</td>
<td>0.18</td>
<td>1.19</td>
<td>0.53-2.19</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Qi3 148</td>
<td>0.2</td>
<td>20</td>
<td>0.18</td>
<td>1.07</td>
<td>0.47-2.19</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>Qi4 148</td>
<td>0.2</td>
<td>19</td>
<td>0.17</td>
<td>0.97</td>
<td>1.05-3.86</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Qi5 149</td>
<td>0.2</td>
<td>35</td>
<td>0.31</td>
<td>2.02</td>
<td>1.94-2.75</td>
<td>2.2x10^-4</td>
</tr>
<tr>
<td>Totals</td>
<td>741</td>
<td>112</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| >1yr from diagnosis | Qi1 148 | 0.2 | 86 | 0.16 | 1 | 0.74-1.57 | 0.68 |
|                     | Qi2 148 | 0.2 | 93 | 0.18 | 1.08 | 0.76-1.60 | 0.6 |
|                     | Qi3 148 | 0.2 | 95 | 0.18 | 1.1 | 0.75-1.59 | 0.65 |
|                     | Qi4 148 | 0.2 | 92 | 0.17 | 1.09 | 1.36-2.75 | 2.2x10^-4 |
|                     | Qi5 149 | 0.2 | 162 | 0.31 | 1.94 | 1.94-2.75 | 2.2x10^-4 |
| Totals             | 741      | 528       |                |         |             |              |         |

Test for Heterogeneity of effects by time from blood to Dx p= 0.93 for Qi5 vs Qi1

* Cut-off values determined by quintiles of Z-scores based on the distribution in the combined control population. Z-score cut off values are -0.83, -0.24, 0.19 and 0.64.

**ORs adjusted by 5-year age categories and study.
A risk biomarker that is measured at one time point only (as is the case with most case-control studies) would ideally be stable over time. Temporal stability of ATMMvp2a and LINE1 DNA methylation was assessed by methylation analysis of ATMMvp2a within a control population where blood samples were taken 6 years apart from the same individuals in the BGS cohort (n=92 pairs). There was no significant change in either ATMMvp2a (median change = 0.19%, p= 0.51) or LINE1 (median change = 0.27%, p= 0.69) over 6 years, using conditional logistic regression. Conditional logistic regression showed that ATM variation between individuals was much larger than within individuals at two time points (ICC=0.57; between time points correlation R^2= 0.79) and there was no significant difference in median methylation between time points overall (p=0.24 (paired T-test)) (fig 2.5). These data suggest that the ATM methylation is stable for at least ~6 years.
Figure 2.5. Difference in ATMmvp2a and LINE1 DNA methylation between time-points in serially collected blood samples. Methylation of (A) ATMmvp2a and (B) LINE1 in 92 pairs of serially collected blood samples from healthy women from the Breakthrough Generations study (BGS). Samples were collected at T0 and T1 (~6 years later) for each individual. The difference in methylation between T0 and T1 for each individual (Y-axis), is plotted against age (X-axis). Central horizontal line represents 0% difference in methylation. Broken lines represent confidence intervals.
Inter-quartile analysis of the association between ATMmvp2b DNA methylation and bilateral breast cancer prevalence in a retrospective case-control study was previously reported (Flanagan et al., 2009a). Re-analysis of this data for both the ATMmvp2a and ATMmvp2b regions was performed, using quintiles (based on control methylation distribution) rather than quartiles as methylation categories. Consistent with prospective studies, individuals within the highest ATMmvp2a methylation quintile were significantly more likely to be cases compared with individuals within the lowest quintile (age-adjusted OR=1.90 (95% CI=1.00-3.62), p=0.05). For the ATMmvp2b region, case-status was also significantly more likely for women within the highest DNA methylation quintile compared with women in the lowest (age-adjusted OR=3.07 (95% CI=1.58-5.93), p=8.8x10^{-4}) (Table 2.8). Meta-analysis of odds ratio estimates for all three prospective studies and the retrospective BBC study indicated that ATMmvp2a methylation within the highest quintile was associated with an OR for breast cancer risk of 1.89 ratio (95% CI 1.36-2.64, p=1.64x10^{-4}) compared with women in the lowest quintile (fig 2.6), with no significant heterogeneity between studies (p=0.15).
Test for Heterogeneity: $Q(df = 3) = 5.3232$, $p$-val = 0.1496

**Figure 2.6. Meta-analysis of ATM methylation in four case–control studies.** Methylation levels at ATMmvp2a were separated into control quintile ranges (using identical ranges from combined analysis), for the previously published report on the British Breast Cancer Study (BBCS ‡)(Flanagan et al., 2009) and the three new case-control studies. Odds ratio of highest vs. lowest methylation quartile is shown. Combined overall analysis was determined using a random effects model.
**Table 2.8.** Association between methylation levels at the ATM locus and breast cancer risk in the British Bilateral Breast Cancer Study (BBCS) (Flanagan et al., 2009b)

<table>
<thead>
<tr>
<th>Quintiles</th>
<th>Methylation range</th>
<th>Controls (n)</th>
<th>Frequency of controls</th>
<th>Cases (n)</th>
<th>Frequency of cases</th>
<th>OR</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ATMmvp2a</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qi1</td>
<td>61.6 - 74.3</td>
<td>38</td>
<td>0.2</td>
<td>28</td>
<td>0.15</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qi2</td>
<td>74.3 - 78.0</td>
<td>38</td>
<td>0.2</td>
<td>34</td>
<td>0.18</td>
<td>1.22</td>
<td>0.62-2.38</td>
<td>0.57</td>
</tr>
<tr>
<td>Qi3</td>
<td>78.0 - 80.4</td>
<td>38</td>
<td>0.2</td>
<td>33</td>
<td>0.17</td>
<td>1.18</td>
<td>0.60-2.31</td>
<td>0.64</td>
</tr>
<tr>
<td>Qi4</td>
<td>80.4 - 83.6</td>
<td>38</td>
<td>0.2</td>
<td>41</td>
<td>0.22</td>
<td>1.46</td>
<td>0.75-2.82</td>
<td>0.26</td>
</tr>
<tr>
<td>Qi5</td>
<td>83.6 - 96.1</td>
<td>38</td>
<td>0.2</td>
<td>53</td>
<td>0.28</td>
<td>1.9</td>
<td>1.00-3.62</td>
<td>0.05</td>
</tr>
</tbody>
</table>

| **ATMmvp2b** |                   |              |                       |           |                    |     |        |        |
| Qi1       | 53.0 - 86.8       | 38           | 0.2                   | 22        | 0.12               | 1   |         |        |
| Qi2       | 86.8 - 89.8       | 38           | 0.2                   | 30        | 0.16               | 1.36| 0.67-2.77| 0.39   |
| Qi3       | 89.8 - 91.7       | 37           | 0.2                   | 35        | 0.18               | 1.64| 0.82-3.31| 0.17   |
| Qi4       | 91.7 - 93.5       | 38           | 0.2                   | 36        | 0.19               | 1.63| 0.81-3.27| 0.17   |
| Qi5       | 93.5 - 98.2       | 38           | 0.2                   | 67        | 0.35               | 3.07| 1.58-5.93| 8.8x10\(^{-4}\) |

*ORs for each marker are adjusted by 5-year age categories.

** Cut-off values determined by quintiles of Z-scores determined based on the distribution in the control population. Z-score cut off values are -0.77, -0.13, 0.29 and 0.83 for ATMmvp2a, and -0.63, 0.03, 0.38 and 0.75 for ATMmvp2b.
2.4 Discussion

Our findings indicate that ATMvmp2a hypermethylation may represent a marker of breast cancer risk, potentially making this of the first known examples of epigenetic variability associated with breast cancer risk.

Previously reported epigenetic cancer risk markers include rare hypermethylation events within gene promoters that are normally unmethylated (epimutations) (Hansmann et al., 2012; Hesson et al., 2010), and significantly decreased levels of genome-wide DNA methylation (Woo and Kim, 2012). ATM hypermethylation is distinct in that it represents a slightly increased frequency of methylated cells in WBCs of cancer cases relative to controls, detectable several years before diagnosis. Furthermore, all previous studies investigating gene-specific DMRs used retrospectively collected samples, often with small sample size, and using non-quantitative methylation-analysis methods (Widschwendter et al., 2008), therefore this study utilised a more thorough approach to investigation of DNA methylation and cancer risk. Candidate cancer risk markers similar to the ATMvmp2a DMR have been reported by retrospective discovery studies using candidate gene (Widschwendter et al., 2008) and array-based (Langevin et al., 2012) approaches. External validation of loci identified in these studies may provide additional cancer risk biomarkers, as our investigation of ATM methylation supports the notion that associations between DNA methylation and cancer risk in retrospective studies (Flanagan et al., 2009b) may be reproducible in prospective studies. A recent discovery study of DNA methylation associated with breast cancer risk using prospectively collected DNA samples identified several significantly loci which remain to be validated (Xu et al., 2013b). Discovery studies for breast cancer risk are discussed further in chapter 6.
Whereas findings for ATMv2a were consistent between the retrospective BBC study and prospective studies, the previously reported (Flanagan et al., 2009b) strong association between methylation at the neighbouring ATMv2b DMR in the BBC study was not replicated in prospective studies. Differences in findings could be due to differences in study populations including the use of pre-diagnostic versus post-diagnostic blood samples, or the investigation of (mostly) unilateral breast cancer rather than bilateral breast cancer. Alternatively, the association between ATMv2b methylation and bilateral breast cancer prevalence may have occurred by chance. Methylation analysis of ATMv2b in an independent set of blood samples collected from bilateral breast cancer patients at diagnosis would be required for external validation, and to assess the potential value of ATMv2b methylation for diagnostic biomarker development.

Although we did not find significant heterogeneity in ATMv2a risk associations across studies, the evidence for association of ATM DNA methylation with breast cancer risk was strongest for KConFab and weaker for the BGS study, and, weaker still for the EPIC cohort. The stronger association in KConFab may be due to the inclusion of cases with very strong family histories, and/or to the choice of controls that were selected from best friends of KConFab participants, and had no family history of breast cancer. In contrast, BGS and EPIC studies were nested case-control studies within general population cohorts, thus ensuring that cases and controls come from the same source population. There was a correlation between increasing methylation and increasing age at blood draw in the control EPIC population, as has been previously reported for DNA methylation in blood (Gomes et al., 2012). The age correlation was particularly strong in the EPIC cohort, which might explain that the risk association in this cohort was not seen in women with bloods collected at older ages. The use of pre-diagnostic samples from incident breast cancer cases reduces the possibility that methylation variability in WBC DNA was influenced by the presence of clinical
cancer or treatment in these patients. Furthermore, *ATM* methylation was not associated with TTD, indicating that *ATM* hypermethylation may represent a disease susceptibility marker rather than a marker of pre-clinical disease (range of TTD in included studies was <1 month to 11 years). Furthermore, a longitudinal study did not show a significant difference in average ATMmvp2a methylation between blood samples collected at two time-points from the same individuals, suggesting that ATMmvp2a methylation may be temporally stable, and may thus confer long-term breast cancer susceptibility.

The mechanism by which methylation at ATMmvp2a could increase risk is not known, however the association between breast cancer risk and methylation of an intragenic repetitive element in *ATM* is consistent with a functional role for intragenic DNA methylation (Aporntewan et al., 2011; Flanagan and Wild, 2007; Kulis et al., 2013; Shenker and Flanagan, 2012). ATMmvp2a resides within an intragenic *Alu* repetitive element ~4kb downstream of the *ATM* transcription start site (genomic location of the ATMmvp2a shown in Fig 2.7). Potential functional roles of intragenic DNA methylation include regulation of enhancers, non-coding RNAs, gene expression, alternative splicing and alternative gene promoters (Akhtar-Zaidi et al., 2012; Illingworth et al., 2008; Kulis et al., 2013; Maunakea et al., 2010). Several recent studies have confirmed a highly-conserved role of intragenic DNA methylation in regulation of alternative RNA splicing (Flores et al., 2012; Maunakea et al., 2013; Wan et al., 2013) and experimentally induced aberrant intragenic methylation can result in aberrant alternative splicing (Maunakea et al., 2013). Furthermore, tissue-specific DNA methylation is enriched at alternatively spliced genes (Wan et al., 2013), and intragenic DNA methylation and alternative splicing may be associated with gene length (Flores et al., 2012). It is plausible that intragenic *ATM* methylation may play such a role, as *ATM* is a particularly long gene and undergoes alternative splicing (Pagani et al., 2002). Interestingly, recent studies have identified a cryptic enhancer within an intragenic *Alu* element in the *ATM*
gene, which regulates inclusion of an adjacent exon (Pastor and Pagani, 2011; Pastor et al., 2009). Consistently, Alu elements preferentially flank alternatively spliced exons, and may play a number of roles in human gene regulation including regulation of alternative splicing (Lev-Maor et al., 2008; Lev-Maor et al., 2003). Given the location of the ATMmvp2a DMR within an intragenic Alu element, the possibility that DNA methylation at this region regulates alternative splicing may be worth investigating.
Figure 2.7. Single nucleotide resolution methylation of the entire *ATM* gene in PBMCs. Whole genome methylome was obtained by deep sequencing of bisulphite treated DNA from PBMCs of a single individual. Methylation (%) of individual CpG sites are shown (grey dots), with a smoothed average (black line), exonic structure at the top of the graph (shown as black dots) and *ATM* risk locus marked by the vertical red line at +4kb (data obtained from (Li et al., 2010)).
It has been previously reported (LaBreche et al., 2011) that \textit{ATM} expression was decreased in WBCs of breast cancer patients taken at the time of mammographic screening (\(n=51\)), compared with healthy controls (\(n=31\)). Whereas regulation of \textit{ATM} at the protein level is relatively well understood, very little is known about transcriptional regulation of \textit{ATM}. Investigations to date have shown that \textit{ATM} transcription is regulated by binding of E2F1 to the \textit{ATM} promoter \textit{in vitro} (Berkovich and Ginsberg, 2003), and binding of BRCA1, E2F1 and CTIP to the \textit{ATM} proximal promoter is reported to activate \textit{ATM} transcription in response to DNA damage (Moiola et al., 2012). Intragenic \textit{ATM} methylation may play a role in altering gene expression, as hypermethylation of the ATM\textit{mvp2b} DMR was associated with reduced \textit{ATM} expression within a panel of cell lines in a previous study (Flanagan et al., 2009a). Analysis of potential association of ATM\textit{mvp2a} methylation with \textit{ATM} expression remains to be carried out.

Inconsistent with the hypothesis that cancer-risk associated DNA methylation variability may represent an intermediate marker of environmental disease risk factors (Christensen and Marsit, 2011), \textit{ATM} methylation was not associated with any life-style/environmental factor analysed, including smoking or alcohol consumption, as well as hormonal risk factors such as age at menarche, oral contraceptive use, parity, menopausal status and HRT use. Further investigation of factors underlying the association between ATM\textit{mvp2a} methylation and breast cancer risk is reported in chapter 4.

The possibility that \textit{ATM} methylation may confer susceptibility to specific breast cancer subtypes was not supported by our findings, as methylation was not associated with tumour features or subtypes such as triple negative cancer, however, this may be due to limited subject numbers in this study.

Associations of WBC genome-wide DNA methylation (and surrogate measures of genome-wide methylation, such as methylation of LINE1 repetitive elements) with cancer have been
widely reported (Di et al., 2011), including studies investigating breast cancer (Xu et al., 2012b). We found no evidence, however, of association between \textit{LINE1} methylation and breast cancer risk in any study within this report, and found that there was very little variability in \textit{LINE1} methylation overall. This is consistent with a retrospective study reported by our own group (Flanagan et al., 2009b). The collective evidence for association of genome-wide DNA methylation and cancer risk will be further discussed in chapter 3.
CHAPTER 3

GENOME-WIDE METHYLATION AS A CANCER RISK FACTOR
3.1. Introduction

Genome-wide DNA methylation is the epigenetic event most widely studied in blood in relation to cancer risk, and in relation to epigenetic consequences of cancer-related exposures (Nelson et al., 2011). Genome-wide or ‘global’ DNA hypomethylation was first described in tumours as the lower net percentage of 5meC in tumour tissue compared with equivalent normal tissue, and is now known to occur frequently in all cancer types (Feinberg and Vogelstein, 1983; Lapeyre et al., 1981). In recent years, genome-wide hypomethylation in histologically normal tissue and blood of cancer patients has been extensively reported, leading to widespread speculation that genome-wide hypomethylation may be a cancer risk factor, and that measurement of genomic 5meC (and surrogate measures thereof) may represent useful cancer risk biomarkers. However, our prospective investigation of LINE1 methylation in relation to breast cancer risk (chapter 2), and other reports (Cash et al., 2011; Gao et al., 2012; Xu et al., 2012b), fail to reproduce these findings.

3.1.1 Genome-Wide DNA methylation

Despite considerable research, the causes, consequences, and exact nature of genome-wide DNA hypomethylation remain poorly defined. An explanation for the paradoxical coincidence of genome-wide hypomethylation with CpG island hypermethylation in cancer has been long sought, yet it remains unclear whether these two phenomena are independent or mechanistically linked (Estecio et al., 2007; Wild and Flanagan, 2010). In recent years, whole-methylome analysis has revealed that rather than representing a ‘global’ reduction of 5meC, tumour-associated hypomethylation is confined to large genomic ‘hypomethylation blocks’ (Hansen et al., 2011; Pujadas and Feinberg, 2012), that tend to occur in regions that display intermediate methylation levels in normal tissue, termed partially methylated domains (PMDs) (Hansen et al., 2011; Lister et al., 2009), and regions
of low CpG density (Ruike et al., 2010). The prevailing idea that genome-wide hypomethylation reflects loss of methylation primarily at repetitive elements (Aporntewan et al., 2011; Estecio et al., 2007; Yang et al., 2004) has been called into question with only a modest enrichment of repeats in hypomethylation blocks (Hansen et al.; Hon et al., 2012). Whereas genome-wide hypomethylation has long been implicated in loss of transcriptional repression, recent evidence suggests that hypomethylation at PMDs or gene-bodies may be also associated with gene repression through formation of repressive chromatin (Hon et al., 2012).

3.1.2 Causes and Consequences

The cause of genome-wide hypomethylation remains unknown, and several potential contributing factors have been identified (Wild and Flanagan, 2010; Wilson et al., 2007). Exogenous exposures such as carcinogenic compounds may influence DNA methylation by inducing DNA damage or by affecting the activity of DNMT enzymes (Feinberg, 2007; Tajuddin et al., 2013; Yang et al., 2012). Lack of methyl donors required for DNA methylation, due to dietary insufficiency of folate, choline, methionine or cobalamin may induce hypomethylation (Friso et al., 2013; Nelson et al., 2011; Vineis et al., 2011). Hypomethylation has also been linked with age-associated loss of efficiency of DNA methylation maintenance (Rakyan et al., 2010; Teschendorff et al., 2010). Genetic factors, such as polymorphisms in one-carbon metabolism genes, DNMTs, and other genes (Friso et al., 2013; Haggarty et al., 2013; Inoue-Choi et al., 2013; Vineis et al., 2011; Wilson et al., 2007) are associated with hypomethylation. Mutation of the MILI, MIW12 (Di Giacomo et al., 2013), Maelstrom (Soper et al., 2008), ATRX (Gibbons et al., 2000), LSH (Dennis et al., 2001) and CXX1 (Carlone and Skalnik, 2001) genes have been shown to induce hypomethylation of repetitive elements in different tissues and developmental contexts (reviewed by Ooi et al (Ooi et al., 2009)). Lastly, ‘field-effect’ hypomethylation in cancer and
surrounding histologically normal tissues may occur due to sequestration of DNA methylation machinery and substrates by rapidly proliferating cancer cells; therefore, hypomethylation in normal tissues may reflect the presence of preclinical disease within an individual (Lim et al., 2008; Suter et al., 2004; Wilson et al., 2007).

Several tumorigenic consequences result from genome-wide hypomethylation, including chromosomal instability, potentially leading to gene-dosage alterations, increased mutation rates, genetic recombination events, large deletions or translocations (Laird, 2005; Matsuzaki et al., 2005; Wilson et al., 2007). Hypomethylation may also lead to altered expression of oncogenes, tumour suppressor genes, and spurious transcription of non-coding RNAs via transcriptional read-through subsequent to a loss of repression at repetitive DNA (Aporntewan et al., 2011; Faulkner et al., 2009; Slotkin and Martienssen, 2007). Transcription of repetitive elements is a feature of many cancers, and occurs frequently in brain tissue, with largely unknown consequences (Faulkner et al., 2009; Rangwala et al., 2009). Whether genome-wide hypomethylation represents an early or causative event in cancer or a passive consequence remains unknown (Wild and Flanagan, 2010). The detection of hypomethylation in cancer precursor lesions and normal adjacent tissue and the induction of cancers in animal models with experimentally induced hypomethylation suggest a causative role (Flatley et al., 2009; Gaudet et al., 2003; Suter et al., 2004; van Hoesel et al., 2012). However, the occurrence of many cancers in the absence of genome-wide hypomethylation, and the progression of hypomethylation with cancer stage suggest a passive role (Estecio et al., 2007; Wild and Flanagan, 2010).

### 3.1.3 Repetitive element hypomethylation

Due to the long held view that genome-wide hypomethylation was associated with repetitive elements (Yang et al., 2004); numerous surrogate assays for genome-wide methylation have been developed specifically targeting consensus sequences for repetitive elements. These
include the long interspersed nuclear element 1 (LINE1), Alu elements (a family of short interspersed nuclear elements (SINEs), and Sat2, a satellite repeat element (Laird, 2010; Yang et al., 2004). LINE1 (hereafter referred to as ‘L1’) is the only autonomous (capable of independent retrotransposition (transposition via an RNA intermediate)), and most highly expressed TE in the human genome, comprising ~17% of the human genome, with over 500,000 copies (Beck et al., 2010; Cordaux and Batzer, 2009; Wilson et al., 2007). A full length L1 element is ~6kb long with a bi-directional, non-canonical promoter and two open reading frames (ORFs) coding for an endonuclease and retrotransposition machinery proteins (Beck et al., 2010; Cordaux and Batzer, 2009; Wilson et al., 2007). L1 transcription is largely regulated by DNA methylation of the 5’ promoter; however, the majority of L1 elements are truncated and cannot be transcribed (Faulkner et al., 2009; Rangwala et al., 2009; Slotkin and Martienssen, 2007). Less than 100 L1 elements are functionally capable of retrotransposition, only a few of which contribute to the vast majority of retrotransposition events (Beck et al., 2010). Alu elements, of which there are multiple families, are the most common TE in the human genome, with ~1.1 million copies, comprising roughly 11% of the genome (Cordaux and Batzer, 2009; Wilson et al., 2007). Alu elements are non-autonomous and their transposition requires the transposition machinery of L1(Wilson et al., 2007). Satellite repeats (including Sat2) are short tandemly repeated non-coding DNA, frequently in centromeric and heterochromatic regions of chromosome 1 (Wu et al., 2012a). Both L1 and Alu elements are heavily methylated in normal somatic tissue, however, hypomethylation of both, especially L1, is often detectable in tumours (Estecio et al., 2007; Figueiredo et al., 2009; Suter et al., 2004). Whereas transcription of TEs are required for their transposition, transposition-independent consequences of transcription may have functional consequences, as an estimated 7% of the human transcriptome is derived from transcription start sites (TSS) within L1 elements (Faulkner et al., 2009) and hypomethylation induced transcription of L1 elements within host gene introns can effect host gene transcription.
through RNA interference (Aporntewan et al., 2011) and through driving host gene ectopic expression from the L1 antisense promoter (Wolff et al., 2010).

3.1.4 Methods for Investigating Genome-Wide Methylation

Several methods have been designed to investigate genome-wide DNA methylation. Some methods capture overall genomic DNA 5-methyl-cytosine (5meC) content, using 5meC-specific antibodies (Weber et al., 2005), methyl-acceptance (Pufulete et al., 2003) and High Performance Liquid Chromatography (HPLC), often combined with mass spectrometry (MS) (Choi et al., 2009). These methods provide accurate measurement of whole-genome methylation, but with no information about the spatial arrangement or genomic location of DNA methylation, and require large amounts of input DNA, making them unsuitable for many population studies using precious (such as prospectively collected) patient samples (Brennan and Flanagan, 2013). Widespread DNA methylation patterns may be measured using methylation sensitive restriction enzymes with relatively high resolution; however, methylation analysis is biased towards regions of high CpG density, as the enzyme restriction sites occur at sequences such as CCGG, and CGCG for the HpaII and HhaI restriction enzymes, respectively (Laird, 2010; Xu et al., 2012a). Many other genome-wide methylation assays are based on bisulphite sequencing (Beck and Rakyan, 2008; Rakyan et al., 2011), however, real-time PCR, restriction enzyme based (combined bisulphite restriction analysis (COBRA)(Hsiung et al., 2007)), and microarray-based methods (Beck and Rakyan, 2008) are also used. Most popular and convenient for population studies is the measurement of methylation at repetitive elements such as L1, Alu and Sat2, which are distributed at high frequency throughout the human genome (Wilson et al., 2007; Yang et al., 2004). These are considered ‘surrogate’ measures of genome-wide DNA methylation as their methylation is thought to reflect genome-wide methylation levels (Yang et al., 2006).
However, the efficacy of these surrogate assays for genome-wide methylation has been questioned (Brennan et al., 2012a; Nelson et al., 2011), and a re-interpretation of the results reported using these assays is warranted.

In order to assess the current evidence for association of genome-wide DNA methylation with cancer risk, and to identify the most promising research avenues and methods for development of cancer risk biomarkers, we conducted a meta-analysis of all available studies investigating blood genome-wide DNA methylation in relation to cancer risk.
3.2 Methods

3.2.1 Search strategy

Eligible studies included prospective and retrospective case-control studies published between 01/01/2000 and 08/06/2012 that investigated genomic DNA methylation in blood, and its potential association with either cancer incidence or prevalence (Details provided in appendix table1). Studies using any quantitative measure of genome-wide DNA methylation, or any surrogate measurement thereof, were included. Required for meta-analysis was the reporting of categorical analysis of the relationship between DNA methylation and the cancer outcome, including odds ratio (OR) and confidence intervals. Cancer types included malignant disease and colorectal adenoma, a non-malignant colorectal cancer precursor. Controls included individuals without cancer, with variable control selection criteria. Literary searches were performed using Pubmed, by Brennan K. B.Sc and Flanagan J.M. PhD. Pubmed Search terms used were ‘LINE1 blood risk’, ‘DNA methylation blood risk’, ‘methylation blood cancer’ and ‘hypomethylation blood cancer’. Reference lists from all included studies and relevant reviews were screened for additional studies, but yielded no additional reports. No unreported studies or reports in languages other than English were identified, therefore all relevant studies, both included and excluded, are represented appendix table1.

3.2.1.1 Excluded studies

All relevant studies found are included in Appendix table1. Inclusion criteria were kept as broad as possible in order to avoid selection bias. Six studies (Appendix table1) were excluded due to non-reporting of categorical analysis only.

3.2.1.2 Data abstraction
All data included in meta-analyses (except for our own unpublished data), including effect estimates and study details, was abstracted directly from the reports referenced in Appendix table 1. Inverse odds ratios were calculated for studies reporting odds ratios for the highest methylation category compared with the lowest (reference) category, and standard errors (SE) for ORs were calculated from confidence intervals (CI) using the formula $SE = \frac{(\log(upper\ CI) - \log(lower\ CI))}{(2*1.96)}$. ORs for L1 studies not reporting categorical analysis was performed by inference based on the sample size, mean methylation and standard deviation. While an ideal meta-analysis would extract raw data and perform the same analysis on all data, this was not possible due to raw data not being available for most published studies.

3.2.1.3 Inclusion of unreported data

The only included studies that were not publicly available were those conducted by ourselves, for which categorical analysis was conducted and included. Study populations and laboratory methods were previously reported (Brennan et al., 2012a). LINE1 methylation, measured by bisulphite pyrosequencing, was split into quartiles based on control methylation distribution. Unconditional logistic regression was used to determine the OR for disease, comparing the lowest quartile to the highest (reference) category.

3.2.2 Meta-analysis

Meta-analysis was carried out using the R statistical program. The 'meta.summaries' command within the 'rmeta' R package was used to generate average summary estimates, weighted by sample size, using random-effects models due to significant inter-study effect heterogeneity, as indicated by Woolf’s test for heterogeneity (p<0.05).
3.2.3 Funnel plot

A funnel plot for publication bias was generated using the ‘funnel’ command within the ‘meta’ R package, with standard error on the Y axis.

3.3 Results

3.3.1 Meta-analysis 01/01/2000-08/06/2012

Twenty three publications reported population-based cancer case-control studies investigating blood genomic DNA methylation in relation to cancer risk (Appendix table 1). It is important to note that the comparability of these studies is limited by a wide range of differences between them; however, such differences may also help to determine the characteristics of genome-wide hypomethylation that are potentially associated with cancer risk. Major variables between these reports included cancer type, assay used, study design (prospective/retrospective), sample size, sex (male/female/mixed), ethnicity and analytical/statistical methods. Furthermore, populations at different cancer risk are included, for instance one report included elderly men at high risk (Zhu et al., 2011), whereas another included Asian women who are at lower risk of breast cancer (Cash et al., 2012). Some of these reports included more than one “study”, due to use of different assays, different populations or different study designs, so altogether there were 34 individual studies of genome-wide methylation and cancer risk (Appendix table 1). The greatest limitation to comparison between these studies was reporting of data analyses. Some studies compared average DNA methylation between cases and controls only at the mean or median level, whereas most studies also included categorical analysis, generating an OR for disease for individuals in the lowest methylation category compared with the highest (reference) category. The OR for disease within the risk-associated methylation category was the most consistently reported, comparable, and representative (of overall results) factor reported between studies, and was, therefore used for comparison between studies. Therefore, eight
studies that did not report categorical analyses and indeed were not significantly different between cases and controls, could not be included resulting in an over-estimate of any positive effects (Cho et al., 2010; Choi et al., 2009; Patchsung et al., 2012; Pobsook et al., 2011; Widschwendter et al., 2008; Zhu et al., 2011) (Appendix table1). We have also included categorical analyses of L1 methylation in two of our own population-based prospective case-control studies for breast cancer risk (chapter 2), for which categorical analysis was not previously reported. As four studies (three reports) (Liao et al., 2011; Lim et al., 2008; Xu et al., 2012b) used the lowest, rather than the highest methylation category as the reference category, inverse ORs were calculated for these studies. Importantly, the methylation ‘split’ (tertile, quartile, quintile or decile), used for categorical analysis varied between reports, and is a potential effect modifier, however, this cannot be easily corrected without obtaining the raw data for each of the studies. A meta-analysis of many of the relevant studies was recently reported by Woo et al (Woo and Kim, 2012), which did not include five relevant, recent and predominantly null reports (Brennan et al., 2012b; Di et al., 2011; Mirabello et al., 2010; Wu et al., 2012b; Xu et al., 2012b). Therefore, this analysis represents a revised meta-analysis including all reported studies.

The overall summary estimate OR for all genomic DNA methylation studies using a random effects model was OR=1.4 (95% CI= 0.9-1.9), suggesting no overall association (Fig 3.1). Summary ORs for studies using L1 (OR=1.24, 95% CI=0.76-1.72), Alu (OR=1.31, 95% CI=0.93-1.68) and Sat2 repetitive elements (OR=1.55, 95% CI=0.99-2.10) were not significantly associated with cancer risk. Significant heterogeneity between studies was evident (p=<0.001), likely due to the use of different assays and investigation of different diseases, including both cancers and CRA (Lim et al., 2008; Pufulete et al., 2003).
Figure 3.1. Meta Analysis of Studies investigating Genome-wide DNA methylation in peripheral blood DNA for cancer risk (studies published between 01/01/2000 and 08/06/2012). Test for heterogeneity showed highly significant heterogeneity across all studies (p<0.001), and specifically in the analysis of 5meC (p<0.001) and LINE1 (p<0.001), but not significant for Alu (p=0.121) or Sat2 (p=0.827). Random Effects (RE) model was used for all summary analyses. CRC=colorectal cancer, CRA=colorectal adenoma, BGS=Breakthrough Generations Study, EPIC=European Prospective Investigation into Cancer and Nutrition. Study reference numbers correspond to appendix Table 1.
We investigated whether there was evidence of publication bias among studies investigating L1 methylation by generating a funnel plot (Stroup et al., 2000; Thornton and Lee, 2000) using all available studies (Fig 3.2). We did not observe funnel-plot asymmetry (evidence of publication bias); however, the ability of this plot to indicate publication bias may be limited by the small number of studies ($n = 14$), and variable direction of effect in studies showing significant associations. All of the included reports showing significant association between L1 methylation and cancer included investigations of L1 methylation only; however, all reports showing negative results for L1 also included studies showing significant associations between another methylation marker and cancer incidence/prevalence, consistent with a bias toward publication of significant associations. There appears to be a trend toward lower effect size with later publication date, as OR is significantly correlated with publication year among L1 methylation studies ($r = 0.49$, $P = 0.05$). This appears to be independent of sample size, as the correlations between OR and sample size ($r = -0.19$), and sample size and year of publication ($r = 0.2$), are not significant. Inclusion of all relevant reports is critical to the accuracy of meta-analyses (Thornton and Lee, 2000), therefore, publication and reporting biases may affect our ability to identify the true effect size.
Figure 3.2 Funnel plot for potential publication bias in LINE1 reports. Log Odds Ratio is shown on the X-axis, with Standard Error on the Y-axis and summary effect size and 95% confidence intervals are shown as dotted vertical lines. Studies are shown with shape representing the categorical split reported in the study, size representing the size of the study and open shapes representing retrospective studies and filled shapes representing prospective studies. The light grey area represents 99% confidence limits and dark grey represents the 95% confidence limits.
A reporting bias exists whereby categorical analysis of methylation-risk associations tends only to be reported if it reveals statistically significant results, and if the authors deem categorical analysis appropriate. In order to address this bias, we conducted a meta-analysis of all published L1 studies (01/01/2000-08/06/2012), including estimated ORs for those studies that did not include categorical analysis, by inferring OR based on sample size, mean methylation and standard deviation of methylation. This was in addition to inclusion of our own L1 studies, for which we had not previously reported categorical analysis (Brennan et al., 2012a). Meta-analysis of all L1 studies shows that blood L1 methylation is not associated with cancer risk (summary OR= 1.10, 95% CI=0.92-1.32), and there was significant heterogeneity across studies (p<0.001) (Fig 3.3).
Figure 3.3. Meta Analysis of Studies investigating LINE1 DNA methylation in peripheral blood DNA for cancer risk. Test for heterogeneity showed highly significant heterogeneity across all studies (p<0.001), and therefore random effects (RE) model was used for summary analyses. Four additional studies that showed no significant difference between cases and controls and did not report categorical analysis have been included showing an estimated odds ratio of 1 and standard error estimated from sample sizes.
3.3.2 Update meta-analysis (08/06/2012-05/09/2013)

Since compilation of the above meta-analysis, 8 publications have reported 11 studies of genome-wide DNA methylation of cancer risk (Appendix table1). An update meta-analysis was conducted for all studies reporting categorical analysis published within this time. Of 11 studies overall, only seven reported ORs for categorical analysis of the association between DNA methylation and cancer risk, including studies using HPLC, methyl-acceptance assay, LUMA, MethyLight methylation analysis of ALU, Sat1, and L1, and pyrosequencing of L1. Consistent with the previous meta-analysis, there was no overall association between DNA methylation within the lowest methylation category and cancer risk (OR=0.82, 95% CI=0.56-1.19) (Fig 3.4). Furthermore, there was significant heterogeneity variance between study estimates (Woolf’s test of heterogeneity) (p=0.003).
Figure 3.4. Update meta Analysis of Studies investigating Genome-wide DNA methylation in peripheral blood DNA for cancer risk (studies published between 08/06/2012 and 05/09/2013). Test for heterogeneity showed highly significant heterogeneity across all studies (p=0.003) Random Effects (RE) model was used for all summary analyses. Reference numbers for studies correspond to those shown in Appendix Table 1.
3.4 Discussion

According to the STROBE-ME (Strengthening the reporting of observational studies in epidemiology-molecular epidemiology) guidelines, all basic statistics of a biomarker measure distribution (mean, median, range and variance) and details of all other analyses should be reported in studies investigating molecular biomarkers for disease risk (Gallo et al., 2011). However, many reports do not include these basic statistics making comparison between studies difficult, and making thorough meta-analysis of the collective evidence impossible.

Overall, we did not find a significant association between genome-wide DNA methylation and cancer risk, either in a meta-analysis of 28 studies published between 01/01/2000 and 08/06/2012, or in an update meta-analysis of 11 studies published since then. Therefore, that genome-wide hypomethylation is detectable in blood of cancer patients is not supported by the data overall. Consistently, two studies reporting hypomethylation detectable in tumour and normal adjacent tissue of bladder (Wolff et al., 2010) and colon (Suter et al., 2004) failed to detect hypomethylation in blood in the same sample sets (Suter et al., 2004; Wolff et al., 2010), suggesting that hypomethylation may be restricted to the disease affected tissue. Indeed, a number of studies have indicated that methylation of repetitive elements may be tissue specific, most pronounced in tumour, and not correlated between tumour and blood (Cho et al., 2010; Piyathilake et al., 2011; van Bemmel et al., 2012). Furthermore, ~36% of L1 elements display tissue-specific expression (Faulkner et al., 2009).

The current findings are inconsistent with a previous meta-analysis (Woo and Kim, 2012) which reported a significant inverse association between both genome-wide DNA methylation and cancer risk overall, as well as a significant inverse association between L1 methylation an cancer risk. The current analysis, however, included five additional studies (Brennan et al., 2012b; Di et al., 2011; Mirabello et al., 2010; Wu et al., 2012b; Xu et al.,
2012b) within the primary meta-analysis, and seven additional studies within the update meta-analysis, therefore representing an updated revision of the evidence.

There is great inconsistency between findings reported by different studies, as is evident from the highly significant heterogeneity variance between effect estimates between studies found in the primary meta-analysis, L1 study analysis, and update meta-analysis. Many factors may account for this variability, as follows.

3.4.1 Cancer heterogeneity

One consideration is that heterogeneity between cancer types may affect the overall findings from this meta-analysis. For example, in tumour tissue, L1 hypomethylation frequently occurs in cancers displaying chromosome instability, but rarely in cancers with microsatellite instability (Estecio et al., 2007; Matsuzaki et al., 2005; Ogino et al., 2008). Consistently, L1 hypomethylation shows considerable inter-individual variability in colon cancer, and is frequently undetectable in tumours (Estecio et al., 2007). Therefore, it might be expected that DNA methylation variability in blood may predict risk of some, but not all cancers. In breast cancer, however, two studies (Choi et al., 2009; Xu et al., 2012b) investigating blood L1 methylation failed to find any associations, consistent with our own studies.

3.4.2 Methylation measurement/assay-type

There appears to be some evidence for association of genome-wide DNA hypomethylation and cancer prevalence within studies using measures of 5meC content, as our primary
meta-analysis (studies published between 01/01/2000 and 08/06/2012) showed a significant
association with cancer risk overall. Furthermore, all five studies included reported significant
associations between hypomethylation and cancer prevalence in categorical analysis, four of
which also found significant association at the mean level. However, of four studies
published since 08/06/2012, only two showed evidence of association between 5meC
content and cancer risk (Friso et al., 2013), whereas two studies (Huang et al., 2012; Tahara
et al., 2013), including the first prospective investigation (Huang et al., 2012) found no
association. A recent study (Friso et al., 2013), which could not be included in our analysis
due to incomplete reporting of categorical analysis statistics, reported very strong
association between HPLC-measured hypomethylation and gastric cancer risk using inter-
quartile analysis, yet the ORs and 95% CIs were not reported. Instead, ORs for cancer risk
within the retrospective study (OR=45.9, 95% CI= 14.3-147.6), and prospective study
(OR=36, 95% CI=3.9-329.9) were calculated using categorical analysis with subjects
dichotomised at methylation cut-offs calculated to best discriminate cases from controls
using ROC curves. As the proportions of cases and controls within each methylation quartile
were reported for the retrospective study (but not the prospective study) the OR for
individuals within the lowest methylation quartile (n=64) compared with the highest quartile
(n=21) was 19.6, however confidence intervals or standard error could not be calculated
based on reported data. Overall, it appears likely that 5meC content is reduced in blood of
cancer patients; however this may not be evident in all cancers and populations, and may
not be reproducible in prospective studies, therefore, 5meC loss may be a marker of disease
prevalence rather than risk.

Two studies to date have investigated genome-wide DNA methylation in blood of breast
cancer patients using the LUMA method, with one study (Xu et al., 2012b) reporting a
strong protective effect of genomic DNA hypomethylation on breast cancer risk; however, a more recent study (Wu et al., 2012a) failed to replicate this. The identification of hypermethylation, rather than hypomethylation associated with cancer may be due to the measurements of methylation at CCGG restriction sites, which are enriched at CpG islands. These regions tend to be unmethylated in normal tissue, meaning that the only change detectable would be an increase in methylation (Xu et al., 2012b).

Consistent with our own research findings (chapter 2), there was no overall association between blood L1 methylation and cancer. Whereas early studies suggested an association of blood L1 hypomethylation with cancer, later studies have failed to replicate this. Experimental improvements and advances in larger epigenetic epidemiological studies may account for this. Furthermore, the direction of effect on cancer prevalence of L1 hypomethylation was inconsistent between studies (Liao et al., 2011), suggesting that these associations may have occurred by chance.

Methylation of TEs, particularly L1, is often reported as genome-wide or ‘global’ DNA methylation. This is confusing and inappropriate as L1 pyrosequencing measures methylation at only three to four CpG sites within the promoters of a pool of L1 elements (Yang et al., 2004), which cannot be considered representative of genome-wide methylation (Nelson et al., 2011). The detection of genome-wide hypomethylation by these assays in tumour DNA is likely due the occurrence of a proportion of repetitive elements within hypomethylated domains, rather than to a specific enrichment of hypomethylation at repetitive elements (Hansen et al., 2011; Hon et al., 2012). Furthermore, whereas an early study reported that L1 and Alu methylation were correlated with 5meC levels in blood, a more recent report was unable to validate this (Choi et al., 2009). This lack of correlation is supported by the stronger association with cancer risk of blood methylation detected with
5meC measures than with surrogate assays such as L1. Two prospective cancer risk studies detected blood hypomethylation at Alu (Gao et al., 2012) and Sat2 (Wu et al., 2012b), but not at L1 in the same samples. This is inconsistent with these assays detecting "genome-wide methylation", but suggests that hypomethylation may be restricted to specific genomic sequences. Technical biases, such as the preferential amplification of different elements due to PCR annealing temperature may be a problem for primers with multiple binding sites (El-Maarri et al., 2011), as L1 elements are differentially methylated at different genomic loci (Phokaew et al., 2008; Singer et al., 2012; Wolff et al., 2010). The precise selection of CpG sites within the L1 consensus sequence is also important, as we and others (Piyathilake et al., 2011) have noticed that L1 methylation varies significantly between CpG sites, a factor that may explain some of the disparities between studies. L1 sequence heterogeneity poses yet another bias, as this may cause cellular and allelic heterogeneity in L1 methylation (Burden et al., 2005), and may lead to underestimation of methylation levels (Yang et al., 2006). Differential methylation of individual L1 elements may yet be important, as a recent study showed that potentially pathogenic hypomethylation of an individual L1 within gene introns was much more pronounced than hypomethylation of global L1 elements, and was detectable in normal tissue, where global L1 hypomethylation was not (Wolff et al., 2010). Furthermore, hypomethylation of a specific L1 element, LRE1 was detected in blood of head and neck cancer patients (Hsiung et al., 2007), suggesting that methylation analysis of individual TEs may yet be useful as biomarkers of cancer risk.

3.4.3 Prospective versus retrospective studies

Of four prospective studies investigating L1 methylation and cancer risk that were included in the primary meta-analysis, only one (Zhu et al., 2011) identified an association between low
L1 methylation and cancer incidence, however, small sample size (n=30 cases, 487 controls), combining of multiple cancer types, and selection of a very high-risk population (elderly males) may all be confounding this finding. A recent study (Li et al., 2013), which showed statistically significant L1 hypomethylation (pyrosequencing) in cancer patients (mixed cancers), could not be included in the meta-analysis due to non-reporting of categorical analysis. This association, however, was not adjusted for age, smoking and occupational exposures, all of which were also associated with L1 methylation and cancer risk. The four most recent, relatively large studies, have failed to identify any such associations (Brennan et al., 2012b; Cash et al., 2011; Wu et al., 2012b). Therefore, the current evidence suggests that L1 methylation is not a risk factor for cancer. Of two prospective studies to date investigating 5meC (measured by HPLC) in relation to cancer risk, one small study identified a significant association with gastric cancer risk (Friso et al., 2013), whereas a larger study (Huang et al., 2012) found no association with colorectal cancer.

3.4.4 Other study design factors

Differences between study designs tend to contribute to heterogeneity and bias in meta-analyses (Stroup et al., 2000; Thornton and Lee, 2000). These analyses included studies of different study designs, included familial studies, hospital-based studies and nested case-control studies. Sample size is another likely bias, and varies largely between the reports, including two studies with very small sample size of less than 40 cases (Pufulete et al., 2003; Zhu et al., 2011). Small sample sizes are only appropriate where the difference in methylation between groups is large, as such studies do not have the power to detect subtle methylation variability, perhaps leading to an interpretation bias towards non-significance, or
to chance detection (Rakyan et al., 2011). Control for potential confounding factors, including any factor that may influence both exposure and outcome independently, is a basic requirement for all molecular epidemiological studies, and may be achieved by matching of cases with controls based on potential confounding factors, or by adjustment within statistical models. Several cancer risk factors, including age (Christensen et al., 2009; Rakyan et al., 2010), gender (Cash et al., 2012; Wilhelm et al., 2010) (El-Maarri et al., 2011), ethnicity (Cash et al., 2012; Hou et al., 2010; Zhang et al., 2011a) (Fraser et al., 2012; Zhang et al., 2011a) and environmental carcinogen exposures (Bollati et al., 2007; Cash et al., 2011) are reported to influence genome-wide DNA methylation patterns. Therefore, overall findings of included studies may depend on the influence of study-specific confounding factors, and on the degree to which they were controlled for.

3.4.5 Statistical factors

Publication bias, whereby studies reporting significant results are more likely to be published, is one of the most significant potential biases in meta-analysis of scientific studies (Thornton and Lee, 2000). A funnel plot of studies investigating L1 methylation in relation to cancer risk did not show evidence of publication bias; however publication of negative L1 studies only within papers reporting other significant findings suggest publication bias. Furthermore, as categorical analysis is generally only included in reports if it shows significant results, an over-estimation of the true summary effect size in meta-analysis is likely.

A prominent bias between studies was the use of different numbers of categories for categorical analysis, which included tertiles, quartiles, quintiles and deciles. Categorical
analysis may not have been appropriate for these studies, especially with use of larger numbers of quantiles, as the narrow ranges of DNA methylation reported, especially at L1, would mean that the difference in percentage methylation between categories would be far below the detection sensitivity of the assay. for instance the technical variation for pyrosequencing is around 2-3%, and we observed an intra-class correlation coefficient (ICC) for LINE1 in blinded duplicate samples of 0 (95% CI 0-0.61) which suggests higher within individual variability than between individual variability (chapter 2).

3.4.6 Conclusions

We conclude from our meta-analysis that that genome-wide DNA methylation, as measured by surrogate repetitive element assays, is not associated with cancer risk, demonstrated most appropriately by several prospective cohort studies. In contrast, associations of total methylation levels observed with HPLC and LUMA based measurements appear to be associated with cancer risk, but remain to be validated.
CHAPTER 4

Investigation of Factors Influencing ATM Methylation
4.1 Introduction

Many studies have identified significant association between DNA methylation and diseases; however, few have sought to determine the mechanism, or potential factors underlying the association. Our research indicated that ATMmvp2a hypermethylation is associated with breast cancer risk, however whether ATM hypermethylation represented an innate epimutation-like risk factor, or an intermediate marker of other breast cancer risk factors, such as age, BMI, or dietary factors, was not clear.

DNA hypermethylation associated with epimutations and repeat instability disorders may be caused by underlying genetic variability, however it is epigenetic silencing of gene expression that confers disease risk (Dolinoy et al., 2006; Hesson et al., 2010; Hitchins and Ward, 2009). The stronger association between ATMmvp2a methylation and cancer risk in the familial KConFab study than the sporadic studies suggests underlying genetic factors, consistent with the genetically mediated heritability of aberrant epigenetic patterns associated with cancer risk (Hitchins et al., 2011; Ligtenberg et al., 2009).

Conversely, epigenetic epidemiological research postulates that epigenetic mechanisms mediate aberrant gene expression through alteration by age and environmental risk factors (Bollati et al., 2007; Brennan and Flanagan, 2012a; Christensen and Marsit, 2011; Shenker et al., 2013). This is supported for breast cancer by the finding that environment-linked metabolic diseases associated with increased breast cancer risk, including diabetes and obesity, are associated with epigenetic changes (Bell et al., 2010; Cheraghi et al., 2012; Milagro et al., 2011; Xu et al., 2013a). The association of ATMmvp2a methylation with age in control supports the idea of non-genetic modifiers of ATMmvp2a methylation; however the temporal stability of ATMmvp2a methylation between time-points ~6 years apart suggests that methylation is not strongly altered by age or environmental factors.
Lastly, in the case of agouti mice, genetic factors (the presence of an intracisternal A particle retrotransposon within the agouti gene) and environmental factors (maternal dietary folate insufficiency) are required to bring about the agouti phenotype associated with metabolic disease and cancer risk (Dolinoy et al., 2006), therefore complex genetic-epigenetic-environmental interactions may contribute to disease susceptibility.

In order to determine the factors driving the association, we investigated the effect on ATMmvp2a methylation of genetic and environmental factors.
4.2 Materials and Methods.

4.2.1 Subjects

The KConfab study population, described in chapter 2 (materials and methods, study populations) was used to investigate the potential influence of haplotype on DNA methylation. The study consisted of 166 blood samples collected prospectively from breast cancer cases within high-risk families, and 225 samples collected from healthy unrelated control at the same time. Cases were invasive breast cancers collected at a median 45 months (range 1 – 140) prior to diagnosis.

The EPIC study was designed to enable prospective investigation of dietary and metabolic factors in relation to cancer risk (Riboli and Kaaks, 1997), and was therefore used to assess the effect of serum metabolites, fasting status, BMI, age, and MTHFRc677t genotype on ATMmvp2a methylation. Five EPIC sub-groups (labelled EPIC1, EPIC2, EPIC3, EPIC4 and EPIC5 (table 4.1)) were included in the analyses of ATMmvp2a methylation. These included two healthy control populations for which serum metabolite data for 24 metabolites related to one carbon metabolism, inflammation, and tryptophan metabolism were measured. Metabolites were measured by Bevital A/S (www.bevital.no) using liquid chromatography and gas chromatography with tandem-mass spectrometry, and microbial assays to measure folate and cobalamine (Johansson et al., 2010). Three prospective breast cancer case-control sub-groups were used, including one sub-group (EPIC3, n=277 cases, 282 controls) that was included in the investigation of ATMmvp2a methylation in relation to breast cancer risk (chapter 2), and two new sub-groups (EPIC4 (n=92 cases, 92 controls) (Shenker et al., 2012) and EPIC5 (n=109 cases, 119 controls) (van Veldhoven et al., unpublished data)) (Table 4.1). 4 cases and 8 controls within the EPIC4 subgroup were removed from analysis due to sample overlap with the EPIC5 subgroup. These samples were removed from EPIC4
rather than EPIC5 prior to analysis, as EPIC4 samples had been bisulphite converted ~1 year before EPIC5 samples, therefore EPIC5 samples are likely to be of higher quality according to the Zymo-research bisulphite conversion kit manual (EPIC4 data in table 4.1 shown for included samples only). Methylation values from EPIC4 were removed, rather than calculating average methylation values for methylation analysis within EPIC4 and EPIC5 subgroups in order to avoid introducing batch-effects. Microarray-based investigation of fasting on DNA methylation was carried out using all samples from EPIC5, and an addition 104 samples, including 49 controls and 55 cases (see table 4.1(B)).

Fasting-status was self-reported as part of a standardised dietary questionnaire completed at the time of blood draw. Fasting time (time since last meal to blood draw) was available for 98 subjects with median length of time for fasting individuals 12.8 (IQR 12-13.7) hours compared with 2.4 (IQR=1.5-3.2) hours in non-fasting individuals.
Table 4.1 Characteristics of EPIC study subgroups used in chapter 4

4.1 (A) Characteristics of EPIC study participants included in analysis of ATMvmp2a DNA methylation

Control studies

<table>
<thead>
<tr>
<th>Study</th>
<th>EPIC1</th>
<th>EPIC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fasting status</td>
<td>Non-fasting</td>
<td>Fasting</td>
</tr>
<tr>
<td>N</td>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td>Age (median (IQR))</td>
<td>56 (53.6-56.1)</td>
<td>56.2 (53.3-59.1)</td>
</tr>
</tbody>
</table>

Prospective case control breast cancer studies

<table>
<thead>
<tr>
<th>Study</th>
<th>EPIC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-control</td>
<td>Case</td>
</tr>
<tr>
<td>Fasting status</td>
<td>Non-fasting</td>
</tr>
<tr>
<td>N</td>
<td>151</td>
</tr>
<tr>
<td>Age (median (IQR))</td>
<td>56.1 (47.1-63.2)</td>
</tr>
<tr>
<td>BMI (median (IQR))</td>
<td>25.36 (22.66-29.00)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>EPIC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-control</td>
<td>Case</td>
</tr>
<tr>
<td>Fasting status</td>
<td>Non-fasting</td>
</tr>
<tr>
<td>N</td>
<td>51</td>
</tr>
<tr>
<td>Age (median (IQR))</td>
<td>51.04 (44.2-55.3)</td>
</tr>
<tr>
<td>BMI (median (IQR))</td>
<td>23.78 (21.95-26.85)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>EPIC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-control</td>
<td>Case</td>
</tr>
<tr>
<td>Fasting status</td>
<td>Non-fasting</td>
</tr>
<tr>
<td>N</td>
<td>23</td>
</tr>
<tr>
<td>Age (median (IQR))</td>
<td>54.5 (50.2-57.6)</td>
</tr>
<tr>
<td>BMI (median (IQR))</td>
<td>24.42 (22.32-25.91)</td>
</tr>
</tbody>
</table>
Table 4.1 (B) Characteristics of EPIC study participants included in EWAS study of DNA methylation and fasting

Discovery set analysis 1

<table>
<thead>
<tr>
<th>Study</th>
<th>EPIC5</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case-control</td>
<td>Case</td>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting status</td>
<td>Non-fasting</td>
<td>Fasting</td>
<td>Non-fasting</td>
<td>Fasting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>23</td>
<td>86</td>
<td>23</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (median (IQR))</td>
<td>54.37 (45.85-57.9)</td>
<td>54.70 (48.65-57.57)</td>
<td>55.2 (51.53-59.24)</td>
<td>54.23 (47.25-57.63)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI (median (IQR))</td>
<td>24.27 (22.45-25.99)</td>
<td>25.32 (22.56-27.62)</td>
<td>24.72 (21.40-26.54)</td>
<td>24.38 (22.34-26.56)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discovery set analysis 2

<table>
<thead>
<tr>
<th>Study</th>
<th>EPIC5/EPIC6</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case-control</td>
<td>Case</td>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting status</td>
<td>Non-fasting</td>
<td>Fasting</td>
<td>Non-fasting</td>
<td>Fasting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>23</td>
<td>86</td>
<td>30</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (median (IQR))</td>
<td>53.84 (46.37-58.17)</td>
<td>54.53 (48.49-57.75)</td>
<td>53.27 (47.06-58.28)</td>
<td>54.32 (47.68-57.75)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI (median (IQR))</td>
<td>24.51 (22.63-26.43)</td>
<td>25.25 (22.67-27.77)</td>
<td>24.64 (21.39-26.55)</td>
<td>24.28 (22.31-26.57)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Control studies (validation set)

<table>
<thead>
<tr>
<th>Study</th>
<th>EPIC7</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fasting status</td>
<td>Non-fasting</td>
<td>Fasting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>104</td>
<td>76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (median (IQR))</td>
<td>48.5 (42.75-54.00)</td>
<td>50 (44.75-55)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI (median (IQR))</td>
<td>23.71 (22.00-27.06)</td>
<td>26.31 (22.77-29.50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2.2 Laboratory methods

4.2.2.1 DNA extraction

DNA samples were extracted from whole blood using Qiagen DNA blood Mini Kits in KConFab. DNA was extracted from buffy coat using DNA Blood Mini Kits (Qiagen, UK) for all EPIC studies except EPIC2, for which DNA was extracted from whole blood using the same kit.

PCR and pyrosequencing for ATM and LINE1 was performed as described in chapter 2, materials and methods. PCR and pyrosequencing for MTHFR was carried out using a single round PCR with a biotinylated reverse primer. Single round semi-nested PCR was used to amplify TOX2, ASB10, ADARB2, PER3 and LGR6, with a common tag incorporated into either the forward or reverse primer, depending on the strand being sequenced, and a universal biotinylated primer, as described previously (Royo et al., 2007). Primers were designed using the Pyromark Assay Design software 2.0 (Qiagen). Primers are shown in table 4.2 (a). Pyrosequencing was carried out as described in chapter 2.

4.2.2.2 Acquisition of DNA from specific blood cell fractions.

DNA isolated from monocytes, B lymphocytes and T lymphocytes from two individuals were provided generated for a previous study (Flanagan et al., 2009b). Blood cell fractionation was carried out using antibody-coated microbeads and MACS MS separation columns (Miltenyi Biotech) following standard manufacturers protocol.
4.2.2.3 Genotyping

Genotyping of ATM to determine the haplotype was carried out by pyrosequencing in bisulphite converted DNA with assays for three SNPs within the ATM gene (rs228589, rs664677 and rs664982), allowing distinction of a putative risk haplotype from the reference haplotypes, and from a further three rare possible haplotypes (Rebbeck et al., 2011). DNA methylation of four CpG sites surrounding rs228589, and methylation of a polymorphic CpG site harbouring a SNP at the guanine (rs664982) was simultaneously measured within the same pyrosequencing runs. The allele frequency for a given SNP was determined by calculating the ratio of peak heights for each allele to reference peak heights for nearby non-polymorphic single nucleotides of the same base types, which represented the ‘expected’ peak heights for individuals homozygous for a given allele. Thresholds of less than 10% difference between observed and expected peak heights were set, so that a sample homozygous for an A allele within a SNP should have an A peak height of within 10% of the peak height of a reference A outside of a SNP, and an individual heterozygous for the A allele should display an A peak height of within 10% of half of the height of the reference A nucleotide. As the frequency of each allele within a SNP were calculated separately, agreement between the allele values was required to ‘pass’ a sample. All samples that were not called automatically we re-called by manually assessing individual peaks. Genotyping of all three SNPs was successful for 307 out of 399 samples. Genotypes were within the Hardy Weinberg Equilibrium (HWE p= 0.789). MTHFRc677t (rs1801133) genotype was measured using the same method, and genotype was within HWE (p=0.62). Primers are shown in table 4.2 (b).
Table 4.2 Primers designed for work within chapter 4

A. Primers for validation of fasting-associated loci from 450k analysis

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence (5’ – 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOX2 F</td>
<td>TTATGGTTGTATAGTATTTTATGTGGTTATG</td>
</tr>
<tr>
<td>TOX2 R</td>
<td>gacgggacaccgctgatctgtaAACTAAATCCAAATCTCTCTAA</td>
</tr>
<tr>
<td>TOX2 S</td>
<td>ATGGATATTGCTGTGGATT</td>
</tr>
<tr>
<td>HDAC4 F</td>
<td>gacgggacaccgctgatctgtaTTGGGAGGTTTGGGAAGTTAGA</td>
</tr>
<tr>
<td>HDAC4 R</td>
<td>ACCTCTTTTTCAACCAATCAA</td>
</tr>
<tr>
<td>HDAC4 S</td>
<td>AAATCTAAACCTACTCTC</td>
</tr>
<tr>
<td>ASB10 F</td>
<td>GTGGAATAGAAGAAAGAGAGAA</td>
</tr>
<tr>
<td>ASB10 R</td>
<td>gacgggacaccgctgatctgtaACTAAAACACCTCAAACCCATAAT</td>
</tr>
<tr>
<td>ASB10 S</td>
<td>AGAAGATGAGAAATATAGGT</td>
</tr>
<tr>
<td>PER3 F</td>
<td>ATGGGAGGAAAAATATTGAGAGG</td>
</tr>
<tr>
<td>PER3 R</td>
<td>gacgggacaccgctgatctgtaATCCCTTTCAACCTTTATTAAAAATTAC</td>
</tr>
<tr>
<td>PER3 S</td>
<td>GGAATTTTAGAAGGT</td>
</tr>
<tr>
<td>ADARB2 F</td>
<td>gacgggacaccgctgatctgtaGAGGAGGAGGAGGTTTGTGTAAGAAGAGAAG</td>
</tr>
<tr>
<td>ADARB2 R</td>
<td>CAACACCTATTCCTCAATCAA</td>
</tr>
<tr>
<td>ADARB2 S</td>
<td>CCCCCCRCRCRCACCA</td>
</tr>
<tr>
<td>LGR6 F</td>
<td>GTGATTTTTGGTGGTAGAAGTGGAAATATATATATATATATATATATATATAT</td>
</tr>
<tr>
<td>LGR6 R</td>
<td>gacgggacaccgctgatctgtaACCTTAATTCCTCCATTCTTAACC</td>
</tr>
<tr>
<td>LGR6 S</td>
<td>GGGGTATAGAGGAGG</td>
</tr>
</tbody>
</table>

B. Primers for genotyping of ATM

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence (5’ – 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATM rs228589 F</td>
<td>GTGGTTTTTTTGGTGGTTTTTGTGATAT</td>
</tr>
<tr>
<td>ATM rs228589 R</td>
<td>gacgggacaccgctgatctgtaACCTTTACACCTTTAATCTTTCTATC</td>
</tr>
<tr>
<td>ATM rs664677 S</td>
<td>ATCCAAAACTCTACC</td>
</tr>
<tr>
<td>ATM rs664677 F</td>
<td>AAGTTTTAAGGAATTTATAGGGTTATTT</td>
</tr>
<tr>
<td>ATM rs664677 R</td>
<td>gacgggacaccgctgatctgtaACCCTTTAATAACCCCCAAAAAAA</td>
</tr>
<tr>
<td>ATM rs664982 S</td>
<td>AGTATTTAGAAAAATTATAGGAAGG</td>
</tr>
<tr>
<td>ATM rs664982 F</td>
<td>ATGGTTATAGAGGTTTGGAGAT</td>
</tr>
<tr>
<td>ATM rs664982 R</td>
<td>gacgggacaccgctgatctgtaTCAACRTCTATAAAGGAAAAACCCACTCTACA</td>
</tr>
<tr>
<td>ATM rs664982 S</td>
<td>ATTTTTTTTTATATTTAATGTTG</td>
</tr>
</tbody>
</table>

*F=forward, R=reverse, S=sequencing

*Lower case letters represent universal tag sequence
4.2.2.3 Ex vivo culturing of PBMCs

PBMCs from six healthy females were extracted from blood using ficoll gradient, and immediately cultured for 24 hours in Dulbecco’s modified eagle medium (Sigma) containing different concentrations of glucose (1.0g/L, 0.5g/L, 0.25g/L and 0g/L), and 10% dialysed (glucose depleted) foetal calf serum. A ‘baseline’ PBMC sample from each individual was frozen overnight and processed at the same time as the cultured PBMC samples to control for the effect of in vitro culture on ATMmvp2a methylation. Glucose restriction was carried out in two batches of three participants, with three glucose concentrations (1g/L, 0.5g/L and 0g/L) used for all individuals, and an extra concentration (0.25g/L) for the second batch (participants D, E and F). Glucose concentrations were based on a standard serial dilution, where 1g/L represented the 100% glucose concentration recommended by the manufacturer for the cell culture medium. Cell proliferation and viability was determined using the CellTiter 96 Aqueous One solution (Promega). Cells were plated in triplicate at 1,000 cells per well of a 96-well plate. The CellTiter solution was added at 24 h using 30 uL/well and further incubated for 3 h at 37°C, and the absorbance was read at 490 nm to quantify the formazan product.

4.2.2.4 Microarray

Genome-wide investigation of the effect of fasting on DNA methylation was carried out using the Illumina Infinium HumanMethylation450 BeadChip microarray (450k array). A detailed description of this array has been published previously (Dedeurwaerder et al., 2011). The 450k array measured DNA methylation at 482,421 CpG sites and 3091 non-CpG sites genome wide, including sites within intragenic and intergenic regions as well as dense
coverage of CpG islands, island ‘shore’ and ‘shelf’ regions, imprinted regions, and known DMRs. DNA methylation is measured using two different assay types known as Infinium I and Infinium II assays. Both assays measure DNA methylation based on DNA sequencing of the methylation-dependent C>T polymorphism created at CpG sites on bisulphite conversion. The type I Infinium assay utilises two different probe types, one which binds to unmethylated CpG sites and the other which binds to methylated CpG sites, each of which is attached to a different type I bead type. Base extension is identical for unmethylated and methylated probe types, and the signal from each is read within the same colour channel.

The type II utilises a single probe type that binds to both unmethylated and methylated DNA due to mismatches at CpG sites within the probe sequence. This probe binds directly adjacent to the target cytosine, and differential incorporation of fluorescently labelled nucleotides occurs during base extension (green-labelled C or G nucleotides (depending on which strand is being measured) bind to unconverted sites, whereas red-labelled A or T nucleotides bind to converted sites). The methylated (green) and unmethylated (red) signals are generated in different colour channels. 500ng of each bisulphite-converted DNA was hybridized to the array following the Illumina Infinium HD methylation protocol. Microarray methylation values are represented as β-values, which represent ratios of the fluorescence intensity of the methylated bead type to the fluorescence intensity of the unmethylated bead type for a given probe, so that 0 equates to a fully unmethylated CpG site and 1 represents a fully methylated CpG site. Methylation values were background-subtracted using Genome studio from Illumina. Samples in which 20% of probes failed (probe detection p-value >0.05) were removed and probes that failed in 20% of samples were removed from analysis. After quality control, 484,804 probes were included in analysis. Methylation data for each sample was normalised using quantile normalisation. Peak-based correction was used to adjust for differences in methylation distribution between type I and type II probes (Dedeurwaerder et
al., 2011), using type I probes as the reference type. This is because measurements of type I probes have been shown to be more accurate and reproducible, and to display less technical variability, compared with type II probes (Dedeurwaerder et al., 2011). The ComBat package (Johnson et al., 2007) was used to adjust for batch effect between experimental batches (array chips).

4.2.3 Statistical analysis

Pyrosequencing Data analysis

The Wilcoxon Rank Sum test was used as a non-parametric test for difference in mean methylation between ATM haplotypes. Associations between DNA methylation and MTHFR genotype were conducted using linear regression models treating MTHFR genotype as a numerical variable when testing for a dose-effect association with the minor allele, and as a categorical variable when testing for differences in methylation between individual genotypes. A p-value for deviation from Hardy-Weinberg equilibrium was calculated using a non-parametric permutation Chi-squared test with 10000 permutations. Linear regression models were used to test for associations between ATMmvp2a methylation and serum metabolites, using univariate models and models adjusted for age, fasting status and MTHFR genotype as there was evidence that these factors affect ATMmvp2a DNA methylation, and as these factors may affect metabolic profiles. The association between ATM methylation and fasting was tested using non-parametric Wilcoxon Rank Sum test, and linear regression models adjusted for age, BMI, study centre, blood-donor status, and breast cancer case-control status were used. Logistic regression was used to generate odds ratios (ORs) for having below-median ATM methylation associated with fasting. A meta-analysis,
using effect estimates (ORs) and standard error for each study, was conducted, using a random-effects model weighted by study size. Two-tailed paired student’s T-tests were used to test differences in DNA methylation associated with glucose restriction, comparing mean methylation between conditions and paired for each participant. Unpaired two-tailed student’s T-tests were used to test for differences in cell proliferation in different glucose conditions for the cell proliferation (MTT) assay. DNA methylation was log-transformed prior to using parametric tests, in order to transform methylation to a more normal distribution, as indicated by histogram-plotting.

4.2.3.1 Microarray data analysis

Individual univariate Linear regression models were used to test if fasting status was associated with potential effect modifiers including age, breast cancer case-control status or EPIC study centre. As methylation was not associated with these factors, univariate linear regression models were used to investigate associations between DNA methylation and fasting status. Wilcoxon rank sum test was used as a non-parametric alternative model for testing associations between methylation and fasting status. FDR corrected p-values <0.05 were considered statistically significant.

Additional probe-filtration parameters used in this analysis included removal of probes harbouring SNPs within the CpG site at which DNA methylation is measured (polymorphic CpG sites), as DNA methylation at these sites is affected by genotype, which may be unevenly distributed between sample groups, and probes binding to >1 genomic locus in silico, as methylation at these sites may be affected by differential binding between samples (Price et al., 2013). Information for both of these factors is given for all probes in the 450k array in a recent ‘re-annotation’ of the 450k beadchip array (Price et al., 2013) (NCBI GEO
accession number [GSE:42409]). All statistical analysis was conducted using the R statistical program (version 2.15.1).
4.3 Results

4.3.1 ATM Haplotype

We first investigated if ATMmvp2a methylation was associated with genetic haplotype as a previous study had reported that a common ATM haplotype is associated with increased risk of breast cancer in BRCA1 mutation carriers (Rebbeck et al., 2011). Furthermore, recent reports have indicated that haplotype-associated DNA methylation is relatively common across the genome (Kerkel et al., 2008; Schalkwyk et al., 2010; Shoemaker et al., 2010). ATM haplotype was measured for cases (n=166), and controls (n=225) within the kConFab familial breast cancer study, in which ATM methylation was most strongly associated with breast cancer risk (OR=3.06 (95% CI 1.53-6.10) (chapter 2) (Brennan et al., 2012b). There were 89 individuals with the reference haplotype, defined as having the major allele for all three SNPs, 139 individuals heterozygous for all SNPs and 58 individuals homozygous for the minor alleles at all three SNPs (risk haplotype). ATMmvp2a methylation did not differ significantly between risk and reference haplotypes (p=0.29) (Fig4.1), or within BRCA1 mutation carriers only (data not shown). Furthermore, methylation of CpG sites surrounding rs228589 did not differ between genotypes (data not shown). Therefore, methylation variability at ATMmvp2a is unlikely to be driven by genetic factors in cis. As expected, due to the small numbers, neither haplotype, nor any of the three SNPs individually were significantly associated with case control status.
**Figure 4.1.** ATMmvp2a DNA methylation stratified by ATM haplotype. ATM Haplotype and ATMmvp2a methylation analysed in blood DNA samples from the the KConFab prospective breast cancer case-control study (n=166 cases, 225 controls). Pyrosequencing was used for methylation analysis and genotyping of three ATM SNPs (rs228589, rs664677 and rs664982) that were informative of haplotype. TTTTA: individuals homozygous for the a common 'reference' haplotype (n=89), AACCGG: individual homozygous for the allele putatively associated with breast cancer risk (Rebbeck et al, Cancer Res 2011) (n=58), TATCAG: Individuals heterozygous for reference and risk alleles (n=139).
4.3.2 MTHFRc77t genotype

DNA methylation can also be affected by genotype in trans, through alteration of OCM efficiency by genetic polymorphisms within genes encoding enzymes required for OCM, including a well-studied SNP within the methylenetetrahydrofolate reductase (MTHFR) gene, known as MTHFRc677t (rs1801133) (Friso et al., 2013). To investigate the potential interaction of MTHFRc677t genotype with ATMv2p2a methylation and one carbon metabolism, rs1801133 was genotyped in all samples from EPIC1, EPIC2, EPIC4, and EPIC5 sub-groups (n= 615 subjects combined). In controls, ATMv2p2a methylation was significantly increased in individuals with genotype CC (n=124) compared with combined genotypes CT and TT (n=296, p=0.003), after adjusting for age and fasting status. There was a significant positive linear relationship between ATMv2p2a methylation and presence of the C allele in controls (p=0.04) (fig4.2). In cases (EPIC4 and EPIC5 only), ATM methylation was not different between genotype CC (n=76) and genotypes CT and TT (n=123, p=0.92), and there was not a linear association between ATMv2p2a methylation and MTHFR genotype (p=0.863). The association in controls was mainly driven by a very strong association in the EPIC1 and EPIC2 subgroups (table 4.3).
Figure 4.2. ATMmvp2a DNA methylation stratified by MTHFRc677t genotype (both measured using pyrosequencing) in blood DNA samples of (A) healthy female controls and (B) prospectively collected breast cancer cases, within four combined subgroups of the European Prospective Investigation into nutrition and Cancer (EPIC) study. DNA methylation values from all subgroups were combined as Z-scores (Y-axes). Numbers of individuals with each genotype were (control) TT n=78, CT n=218, CC n=124, (cases) TT n=34, CT n=98, CC n=76. ** p-value <0.01 (logistic regression) for a difference in methylation between genotypes CC and combined genotypes CT and CT in controls.
4.3.3 Serum Metabolites

ATMmvp2a methylation was analysed in PBMC DNA from EPIC1 (n=72) and EPIC2 (n=146) for which 24 serum metabolite levels had been measured (Table 4.3). These included 15 metabolites involved in one carbon metabolism (OCM), due to the known limiting role of OCM on DNA methylation genome wide (Chiacchiera et al., 2013), neopterin, and 9 tryptophan metabolites, due to their roles in ageing and immune cell activation (Frick et al., 2004). ATMmvp2a methylation was not consistently associated with serum levels of any OCM metabolite. Methylation was associated with serum folate levels in EPIC1 (p=0.005), but not in the EPIC2 (p=0.97), nor in combined analysis (p=0.09). Serum folate was significantly correlated with homocysteine levels in EPIC1 (p=8.022e-06, rho=-0.49) and EPIC2 (p=3.722e-06, rho=-0.37). The MTHFR C allele was negatively associated with serum homocysteine, with both heterozygotes and homozygotes showing significant associations in both EPIC1, EPIC2 and combined analysis, especially after adjustment for serum folate levels (CT p=0.00011, CC p=1.52e-05). However, adjustment for MTHFR genotype did not significantly alter associations between ATMmvp2a methylation and other serum metabolites (Table 4.3).
Table 4.3. General linear models for investigation of association between ATM methylation and serum metabolites in test and validation sets

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Test glm p (n=72)</th>
<th>Validation (n=146)</th>
<th>Z-scores (combined)</th>
<th>Adjusted for MTHFR</th>
<th>Metabolite</th>
<th>Test glm p (n=72)</th>
<th>Validation (n=146)</th>
<th>Z-scores (combined)</th>
<th>Adjusted for MTHFR</th>
<th>Adjusted for fasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>homocysteine</td>
<td>0.081</td>
<td>0.943</td>
<td>0.277</td>
<td>0.477</td>
<td>0.32</td>
<td>neopterine</td>
<td>0.679</td>
<td>0.056</td>
<td>0.217</td>
<td>0.301</td>
</tr>
<tr>
<td>methylmalonic Acid</td>
<td>0.274</td>
<td>0.851</td>
<td>0.403</td>
<td>0.526</td>
<td>Tryptophan_B</td>
<td>0.935</td>
<td>0.113</td>
<td>0.171</td>
<td>0.151</td>
<td>0.199</td>
</tr>
<tr>
<td>cysteine</td>
<td>0.16</td>
<td>0.318</td>
<td>0.1</td>
<td>0.139</td>
<td>0.113</td>
<td>Kynurenine_A</td>
<td>0.84</td>
<td>0.138</td>
<td>0.239</td>
<td>0.194</td>
</tr>
<tr>
<td>methionine</td>
<td>0.217</td>
<td>0.793</td>
<td>0.529</td>
<td>0.782</td>
<td>0.519</td>
<td>Tryptophan_D</td>
<td>0.614</td>
<td>0.396</td>
<td>0.34</td>
<td>0.208</td>
</tr>
<tr>
<td>serine</td>
<td>0.283</td>
<td>0.629</td>
<td>0.32</td>
<td>0.505</td>
<td>0.288</td>
<td>Kynurenine_D</td>
<td>0.302</td>
<td>0.024</td>
<td>0.196</td>
<td>0.31</td>
</tr>
<tr>
<td>glycine</td>
<td>0.942</td>
<td>0.931</td>
<td>0.975</td>
<td>0.991</td>
<td>0.937</td>
<td>hydroxyknurenine</td>
<td>0.538</td>
<td>0.046</td>
<td>0.04</td>
<td>0.063</td>
</tr>
<tr>
<td>sarcosine</td>
<td>0.392</td>
<td>0.088</td>
<td>0.404</td>
<td>0.616</td>
<td>0.396</td>
<td>kynurenic acid</td>
<td>0.389</td>
<td>0.032</td>
<td>0.023</td>
<td>0.022</td>
</tr>
<tr>
<td>cystathionine</td>
<td>0.868</td>
<td>0.937</td>
<td>0.874</td>
<td>0.635</td>
<td>0.794</td>
<td>xanthurenic acid</td>
<td>0.51</td>
<td>0.205</td>
<td>0.153</td>
<td>0.1</td>
</tr>
<tr>
<td>folate</td>
<td>0.005</td>
<td>0.972</td>
<td>0.218</td>
<td>0.143</td>
<td>0.089</td>
<td>anthralinic acid</td>
<td>0.855</td>
<td>0.356</td>
<td>0.469</td>
<td>0.351</td>
</tr>
<tr>
<td>cobalamin</td>
<td>0.808</td>
<td>0.423</td>
<td>0.492</td>
<td>0.389</td>
<td>0.796</td>
<td>3-hydroxyanthranilic acid</td>
<td>0.791</td>
<td>0.022</td>
<td>0.057</td>
<td>0.098</td>
</tr>
<tr>
<td>pyridoxal Phosphate</td>
<td>0.877</td>
<td>0.403</td>
<td>0.51</td>
<td>0.562</td>
<td>0.486</td>
<td>fasting_status</td>
<td>0.076</td>
<td>0.186</td>
<td>0.034</td>
<td>0.071</td>
</tr>
<tr>
<td>pyridoxal</td>
<td>0.549</td>
<td>0.199</td>
<td>0.456</td>
<td>0.306</td>
<td>0.492</td>
<td>MTHFR TT*</td>
<td>0.126</td>
<td>0.024</td>
<td>0.004</td>
<td>0.006</td>
</tr>
<tr>
<td>pyridoxic acid</td>
<td>0.104</td>
<td>0.387</td>
<td>0.118</td>
<td>0.248</td>
<td>0.127</td>
<td>MTHFR CT**</td>
<td>0.401</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>pyridoxamine</td>
<td>0.062</td>
<td>0.57</td>
<td>0.318</td>
<td>0.562</td>
<td>0.37</td>
<td>MTHFR***</td>
<td>0.065</td>
<td>0.97</td>
<td>0.22</td>
<td>0.18</td>
</tr>
</tbody>
</table>

* MTHFR TT: General linear model testing for significantly different methylation between MTHFRc677t genotypes TT and CC

** MTHFR CT: General linear model testing for significantly different methylation between MTHFRc677t genotypes CT and CC

*** MTHFR: General linear model testing for a linear trend in methylation across all three genotypes.
Surprisingly, ATMmvp2a methylation was significantly associated with serum levels of tryptophan metabolites: kynurenic acid (KA), hydroxykynurenine, kynurenine-D and 3-hydroxyanthranillic acid in EPIC2, and in the combined analysis, independent of MTHFR genotype or fasting status, however, only KA remained significant after adjusting for fasting status (p=0.02) or MTHFR genotype (p=0.03) (Table 4.3). Interquartile analysis of the association indicated that ATMmvp2a methylation was significantly lower within the highest KA quartile (median 73.89%, n=55) compared to the lowest quartile (median 77.90%, n=55) (p=0.009) (Fig 4.3). KA was positively correlated with hydroxykynurenine (p=1.143e-05, rho=0.29), 3-hydroxyanthranillic acid (p=1.424e-08, rho=0.37), and kynurenine (p=0.0014, rho=0.23), and negatively correlated with ATMmvp2a DNA methylation (p=0.007, rho=0.19), but was not correlated with tryptophan or OCM metabolites, folate or homocysteine (data not shown).
Figure 4.3. ATMmvp2a DNA methylation measured by pyrosequencing, stratified by quartiles of kynurenic acid (KA) measured using liquid-chromatography-mass-spectrometry in healthy female control blood samples. Samples represent two subgroups (EPIC1 and EPIC2) within the European Prospective Investigation into Nutrition and Cancer (EPIC). DNA methylation is represented by z-scores (Y-axis), enabling combined analysis of subgroups.
4.3.4 Tissue-specific DNA methylation

Measurement of DNA methylation variability in whole blood cell populations is often confounded by the occurrence of tissues-specific DNA methylation between different cell types, leading to apparent differential methylation between individuals with different cell-type proportions (Koestler et al., 2012; Reinius et al., 2012). We measured ATM methylation in monocytes, B lymphocytes and T lymphocytes from a single individual, and monocytes and T cells from a second individual, separated by magnetic bead linked antibody cell separation (MACS) (Flanagan et al., 2009b). T lymphocytes displayed ATMmpv2a methylation of 28% and 43% in the two individuals, whereas monocytes and B cells were heavily methylated (93-98%) in monocytes and B cells (Fig 4.4). The ATMmpv2b repetitive element MVP 381bp downstream of ATMmpv2a also displays hypomethylation within T lymphocytes compared with other cell types, but to a lesser extent (21-28% lower methylation), consistent with higher methylation and lower methylation variability at this region, suggesting that inter-individual variability in blood of ATMmpv2a methylation may be restricted to T lymphocytes.
Figure 4.4 DNA methylation analysis (pyrosequencing) of (A) ATMmvp2a and (B) ATMmvp2b in blood cell fractions for two individuals, including monocytes and T lymphocytes for participant A (blue bars), and B lymphocytes, monocytes and T lymphocytes from participant B (green bars). 100% methylated (MC) and unmethylated (UC) controls (red bars), and whole PBMCs from a third individual, participant C (purple bar) are shown for reference. Error bars represent standard deviation across three technical replicates (PCR and pyrosequencing).
4.3.5 Fasting Status

ATMmvp2a methylation was substantially lower in blood samples from individuals who had fasted compared with non-fasting samples for both EPIC1, and EPIC2 sub-groups, and this difference was significant for the combined analysis (p=0.03) (Table 4.3 and Fig 4.5). The fasting-status variable was also available for three EPIC prospective breast cancer cases-control sub-groups, including the previously reported EPIC3 sub-group (Shenker et al., 2012), and two new sub-groups (EPIC4 and EPIC5). ATMmvp2a methylation was consistently lower in fasting blood samples compared with non-fasting samples in all studies, and this was statistically significant for EPIC3 (p=1.016e-05) and EPIC5 (p=0.005), but not for EPIC4 (p=0.37) (Fig 4.5). Using combined Z-scores for all sub-groups, ATMmvp2a methylation was significantly lower in individuals who had fasted prior to blood draw (n=689) compared to those who had not (n=485) (Wilcoxon signed rank sum test, p=4.572e-07), and this was independent of age, BMI and study centre. The average methylation difference between non-fasting and fasting individuals overall was 5.2% across all study subjects (range 4.0-12.2). In order to determine the effect of fasting on ATMmvp2a methylation, a meta-analysis was conducted, combining the odds ratios (OR) from all five sub-groups, representing the likelihood of a fasting individual displaying ATMmvp2a methylation below the median (Fig 4.5 (F)). The individual ORs (95% confidence intervals (95% CIs) for the subgroups were: EPIC1: 1.57 (0.88-2.91), EPIC2: 1.32 (0.64-2.97), EPIC3: 1.66 (1.38-2.0), EPIC4: 1.25 (0.92-1.68), and EPIC5: 1.65 (1.16-2.38), and these were statistically significant for EPIC 3 (p=9.28E-08) and EPIC5 (p=0.006). The summary OR, using a random effects model, was 1.56 (1.36-1.77), and the estimated heterogeneity variance was not significant (p=0.207). Meta-analysis in controls and cases separately, using only case-control sub-groups (EPIC3, EPIC4 and EPIC5), showed that ATMmvp2a methylation was significantly associated with fasting in both controls and cases, however, the summary effect was greater
in controls (1.68 (95% CI: 1.04, 2.32)) than in cases (1.46 (95% CI: 1.25, 1.66)), and there was significant heterogeneity variance in controls (0.2, p=0.002), but not in cases (0, p=1) (Fig 4.6).
**Figure 4.5:** Association of ATMmvp2a DNA methylation with Fasting Status. (A-E) ATMmvp2a methylation was lower in fasting individuals compared with non-fasting individuals in all studies. (F) Meta-analysis of association between ATMmvp2a methylation and fasting status based on odds ratio (OR) of below median ATMmvp2a methylation in fasting individuals, using a random effects model weighted by study size. Het p=Estimated Heterogeneity Variance p-value.
Figure 4.6 Meta-analysis of association between ATMmvp2a methylation and fasting status stratified by case-control status based on odds ratio (OR) for having <median ATMmvp2a methylation in fasting individuals, using a random effects model weighted by study size. Cases in red and controls in black. het=estimated heterogeneity variance p-value.
4.3.6 Age

As reported in chapter 2 ATMvp2a methylation was significantly associated with age in controls ($r^2=0.27$, $p=0.008$), but not cases ($r^2=0.035$, $p=0.583$) in EPIC3. The association in controls remained significant after adjusting for fasting status ($p=0.0003$). Similarly, in combined new analysis, ATMvp2a methylation was significantly increased with age in controls ($n=418$, $r^2=0.42$, $p=0.004$), but not cases ($n=197$, $r^2=-0.05$, $p=0.99$), after adjusting for fasting (Fig 4.7). A previous report indicated that DNA methylation at the LINE1 repetitive element may be associated with serum glucose (Ulrich et al., 2012). However, we observed no evidence of association with fasting status and DNA methylation at LINE1 (Fig 4.8 (A)). Furthermore, there was no association between DNA methylation at the ATMvp2b region (Fig 4.8 (B)), suggesting that the fasting effect is locus specific.
Figure 4.7. ATMmvp2a DNA methylation in blood, measured by pyrosequencing plotted against age at blood draw in (A) controls (n=719) and (B) prospectively collected breast cancer cases (n=491) from all combined subgroups from the European Prospective Investigation into Nutrition and Cancer study. Rho and p-values shown for spearman correlations between methylation (z-scores) and age.
Figure 4.8. **LINE1 and ATMmvp2b methylation stratified by fasting status** within the EPIC3 subgroup of the European Prospective Investigation into nutrition and Cancer (EPIC) study. DNA methylation measured by pyrosequencing of LINE1 (A) and ATMmvp2b (B) in a population study of non-fasting (n=308) and fasting (n=251) individuals.
4.3.7 *Ex-vivo* Glucose Restriction

Since a response to fasting is rapid (~10 hours difference between fasting and non fasting individuals), and ATM has a role in glucose metabolism, we hypothesised that the effect of fasting on DNA methylation may be mediated by a depletion of a blood glucose or cellular energy. Therefore, we investigated whether the fasting effect could be recapitulated by *in vitro* glucose restriction of peripheral blood mononuclear cells (PBMCs). Consistent with the epidemiological fasting data, average ATMmvp2a methylation across all subjects was significantly lower (5.15%) in PBMCs cultured in 0.5g/L (50%) glucose compared with 1.0g/L (100%, *p*=0.046) (Fig 4.9). There was inter-individual variability in the methylation response, whereby four out of six participants displayed methylation loss, whereas two subjects showed no change (Fig 4.9 (B)). Lower methylation with glucose restriction was consistent for the 0.25g/L glucose condition in two out of three individuals (Table 4.4). There was no change on average in ATMmvp2a methylation from baseline to 1.0g/L, indicating that ATMmvp2a methylation alteration was due to glucose restriction and not *in vitro* culturing alone. Using an MTT cell viability assay we show a significant loss of cell viability with complete glucose restriction (0%, *p*=1.74 e-05), however, there was no change between 1.0g/L (100%) and 0.5g/L (50%), or 0.25g/L (25%) conditions, indicating that partial glucose-restriction-induced ATMmvp2a hypomethylation was independent of cell death or proliferation (Fig 4.10). However, ATMmvp2a methylation was not altered by glucose restriction in cell lines derived from normal breast epithelium (MCF10a), and an ER+ invasive breast cancer (MCF7) (fig 4.11), supporting the idea that ATMmvp2a hypomethylation is cell-type-specific.
Figure 4.9: Reduction of ATM methylation upon glucose restriction. (A) Mean ATMmvp2a methylation, measured by pyrosequencing in PBMCs from six individuals at baseline, and after culture for 24h in media with glucose concentrations of 1g/L (full-glucose), and 0.5g/L (glucose-restricted). Error bars represent standard deviation across six individuals. *p<0.05. (B) Showing ATMmvp2a methylation at 1.0g/L and 0.5g/L glucose conditions for each individual separately, indicating inter-individual variability in epigenetic response to glucose restriction. Error bars represent standard error of the mean across three technical replicates.
Figure 4.10 Cell proliferation in PBMCs in response to glucose restriction. PBMCs collected from three healthy female participants were cultured for 24 hours in different medium glucose concentrations (1g/L (100%) (X-axis). Y-axis represents Percentage Cell Viability measured by MTT assay. Error bars represent standard deviation across three technical replicates. P-values represent t-test compared with 100% (1g/L) control.
Figure 4.11. ATMmvp2a methylation analysis in glucose restricted cell lines. Mean ATMmvp2a methylation, measured by pyrosequencing in an ER+ breast cancer (MCF7) and a normal breast epithelium (MCF10A) cell line, after culture for 24h in media with glucose concentrations of 1g/L, 0.5g/L, 0.1g/L and 0g/L. Error bars represent standard deviation across three technical replicates.
Table 4.4. Mean and standard deviation (SD) of ATMmvp2a methylation in PBMCs of 6 participants (A-F) after culture for 24 hours in media of different glucose concentrations (1.0g/L, 0.5g/L, 0.25g/L, 0g/L).

<table>
<thead>
<tr>
<th>Participant</th>
<th>Baseline</th>
<th>Mean 1.0g/L</th>
<th>SD 1.0g/L</th>
<th>Mean 0.5g/L</th>
<th>SD 0.5g/L</th>
<th>Mean 0.25g/L</th>
<th>SD 0.25g/L</th>
<th>Mean 0g/L</th>
<th>SD 0g/L</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>51</td>
<td>45.4</td>
<td>4.64</td>
<td>36.7</td>
<td>12.04</td>
<td>*</td>
<td>_</td>
<td>49.51</td>
<td>4.61</td>
</tr>
<tr>
<td>B</td>
<td>38.34</td>
<td>41.84</td>
<td>5.11</td>
<td>33.41</td>
<td>3.71</td>
<td>_</td>
<td>_</td>
<td>46.14</td>
<td>4.21</td>
</tr>
<tr>
<td>C</td>
<td>49.71</td>
<td>50.68</td>
<td>5.61</td>
<td>50.54</td>
<td>5.96</td>
<td>_</td>
<td>_</td>
<td>49.56</td>
<td>3.52</td>
</tr>
<tr>
<td>D</td>
<td>63.61</td>
<td>43.87</td>
<td>0.89</td>
<td>45.07</td>
<td>2.01</td>
<td>49.42</td>
<td>2.86</td>
<td>52.34</td>
<td>1.57</td>
</tr>
<tr>
<td>E</td>
<td>38.17</td>
<td>40.97</td>
<td>1.32</td>
<td>36.41</td>
<td>4.11</td>
<td>36.65</td>
<td>4.94</td>
<td>34.8</td>
<td>3.83</td>
</tr>
<tr>
<td>F</td>
<td>47.07</td>
<td>53.47</td>
<td>4.87</td>
<td>44.2</td>
<td>5.32</td>
<td>39.4</td>
<td>4.2</td>
<td>38.61</td>
<td>3.14</td>
</tr>
</tbody>
</table>

* 0.25g/L glucose concentration not measured for participants A-C
4.3.8 Genome-wide Analysis of Effect of Fasting on DNA Methylation

DNA methylation data at ~480,000 loci genome-wide, generated using the Illumina Infinium HumanMethylayion450 BeadChip array, was available for 311 EPIC PBMC DNA samples (controls=156, n cases=155) for which fasting status was available, enabling genome-wide screening for additional loci associated with fasting. This sample set included 210 samples from EPIC5, and an additional 101 samples (Table 4.1).

4.3.8.1 Analysis Method 1

Initial analysis was restricted to the EPIC5 subgroup for which ATMmvp2a methylation data was also available (n=210) as this represented a positive control for the effect of fasting on DNA methylation. Using univariate linear regression, 8112 loci were significantly associated with fasting with q value <0.05 (p value after FDR correction) (Fig 4.12). 149 significant loci had median absolute Δβ-values associated with fasting (difference in median methylation between non-fasting and fasting individuals) >0.05 (5%), with 146 loci showing a loss of methylation in fasting individuals and 3 loci an increase in methylation in fasting individuals. The largest decrease in methylation associated with fasting was 11%, and the largest increase was 5.7%. TOX2_c cg20889774 (Δβ =0.92), ASB10_c cg01954686 (Δβ =0.11) and HDAC4 cg04011897 (Δβ =0.037) were selected for validation in an independent sample set based on low q values, high absolute Δβ values in the case of TOX2 and ASB10, and multiple (11) loci associated with fasting within the same gene for HDAC4 (Table 4.5). The validation sample set consisted of 104 non-fasting and 76 fasting individuals (all healthy control subjects). TOX2 methylation was significantly lower in fasting (median=92.5) compared with non-fasting (median=93.7) individuals in the validation set (p=0.03) (Fig 4.13 (A)), however, the difference in median methylation (1.2%) was much smaller than the Δβ-value for the discovery set, and the methylation distribution was much higher and narrower
in the validation set than the discovery set (Fig 4.13 (B)). Fasting-associated methylation differences did not validate for either the ASB10 or HDAC4 loci, and both displayed very high methylation (90-95%) with narrow methylation ranges compared with array data (Fig 4.13 (B,C)), suggesting a skewing of methylation values caused either by differences between pyrosequencing and Infinium-based methylation analysis, or introduced through array data processing.
Figure 4.12. Top 8378 loci significantly associated with fasting status (FDR <0.05). Each point represents a single CpG site. \(-\log_{10} q\)-values (Y-axis) are plotted against the $\Delta\beta$-values (median $\beta$-value for fasting individuals - median $\beta$-value for non-fasting individuals) (X-axis). Negative values on the X-axis indicate lower methylation in fasting (n=170), compared with non-fasting (n=46) individuals. The solid line at 0.00 (X-axis) represents a 0% difference in methylation. Broken lines represent cut-offs for $\Delta\beta$-values of lesser than -0.05 or greater than 0.05, indicating 5% decreases or increased in DNA methylation, respectively, in fasting versus non-fasting individuals.
Figure 4.13. Discovery of additional candidate fasting-association loci in Illumina 450k array data. Upper box-plot within each panel displays significant differential methylation between non-fasting and fasting individuals within Illumina 450k array data. Lower box-plot within each panel represents pyrosequencing of the same locus in an independent sample set of healthy controls (n non-fasting=104, n fasting=76). For (A) TOX2, (B) HDAC4 and (C) ASB10 (analysis method 1) array data n fasting=84, n non-fasting=46.
Figure 4.13. Discovery of additional candidate fasting-association loci in Illumina 450k array data (continued from previous page). Upper box-plot within each panel displays significant differential methylation between non-fasting and fasting individuals within Illumina 450k array data. Lower box-plot within each panel represents pyrosequencing of the same locus in an independent sample set of healthy controls (n non-fasting=104, n fasting=76). For (D) ADARB2 and (E) PER3 (analysis method 2) array data n fasting=175, n non-fasting=53.* p<0.05, *** p<0.001.
4.3.8.2 Analysis Method 2

In order to maximise the potential of identifying loci associated with fasting, the additional samples for which array data was available (n=49 controls, 55 cases, see table 4.1) were included in the analysis, and a new analytical approach was used. Firstly, all polymorphic CpG sites, i.e. CpG sites harbouring SNPs within the C or G, on the 450k array (listed within a recently reported re-annotation of the array (Price et al., 2013) were excluded from analysis in order to exclude false positives occurring due to chance differences in allele frequency between experimental groups, which bias methylation at those sites. Also excluded were probes with >1 in silico binding sites, listed in the same re-annotation, in order to restrict analysis to loci with high-confidence annotation. Using a non-parametric Wilcoxon-rank sum test to identify loci significantly associated with fasting, 106 loci were associated with fasting after FDR correction (q<0.05), with absolute Δβ-values > 0.05, included 58 loci displaying a decrease of methylation in fasting individuals and 48 loci displaying an increase in fasting individuals (Fig 4.14). Among these 106 were 4 adjacent loci within the promoter of LGR6, all displaying a methylation increase (11.9-21.4%) in fasting, compared with non-fasting individuals, and two adjacent loci within the FAM101A gene, both displaying decreased methylation (5.4-6%) in fasting, compared to non fasting individuals (Fig 4.14 and Table 4.5).
Figure 4.14. Top 106 loci significantly associated with fasting status in analysis 2. Showing significant loci (FDR p-value (q-value) <0.05) with absolute Δβ-values >0.05 (5%). Each point represents a single CpG site. – log10 q-values (Y-axis) are plotted against the Δβ-values (median β-value for fasting individuals - median β-value for non-fasting individuals) (X-axis). Negative values on the X-axis indicate lower methylation in fasting, compared with non-fasting individuals. The solid line at 0.00 (X-axis) represents a 0% difference in methylation. Coloured points represent loci within LGR6 (red), ADARB2 (green), PER3 (dark blue), ASB10 (light blue) and FAM101A (pink).
Table 4.5. 450k array loci associated with fasting in analysis 1 and 2

<table>
<thead>
<tr>
<th>Illumina ID</th>
<th>Chr*</th>
<th>Genomic position</th>
<th>Gene</th>
<th>Gene-location</th>
<th>Estimate (glm**)</th>
<th>p-value (glm)</th>
<th>p-value (glm) FDR</th>
<th>p-value (wilcox***</th>
<th>p-value (wilcox) FDR</th>
<th>median methylation (fasting)</th>
<th>median methylation (non-fasting)</th>
<th>Δβ****</th>
<th>Assay type</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg01954686</td>
<td>7</td>
<td>150885654</td>
<td>ASB10</td>
<td>TSS1500</td>
<td>-0.463</td>
<td>1.41E-10</td>
<td></td>
<td></td>
<td></td>
<td>0.55</td>
<td>0.67</td>
<td>-0.12</td>
<td>II</td>
</tr>
<tr>
<td>cg20889774</td>
<td>20</td>
<td>42548586</td>
<td>TOX2</td>
<td>5UTR</td>
<td>-0.48</td>
<td>1.13E-10</td>
<td></td>
<td></td>
<td></td>
<td>0.31</td>
<td>0.4</td>
<td>-0.09</td>
<td>II</td>
</tr>
<tr>
<td>cg04011897</td>
<td>2</td>
<td>240014762</td>
<td>HDAC4</td>
<td>Body</td>
<td>-0.344</td>
<td>6.76E-11</td>
<td>4.90E-08</td>
<td></td>
<td></td>
<td>0.73</td>
<td>0.76</td>
<td>-0.03</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CpGs selected for validation from analysis 1

<table>
<thead>
<tr>
<th>Illumina ID</th>
<th>Chr*</th>
<th>Genomic position</th>
<th>Gene</th>
<th>Gene-location</th>
<th>Estimate (glm**)</th>
<th>p-value (glm)</th>
<th>p-value (glm) FDR</th>
<th>p-value (wilcox***</th>
<th>p-value (wilcox) FDR</th>
<th>median methylation (fasting)</th>
<th>median methylation (non-fasting)</th>
<th>Δβ****</th>
<th>Assay type</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg12799314</td>
<td>10</td>
<td>1405193</td>
<td>ADARB2</td>
<td>Body</td>
<td>-0.564</td>
<td>5.19E-07</td>
<td>1.57E-05</td>
<td>8.28E-07</td>
<td>7.19E-05</td>
<td>0.83</td>
<td>0.89</td>
<td>-0.06</td>
<td>I</td>
</tr>
<tr>
<td>cg10059324</td>
<td>1</td>
<td>7884824</td>
<td>PER3</td>
<td>Body</td>
<td>-0.357</td>
<td>2.89E-09</td>
<td>3.52E-07</td>
<td>1.95E-08</td>
<td>5.89E-06</td>
<td>0.75</td>
<td>0.81</td>
<td>-0.06</td>
<td>II</td>
</tr>
</tbody>
</table>

CpGs selected for validation from analysis 2

<table>
<thead>
<tr>
<th>Illumina ID</th>
<th>Chr*</th>
<th>Genomic position</th>
<th>Gene</th>
<th>Gene-location</th>
<th>Estimate (glm**)</th>
<th>p-value (glm)</th>
<th>p-value (glm) FDR</th>
<th>p-value (wilcox***</th>
<th>p-value (wilcox) FDR</th>
<th>median methylation (fasting)</th>
<th>median methylation (non-fasting)</th>
<th>Δβ****</th>
<th>Assay type</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg06944096</td>
<td>12</td>
<td>124787364</td>
<td>FAM101A</td>
<td>5UTR</td>
<td>-0.412</td>
<td>6.55E-15</td>
<td>7.52E-11</td>
<td>1.26E-12</td>
<td>3.16E-08</td>
<td>0.7</td>
<td>0.76</td>
<td>-0.06</td>
<td>I</td>
</tr>
<tr>
<td>cg06879608</td>
<td>12</td>
<td>124787404</td>
<td>FAM101A</td>
<td>5UTR</td>
<td>-0.351</td>
<td>2.33E-15</td>
<td>3.65E-11</td>
<td>2.47E-12</td>
<td>4.14E-08</td>
<td>0.64</td>
<td>0.7</td>
<td>-0.06</td>
<td>I</td>
</tr>
</tbody>
</table>

CpGs in genes for which >1 locus was associated with fasting in analysis 2

<table>
<thead>
<tr>
<th>Illumina ID</th>
<th>Chr*</th>
<th>Genomic position</th>
<th>Gene</th>
<th>Gene-location</th>
<th>Estimate (glm**)</th>
<th>p-value (glm)</th>
<th>p-value (glm) FDR</th>
<th>p-value (wilcox***</th>
<th>p-value (wilcox) FDR</th>
<th>median methylation (fasting)</th>
<th>median methylation (non-fasting)</th>
<th>Δβ****</th>
<th>Assay type</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg05044291</td>
<td>1</td>
<td>202172867</td>
<td>LGR6</td>
<td>TSS200</td>
<td>0.353</td>
<td>0.006869</td>
<td>0.030974</td>
<td>0.01426</td>
<td>0.046986</td>
<td>0.32</td>
<td>0.15</td>
<td>0.17</td>
<td>II</td>
</tr>
<tr>
<td>cg06825163</td>
<td>1</td>
<td>202172912</td>
<td>LGR6</td>
<td>1stExon</td>
<td>0.415</td>
<td>0.003163</td>
<td>0.001419</td>
<td>0.001224</td>
<td>0.008708</td>
<td>0.51</td>
<td>0.3</td>
<td>0.21</td>
<td>II</td>
</tr>
</tbody>
</table>

*Chromosome, **univariate general linear model, ***Wilcoxon-rank-sum-test, ****Median methylation in fasting individuals - median methylation in non-fasting individuals
Spatial clustering of CpG sites displaying a consistent pattern of association with the outcome of interest have been validated in previous studies investigating the effect of smoking on DNA methylation (Joubert et al., 2012; Shenker et al., 2012). Spatial clustering of fasting-associated loci was therefore considered a potential indicator of genuine association, and was assessed for all genes to which ‘top hit’ fasting-associated loci were annotated, by plotting the $\Delta\beta$-values for all loci within those genes against genomic location (representative examples shown in Fig 4.15). Clear spatial clustering of fasting associated loci was evident for loci within ADARB2, FAM101A, HTR3A, KLK15, LDHAL6A, LGR6, LIN7B, PER3, PLD6, TMEM151B and YTHDC1, but not other loci.
**Fig 4.15.**

**A.** Fasting-associated $\Delta \beta$ for all LDHALA6 loci measured on 450k array mapped against genomic location

**B.** Fasting-associated $\Delta \beta$ for all GOLGA3 loci measured on 450k array mapped against genomic location

**Figure 4.15. Selecting of loci significantly associated with fasting** in Illumina Infinium HumanMethylayion450 BeadChip array for pyrosequencing validation based on concordant patterns of fasting-associated methylation of neighbouring CpG sites. $\Delta \beta$-values associated with fasting status (median DNA methylation in fasting individuals minus median methylation in non-fasting individuals ($Y$-axis)) for all loci annotated to genes in which ‘top hit’ significantly fasting-associated loci reside, plotted against genomic location ($X$-axis). Each point represent a CpG site represented on the array (significantly associate loci circled in red). (A) LDHAL6A, an example of a locus significantly associated with fasting and displaying clustering of fasting-associated methylation pattern with neighbouring CpG sites, whereby several local CpG sites, though not statistically significantly associated with fasting, display concordant hypomethylation. (B) GOLGA3, example of a locus not displaying clustering, as no neighbouring loci display concordant fasting-associated methylation.
Plotting of fasting-associated regions displaying spatial clustering revealed that the association with fasting of loci within LGR6 and LIN7B were confounded by trimodal methylation patterns that are suggestive of a haplotype-associated DNA methylation pattern that is not due to polymorphic CpG sites (Discussed in chapter 6). Other loci displaying spatial clustering did not display clear trimodal methylation distributions (Fig 4.16); however, haplotype information would be required to determine whether allelic skewing of methylation occurs at these sites.
Figure 4.16. Strip-charts showing DNA methylation distributions at (A) PER3-cg10059324, (B) ADARB2-cg12799314, (C) LIN7B-cg10667338 and (D) FAM101A-cg06879608 in EPIC5/EPIC6 Illumina Infinium HumanMethylation450 BeadChip array data (n=332).
Validation of many loci was limited by their occurrence within CpG islands, as high CpG density prevented pyrosequencing assay design for loci such as LDHAL6a and FAM101A. Validation was attempted for loci within intragenic regions of the ADARB2 (ADARB2-cg12799314) and PER3 (PER3-cg10059324) genes; however neither of these loci validated in the EPIC6 independent sample set. For PER3, as with TOX2, HDAC4 and ASB10, the methylation range was much narrower and higher in pyrosequencing data compared with array data, whereas for ADARB2, the methylation range was similar to the array data, but there was no association with fasting (Fig 4.13).
4.4 Discussion

We aimed to identify the factors underpinning ATMmp2a DNA methylation variability, including genetic, demographic, dietary and immunological factors. Whereas other studies have investigated factors affecting DNA methylation at multiple repetitive elements (Ulrich et al., 2012), and across the genome using microarrays (Lam et al., 2012), we have focused in detail on a single region of inter-individual variability, at which hypermethylation was associated with breast cancer risk in our previous investigation (Brennan et al., 2012b).

We investigated the relationship between ATMmp2a methylation variability and ATM genetic haplotype, as genetic polymorphisms within ATM are associated with breast cancer risk (Fanale et al., 2012; Lavin, 2008), and a common ATM haplotype may be associated with breast cancer risk in BRCA1 mutation carriers (Rebbeck et al., 2011). It seemed likely that ATMmp2a methylation had been associated with breast cancer risk due to haplotype-associated DNA methylation, as allelic skewing of DNA methylation is a frequent event (Schalkwyk et al., 2010; Shoemaker et al., 2010) However, we found no evidence for an association between DNA methylation and ATM haplotype.

DNA methylation is also affected by genetic variability in trans, due to polymorphism in genes encoding enzymes required for methylation (Moarefi and Chedin, 2011; Vineis et al., 2011). Folate-dependent one-carbon metabolism (OCM) provides methyl groups for all methylation reactions via the universal methyl donor S-adenosylmethionine (SAM), derived from dietary folate, choline, and B-vitamins (Chiacchiera et al., 2013); therefore genetic polymorphism in OCM enzymes such as MTHFR, as well as supply of OCM micronutrients, may be directly limiting of DNA methylation (Christensen and Marsit, 2011; Vineis et al., 2011). ATMmp2a methylation was associated with MTHFRc677t genotype in healthy control individuals, but not cases. Increased methylation associated with the C allele is
consistent with previous findings (Friso et al., 2002) therefore the association of ATMmvp2a methylation with MTHFR status is biologically plausible. Given the widespread affect of common MTHFR polymorphisms on DNA methylation and on disease risk (Friso et al., 2002; Liu et al., 2013), further work will be required to determine whether it is necessary to measure and adjust for MTHFR genotype in EWAS studies.

We found a statistically significant association between ATMmvp2a methylation and the tryptophan derivative kynurenic acid (KA). Whereas this requires further validation, the indication that DNA methylation might be altered by kynurenine pathway metabolites has not been previously reported and is potentially very interesting, as levels of blood kynurenic acid increase with age, and elevation of KA is associated with oxidative stress, and several conditions, including inflammatory bowel disease, Alzheimer’s disease and Schizophrenia (Colin-Gonzalez et al., 2013). Kynurenines also play roles in immunity and inflammation, including activation of T cells, as well as a role in activation of the Aryl hydrocarbon receptor, a key mediator of xenobiotic metabolism (Stone et al., 2013). Validation of the association of ATMmvp2a methylation and serum kynurenines, assessment of the degree to which kynurenines affects DNA methylation across the genome, and investigation of the mechanism would be worthwhile. Given the role of KA in immune activation (Stone et al., 2013), it is possible that differential methylation associated with KA levels may reflect underlying variability in leukocyte populations. Otherwise, ATMmvp2a methylation variability may be specific to ATM, due to the role of ATM in mediating response of oxidative stress (Okuno et al., 2012).

To our knowledge, this study is the first to link peripheral blood DNA methylation with fasting status, and coupled with the recapitulation of this effect through ex vivo glucose restriction, suggests that this rapid change in methylation is reproducible and potentially an important
confounding factor to consider for future EWAS studies. Alteration of ATMmp2a methylation by transient exposures such as fasting is inconsistent with the prevailing dogma by which DNA methylation is established during early development and displays long-term mitotic stability (Feinberg et al., 2010). The findings that blood methylation differences associated with smoking are detectable in former smokers many years after smoking cessation (Shenker et al., 2012), and that altered methylation at several loci associated with developmental exposure to famine are detectable in blood decades later (Heijmans et al., 2008; Tobi et al., 2009), support the existence of stable epigenetic ‘memory’.

However, rapid genome-wide DNA demethylation occurs within the paternal genome during preimplantation (Chiacchiera et al., 2013) and occurs by ‘active’ demethylation, a poorly understood process which is thought to occur though sequential conversion of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), and then to 5-flourouracil (5fC) and 5-carboxylcytosine (5CaC) by the ten-eleven-translocase (TET) enzymes, possibly followed by a base-excision repair mechanism to restore 5mC (Chiacchiera et al., 2013). A few examples of active DNA methylation changes have been demonstrated in adult tissues. For example, Brain-derived neurotrophic factor (BDNF) and fibroblast growth factor 1 (FGF1) undergo active demethylation in post-mitotic neurons via oxidation and deamination of 5mC followed by base-excision repair (BER) (Guo et al., 2011; Martinowich et al., 2003). Rapid alteration of DNA methylation in skeletal muscle with exercise has been described (Barres et al., 2012), as has cyclical methylation of the Ps2/TTF1 gene promoter under estrogen activation in breast cells (Kangaspeska et al., 2008; Metivier et al., 2008). It is commonly speculated that epigenetic patterns associated with adult conditions may reflect aberrations in epigenetic programming due to innate influences or exogenous exposures during early development (Feinberg et al., 2006; Keating and El-Osta, 2013; Tobi et al., 2009), however,
greater consideration should be given to the possibility that associated epigenetic patterns represent consequences of disease-related metabolic or physiological states, as many loci, in addition to ATMvp2a, may be responsive to metabolic conditions.

Glucose and energy metabolism are linked with epigenetics mechanisms (Donohoe and Bultman, 2012). Studies investigating the profound effect of caloric restriction (CR) in extending cellular lifespan and preventing carcinogenesis in mammal (Hursting et al., 2010) have revealed that CR is largely mediated by histone deacetylation (Hursting et al., 2010; Willcox et al., 2007), and histone acetylation and DNA methylation are thought to work synergistically (Kalac et al., 2011). Long-term glucose restriction has been shown to alter gene expression, DNA methylation and histone acetylation of P16 and hTERT in vitro (Li et al., 2010b), however, our report of acute alteration of methylation on short-term glucose restriction supports the notion of rapid methylation changes independent of cell proliferation (Kangaspeska et al., 2008; Metivier et al., 2008). DNA methylation differences in blood (Milagro et al., 2011) and adipose tissue (Bouchard et al., 2010), associated with weight loss response to CR provide strong evidence that DNA methylation is implicated in the effects of energy metabolism on health outcomes. Whether variability in DNA methylation response to glucose restriction relates to metabolic disease or cancer risk remains to be investigated.

Short-term fasting induces several acute molecular and physiological changes, including widespread gene expression changes (Bouchard et al., 2010; Bouwens et al., 2007; Caimari et al., 2010), weight-loss, increase in haemoglobin, red blood cell count and growth hormone levels, decrease in bicarbonate levels, and change in lipid profiles (Horne et al., 2012). Prolonged fasting is associated with a decrease in leukocyte count and proinflammatory cytokine levels (Faris et al., 2012).
We do not find an association between ATMmvp2a methylation and breast cancer risk in either of two new prospective case control studies from the EPIC cohort included in this report. The identification of metabolic factors affecting ATMmvp2a methylation indicate that analysis of the breast cancer risk association may have been confounded by additional factors for which information is not available, and highlight the necessity to understand and control for factors influencing DNA methylation in future disease-association studies. The first epigenome-wide association studies (EWAS) have highlighted the potential of investigation of DNA methylation to help understand the contribution of DNA methylation to disease (Rakyan et al., 2011; Shenker et al., 2012). In order to establish the efficacy of this approach, and to identify potential confounding factors, it will be important to determine the extent to which metabolic variables and transient exposures such as fasting may affect genome-wide DNA methylation.

Epigenetic mechanisms are thought to interact with diet to influence breast cancer risk (Haggarty, 2013; Lampe et al., 2013; Li et al., 2003). Many breast cancer risk factors, such as obesity, diabetes, alcohol consumption, birth weight, menopause status, growth hormones, and circadian rhythm patterns, relate strongly to glucose metabolism and diet (Amadou et al., 2013; Choi et al., 2012; Feng and Lazar, 2012; Mauvais-Jarvis et al., 2013; McPherson et al., 2000; Thomson, 2012) and are associated with epigenetic signatures (Barres et al., 2009; Feinberg et al., 2010; Keating and El-Osta, 2013; Seki et al., 2012; Xu et al., 2013a). Biomarkers of response to fasting are frequently associated with metabolic phenotypes; Gene expression changes associated with fasting and re-feeding in PBMCs are blunted in obese rats (Oliver et al., 2013), and a recent study has identified DNA methylation signatures associated with weight-loss response to CR (Hursting et al., 2010); therefore, inter-individual variability in ATMmvp2a methylation fasting response may relate to metabolic
factors such as age or BMI. ATMmp2a methylation response to fasting and glucose restriction displays inter-individual variability and fasting response may be stronger in control blood samples compared to incident breast cancer cases; however the factors underlying this remain unclear. It is possible that ATMmp2a hypermethylation may be sustained in vivo due to hyperglycemia, an independent risk factor for breast cancer (Sieri et al., 2012). Whereas meta-analysis has indicated that the independent effect of hyperglycemia is likely to be mild (RR=1.11 (1.00-1.23)) (Boyle et al., 2012b), hyperglycemia is associated with other metabolism-related risk factors, including diabetes (Boyle et al., 2012a), and obesity (Amadou et al., 2013; Cheraghi et al., 2012), age, or menopausal status. A limitation of our study is that we did not have serial samples from the same individual before and after fasting, which would correct for baseline differences in methylation. A study using such samples, with phenotype data for metabolic factors would allow greater interrogation of the factors underlying response variability. The complex interactions between diet, epigenetics and disease risk are poorly understood, however, a greater understanding of the ways in which metabolic derivatives of dietary factors affect epigenetics will aid in our understanding (Donohoe and Bultman, 2012), as will investigation of variability in epigenetic response to dietary interventions that are known to reduce breast cancer risk (Lampe et al., 2013; Milagro et al., 2011).

There are interesting links between ATM, breast cancer risk, and glucose metabolism, as ATM mutations and haploinsufficiency are associated with diabetes mellitus (Bar et al., 1978; Miles et al., 2007), as well as increased breast cancer risk (Lavin, 2008). Furthermore, a SNP in ATM is associated with improved glycemic response to the antidiabetic drug metformin (Tkac, 2012; van Leeuwen et al., 2012; Zhou et al., 2011), treatment with which is associated with reduced breast cancer risk in diabetic individuals (Decensi et al., 2010).
Immune shifts in cell count represent a potential explanation for the association between ATMmvp2a methylation and glucose metabolism, as inter-individual DNA methylation differences can reflect differences in proportions of cell types that display tissue-specific methylation, rather than differences in the number of cells that are methylated within a given cell type (i.e. a change in percentage methylation) (Koestler et al., 2012). Both ATMmvp2a and ATMmvp2b were is hypomethylated in T cells compared with monocytes and B cells, with more significant hypomethylation in ATMmvp2a, consistent with lower methylation and higher variability at ATMmvp2a than ATMmvp2b in PBMCs overall. This suggests that ATMmvp2a WBC methylation variability may be unique to T lymphocytes, or a subset thereof, and case inter-individual variability may be far greater within this cell subset than in PBMCs overall. ATMmvp2a hypomethylation on fasting may reflect a proliferation or selective survival of hypomethylated T cells with fasting. However, hypomethylation induced by in vitro glucose restriction was independent of cell viability, suggesting active demethylation.

Consistently, T cell activation is associated with active DNA methylation changes, as rapid demethylation of the interleukin-2 gene (IL2) occurs in non-dividing T lymphocytes on activation (Bruniquel and Schwartz, 2003), and active demethylation of Foxp3 is required for activation of regulatory T cells in the thymus (Toker et al., 2013). ATMmvp2a may play a functional role, as T cells display highly distinct methylation patterns relative to other cell types, and loci hypomethylated in CD4+ and CD8+ T cells are enriched for genes involved in lymphocyte activation and T-cell-specific immune function (Reinius et al., 2012). All of the factors associated with ATMmvp2a methylation (except MTHFRc677t genotype, for which the mechanism in affecting DNA methylation is well established (Friso et al., 2002)) relate to T cell activation, as both glucose availability (Jacobs et al., 2008) and kynurenines (Mandi
and Vecsei, 2012) regulate T cell activation, and activation is blunted with age
(Czesnikiewicz-Guzik et al., 2008). Efficient T cell activation influences breast cancer risk as
immunosurveillance mediated by tumour-specific T cells are required to detect and kill
cancer cells (Haabeth et al., 2011). Furthermore ATM mutations are often associated with T
cell proliferation and T cell lymphomas (Miles et al., 2007; Stankovic et al., 1998).

Methylation analysis of blood samples sorted into discrete cell types, and functional assays
will be required to confirm that ATMmvp2a methylation variability is restricted to T cells, and
will determine if methylation is related to T cell activation state.

ATMmvp2a methylation was not altered by glucose restriction in either MCF7 or MCF10,
despite the fact that ATM is activated by serum starvation in vitro (Shi et al., 2012), and
inhibition of ATM activity in MCF7 and MCF10a cells alters glucose metabolism (Zakikhani et
al., 2012). This is possibly due to cell-type-specificity of ATMmvp2a methylation variability.
Alternatively it could be due to the TET enzymes, required for active demethylation, not
being expressed in cells lines that have been cultured long-term, or that cancer cells display
abnormal glucose metabolism (Li et al., 2010b), MCF10A cells requires insulin
supplementation, and DMSO required for cell cryopreservation, has been reported to induce
active DNA demethylation (Thaler et al., 2012). Therefore, several factors may have
impeded our ability to detect ATMmvp2a methylation alteration by glucose restriction in
these cell lines.

Of five fasting-associated loci selected for validation, only TOX2 validated, and four out of
five loci, including TOX2, displayed methylation distributions that were far higher and
narrower than was apparent in the array discovery set data, suggesting a skewing of the
data in array data processing. This was not due to peak-based correction, as this only
affects type II probes, whereas both probe types were represented among the fasting-
associated loci. Normalisation steps such as quantile normalisation or ComBat may have introduced bias into the analysis.

Further validation of TOX2 is required, as is further analysis to identify additional loci associated with fasting. Fasting-associated loci in addition to ATMmvp2a are likely to exist; however, the relative subtlety of the methylation change associated with fasting, in comparison to other factors that have been investigated using the 450k array, such as smoking or haplotype, may limit our ability to identify fasting-associated loci using this system. Among loci apparently associated with fasting, loci within LGR6 and LIN7B displayed trimodal methylation patterns that may indicate allele-specific DNA methylation. Allelic skewing of DNA methylation is frequent in the human genome and occurs predominately at polymorphic CpG sites (Heyn et al., 2013; Kerkel et al., 2008; Schalkwyk et al., 2010; Shoemaker et al., 2010). Allelic skewing represents a potential source of bias for EWAS studies (Price et al., 2013) therefore polymorphic CpG sites may be excluded from analysis. However, haplotype-associated methylation of non-polymorphic CpG sites clearly represents an overlooked potential confounding factor that should be assessed in future EWAS studies, especially where patterns of differential methylation are consistent across multiple adjacent loci.

High frequency of fasting-associated loci within CpG islands impeded validation of loci that displayed strong clustering of fasting-associated CpG sites, but that did not display apparent trimodal methylation distributions. It remains unclear, therefore, whether these regions are associated with fasting, or whether another unknown confounding factor accounts for these associations.
In conclusion, this study demonstrates the potential influence of disease-related metabolic pathways, other than OCM, in influencing DNA methylation, and raises the possibility that transient exposures such as fasting may have profound influence on DNA methylation. These findings may have important implications for epigenome-wide association studies investigating the role of peripheral blood DNA methylation in disease risk.
CHAPTER 5

Microarray-based discovery studies to identify novel breast cancer risk markers
5.1 Introduction

Novel breast cancer risk markers may help improve current disease risk models, improving early diagnosis, and aiding population stratification for screening (Amir et al., 2010). Our prospective investigation of ATM methylation (chapter2) provided proof of principle for the investigation of DNA methylation variability as a potential source of cancer risk biomarkers; however, additional markers, or ideally a panel of risk markers similar to the combination of multiple SNPs into a “polygenic risk score”, would be useful in helping to improve the sensitivity and specificity of breast cancer risk models. In recent years, new technologies have enabled high-throughput, genome-scale investigation of epigenetic patterns associated with phenotypes, exposures and disease outcomes (Langevin and Kelsey, 2013; Rakyan et al., 2011). Such technologies provide an opportunity to perform genome-scale discovery studies for identification of breast cancer risk markers.

It is now possible to carry out genome-wide DNA methylation (whole-methylome) analysis at single-base-resolution (Li et al., 2010c), and future investigation of DNA methylation variability will likely be greatly assisted by this ‘gold-standard’ method (Rakyan et al., 2011). However, the use of whole -genome bisulphite sequencing is currently limited to studies of small sample sizes (Feber et al., 2011; Hansen et al., 2011), due to its high cost. For population studies, the value of genome-coverage must be balanced with the requirements for large sample sizes and highly quantitative measurement. To this end, candidate gene studies and array-based techniques measuring DNA methylation at a small proportion of CpG sites are used for population studies. Several studies have now investigated blood DNA methylation in cancer patients (Marsit et al., 2011; Teschendorff et al., 2009; Widschwendter et al., 2008), and each identified multiple loci significantly associated with disease, indicating that blood DNA methylation may be associated with cancer outcomes, and potentially providing useful biomarkers for cancer diagnosis.
Candidate-gene studies can be used to investigate risk-associated DNA methylation at genes implicated in breast cancer in the literature, but is limited by currently available knowledge. Using the methyLight technique, Widschwendter et al (Widschwendter et al., 2008) identified five genes that were significantly differentially methylated in blood of breast cancer cases, and different gene sets predicting invasive-ductal and invasive-lobular carcinomas, potentially indicating cancer-pathway-specific differential methylation. A recent study identified subtle SFRP1 hypermethylation in blood of gastric cancer patients compared with controls using pyrosequencing (Tahara et al., 2013).

Illumina Infinium BeadChip arrays have become the most popular method for investigation of DNA methylation in population studies. Three retrospective studies have utilised the Illumina Infinium HumanMethylation27 BeadChip array, which measures methylation at ~27,578 CpG sites within gene promoters across the genome, to assess differential methylation in blood of cancer patients compared with controls. Teschendorff et al (Teschendorff et al., 2009) identified 2714 CpG sites that were differentially methylated in blood of ovarian cancer patients compared with unaffected controls in a relatively small sample set (n=113 cases, 148 controls), and 355 of these loci validated in an independent sample set. Marsit et al (Marsit et al., 2011) found 9 CpG sites that were statistically significantly hypermethylated in bladder cancer cases compared with healthy controls. Langevin et al (Langevin et al., 2012) identified a panel of six CpG sites that could predict case-control status for head and neck squamous cell carcinoma. Each of these loci was significantly hypomethylated in cases compared with controls, though with absolute $\Delta\beta$-values of only ~0.01 (1%), and with no evidence of validation presented in these studies.

These studies provide evidence for the utility of DNA methylation variability for development of biomarkers for cancer diagnosis, and some of the loci identified may represent potential markers of cancer risk (Teschendorff et al., 2009; Widschwendter et al., 2008). However;
many factors associated with the presence of active disease may affect DNA methylation, most importantly the possibility that cancer-associated immune responses may alter leukocyte proportions resulting in apparent differential methylation due to cell-type specific DNA methylation (Koestler et al., 2012; Langevin et al., 2012). Another potential confounding factor is the use of case samples derived from individuals who have undergone cancer treatment (Langevin et al., 2012; Teschendorff et al., 2009; Widschwendter et al., 2008), as cancer treatment may affect epigenetic profiles (Rakyan et al., 2011; Xu et al., 2013b). It remains possible; however, that many loci associated with cancer prevalence in retrospective studies may represent risk markers, as ATM hypermethylation identified in a retrospective study was validated as a risk marker in our prospective investigation (Chapter 2).

Recently, the first array-based prospective investigation of blood DNA methylation associated with breast cancer risk was reported (Xu et al., 2013b), and identified 250 loci associated with increased risk using the Illumina 27k array, 75.2% of which were hypomethylated in cases. Furthermore, they reported a >2-fold enrichment of differentially methylated loci near known breast cancer susceptibility genes including ATM, though only one locus was statistically significant after FDR correction. Unfortunately, no external validation of any of the loci was performed; and as all of the differentially methylated loci displayed absolute Δβ-values of <0.05 (5% differential methylation), it is possible that many of the associated loci identified represent statistical or technical artefacts rather than genuine differential methylation (Garcia-Closas et al., 2013b). Some loci reported in this study displayed increasing strength of association with decreasing time to diagnosis (TTD), suggesting that they may represent markers of preclinical disease rather than susceptibility loci.
In order to identify potential risk markers, we aimed to conduct genome-scale discovery studies of DNA methylation loci associated with breast cancer, using blood DNA samples from prospective case-control studies and DNA methylation microarrays, and with an external validation stage using an independent sample set also from the EPIC study.
5.2 Materials and methods

5.2.1 Discovery study 1 - Differential Methylation Hybridisation (DMH)

5.2.1.1 Study samples

A subset of the KConFab study sample set described in chapter 2 were used for the DMH microarray, including whole blood DNA samples prospectively collected from 21 invasive breast cancer cases and 30 healthy unrelated (best-friend) controls (Table 5.1). Reference DNA 100% methylated and 0% methylated controls were also included on the array. Biological validation by pyrosequencing was carried out in the remainder of the KConFab study sample set (Table 5.1). Further pyrosequencing in a sporadic breast cancer study was carried out using the Breakthrough Generations Study (BGS) study described in chapter 2 (Table 5.1).
Table 5.1. KConFab study samples used for investigation of PARP1 methylation

<table>
<thead>
<tr>
<th>Sample set</th>
<th>Discovery set</th>
<th>Validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Controls</td>
<td>Cases</td>
</tr>
<tr>
<td>Case/control status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td>Age at blood draw in years, mean (range/SD*)</td>
<td>59 (49-75/7.22)</td>
<td>55 (36-71/11.38)</td>
</tr>
<tr>
<td>Menopausal status at blood draw, N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-menopausal</td>
<td>4%</td>
<td>61%</td>
</tr>
<tr>
<td>Post-menopausal</td>
<td>96%</td>
<td>39%</td>
</tr>
<tr>
<td>TTD* in months mean (range/SD)</td>
<td>–</td>
<td>21.46 (1-74/23.92)</td>
</tr>
<tr>
<td>Age at diagnosis, mean (range/SD)</td>
<td>–</td>
<td>56(37-72/10.67)</td>
</tr>
<tr>
<td>Alcohol drinker, N (%)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drinker</td>
<td>69%</td>
<td>62%</td>
</tr>
<tr>
<td>Non-drinker</td>
<td>27%</td>
<td>38%</td>
</tr>
<tr>
<td>Smoker (ever), N, (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoker</td>
<td>38%</td>
<td>38%</td>
</tr>
<tr>
<td>Non-smoker</td>
<td>58%</td>
<td>62%</td>
</tr>
</tbody>
</table>

*Time to diagnosis (from blood draw)

**Alcohol drinker status for individuals for which data was available
5.2.1.2 DMH custom tiled microarray

Differential methylation hybridisation (DMH) was carried out by Oxford Gene Technology (OGT). DMH is a semi-quantitative method of DNA methylation analysis which involves methylation-sensitive restriction digestion of DNA, fluorescent-labelling, followed by hybridisation of digestion fragments to an oligonucleotide microarray, allowing selective methylation analysis of candidate genes. Briefly, sample DNA is digested using the restriction enzyme *MseI*, which cuts at the restrictions site TTAA, and therefore, digests DNA into small ~200bp fragments, whilst leaving CpG islands (which are depleted for the TTAA sequence) intact. Cleaved ends of DNA fragments are ligated to universal linkers, followed by methylation sensitive restriction digestion of ligated fragments with enzyme *McrBI*, which cleaves methylated DNA. Restriction with *MseI* and ‘mock-digestion’ is performed on a separate aliquot of the sample for comparison. Linker PCR is then used to amplify fragments from mock-digested and McrBI digested samples, followed by purification of amplicons, and fluorescent labelling with Cy5 and Cy3, respectively. This is followed by competitive hybridisation of mock digested and McrBI digested DNA to a custom tiled microarray, and scanning of fluorescence signal. A DMH ratio (Ratio of the fluorescence signal intensities of Cy5 (mock-digested fragments) to Cy3 (digested fragments)) is used to estimate the methylation of a fragment, where unmethylated fragments have a DMH ratio ~1, whereas a DMH ratio >1 indicates that the fragment is methylated (Dai et al., 2008; Yan et al., 2002). The custom array was designed to cover gene promoters and intragenic regions, tiled every 50bp from -2kb to +8kb around the transcription start sites of 1381 candidate genes resulting in 8758 *MseI* genomic fragments for analysis. Use of a custom array was required for methylation analysis of intragenic regions that were not covered by other available arrays.
Candidate genes were selected from seven pathways commonly deregulated in breast cancer, including the AKT, BRCA1/2, DNA repair (HR, MMR, NHEJ), Fanconi, p53, WNT, and Redox pathways.

5.2.1.3 Statistical analysis

Raw signal intensities from the samples were extracted using Agilent feature extraction software. Two criteria were used to ‘pass’ each probe: 1) Signal intensities must not be higher than 65,000 (signal saturation) or lower than mean+2 standard deviation from negative control, and 2) probes must not be flagged by Agilent extraction software. The data points failing either of these criteria were treated as poor-quality data points, and were removed from analysis. Probes with poor-quality data points in > 50% of samples were removed from analysis. The median difference between log2 of the DMH ratios for the 100% methylated and 0% methylated controls were used to determine efficiency of enzymatic digestion using McrBC. Samples with median difference between 100% and 0% methylated controls less than 1.87 were removed from further analysis resulting in a final dataset of 18 cases and 24 controls. Methylation linear discriminant analysis (MLDA) was used to normalize the DMH ratios by equalising the signals from 100% and 0% methylated controls, as previously described (Dai et al., 2008). DMH ratios were expressed as a magnitude of difference. Data processing and normalisation was performed by the Epigenetics Unit Bioinformatician, Dr. Wei Dai. Wilcoxon rank sum test was used to test for differential methylation between cases and controls. False Discovery Rate (FDR) was used to correct for multiple testing, with FDR p<0.05 considered significant. Inter-quartile analysis and logistic regression was used to calculate an OR for breast cancer incidence for individuals
within the highest DNA methylation quartiles compared with the lowest (reference) quartile, with quartiles specified using controls only.

For PARP1 pyrosequencing data, Wilcoxon rank sum test was used to test for differential methylation between cases and controls. Analysis of variance (ANOVA) was used to test for statistical significance of between cases and control groups for within-group DNA methylation variance. Spearman correlation tests were used to test for significance of correlations between DNA methylation and TTD.

5.2.1.4 PCR and pyrosequencing

We selected one of the top 8 hits mapping to the PARP1 gene for further validation. Pyrosequencing of the PARP1 739bp DMR fragment (NCBI136/hg18 genomic location chr1: 224662084-224662823) in the validation sample set was performed with multiple assays, tiling a 530bp fragment (NCBI136/hg18 genomic location chr1: 224661898-224662428) spanning from base 186bp upstream of the fragment to 395bp into the fragment using six pyrosequencing assays (labelled PARP1s1-PARP1s6) within two PCR amplicons, with each pyrosequencing assay measuring 3-8 CpG sites (primer sequences shown in table 5.2 (A), schematic diagram of PARP1 primer-binding sites Fig 5.1). Performance of assays was variable; however, all six regions were pyrosequenced in order to achieve greater coverage of the DMH fragment. Methylation values were calculated as an average of all CpG sites within a given assay, that were labelled either ‘pass’ or ‘check’, based on the Pyromark pyrosequencing quality controls. ‘Check’ data, for which quality control thresholds were not reached (though the samples were not failed), were used in order to optimise sample numbers, as all assays displayed appropriate methylation values for 100% methylated and
0% methylated controls (i.e. >95% and <5%, respectively). The numbers of samples included in validation analysis was n=153 cases and n=198 controls, after exclusion of NAs, as shown in table 5.2. Three samples with methylation data that representing extreme statistical outlier (>40% methylation, >99.99% confidence intervals) were removed from analyses, due to potential measurement error or contamination with the positive control DNA.
**Table 5.2 Primers for chapter 5**

**A. PARP1 primers**

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence (5’ – 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PARP1 Forward 1</strong></td>
<td>biotin-GGTTAGGTATTAGTAATTATTAGGGAA</td>
</tr>
<tr>
<td><strong>PARP1 Reverse 1</strong></td>
<td>CACCTACACCATATAAACCATC</td>
</tr>
<tr>
<td><strong>PARP1 S1</strong></td>
<td>CTCTCRTCACATTTTCTTACAAAAAAA</td>
</tr>
<tr>
<td><strong>PARP1 S2</strong></td>
<td>CATCCTCCCCCTAACT</td>
</tr>
<tr>
<td><strong>PARP1 S3</strong></td>
<td>CCACRACCTAAAAACAC</td>
</tr>
<tr>
<td><strong>PARP1 Forward 2</strong></td>
<td>biotin-GGATGGTATTATTGTTAGGT</td>
</tr>
<tr>
<td><strong>PARP1 Reverse 2</strong></td>
<td>ACAACACCAACTACAAACTTTATTT</td>
</tr>
<tr>
<td><strong>PARP1 S4</strong></td>
<td>CAACATCAAACACACCT</td>
</tr>
<tr>
<td><strong>PARP1 S5</strong></td>
<td>CCCATAACCCCCAAAT</td>
</tr>
<tr>
<td><strong>PARP1 S6</strong></td>
<td>CCCTCCCCACACCT</td>
</tr>
</tbody>
</table>

**B. Primers for validation of loci associated with breast cancer risk in discovery study 2**

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence (5’ – 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEATR2 F</td>
<td>GTGGGAAAAATTAGGAATGGAATAAAA</td>
</tr>
<tr>
<td>HEATR2 R</td>
<td>biotin-ACCCCAAAAAAACAACAAAAACTAC</td>
</tr>
<tr>
<td>HEATR2 S</td>
<td>GGGTTTTGTGGGAGT</td>
</tr>
<tr>
<td>SLFN12L F</td>
<td>biotin-GTTTTGGGGAAGAGGTTG</td>
</tr>
<tr>
<td>SLFN12L R</td>
<td>CCTCRAACACTTCTATTTCTTCTTCTAT</td>
</tr>
<tr>
<td>SLFN12L S</td>
<td>AAAAACCCRAATACTCCAA</td>
</tr>
</tbody>
</table>
Fig 5.1

Fig 5.1. Genomic loci of PCR amplicons and pyrosequencing assays for validation of PARP1 differential methylation by pyrosequencing. Gene-structure image taken from UCSC genome browser (NCBI136/hg18). Green bar represents CpG island. Coloured arrows represent binding-sites of PCR primers, and white arrows represent binding-sites of pyrosequencing primers, with arrow direction indicating sequencing direction.
5.2.2 Discovery study 2 (Illumina 450k beadchip array)

5.2.2.1 Study samples

Samples used for discovery study 2 were the EPIC4 subgroup described in chapter 4 prior to removal of samples overlapping with EPIC 5 (see chapter 4), and will be referred to in this chapter as EPIC4. These consisted of prospectively collected white blood cell DNA samples from 92 invasive breast cancers cases and 92 healthy unrelated controls (without breast cancer at follow-up) (Table 5.3). Case-control pairs were matched on age (within 5 years), ethnicity and study centre. Sample preparation for the EPIC4 subgroup was as for other EPIC study subgroups described in chapter 2. The validation cohort for discovery study 2 (EPIC3) is as for chapter 2.
### Table 5.3. Discovery and validation sample sets for EWAS study using the 450k beadchip array

<table>
<thead>
<tr>
<th></th>
<th>Discovery microarray cohort [EPIC4]</th>
<th>Validation pyrosequencing cohort [EPIC3]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case</td>
<td>Control</td>
</tr>
<tr>
<td>N</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>Family history (yes)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Age at blood draw in years, mean (range)</td>
<td>51 (34-62)</td>
<td>50 (34-63)</td>
</tr>
<tr>
<td>TTD* in months (range)</td>
<td>55 (24-108)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*Time to diagnosis (from blood draw)
5.2.2.2 Illumina Infinium HumanMethylation 450k beadchip array

The Illumina Infinium HumanMethylation 450k beadchip array (Illumina 450k array) is described in chapter 4, materials and methods.

5.2.2.3 Statistical analysis (450k array)

Background-subtracted $\beta$-values were extracted from genome studio, and normalised using quantile normalisation. Conditional logistic regression was used to identify loci significantly associated with case control status, adjusting for age and experimental batch (array chip), with samples matched for age, study centre and ethnicity. Loci associated with case/control-status, with $p<10^{-5}$ were considered statistically significant, with loci $p<10^{-7}$ reaching genome-wide significance. Wilcoxon rank sum test was used to test associations between DNA methylation and cancer risk in the validation set, and Interquartile analysis was conducted using multinomial logistic regression models to calculate an OR for breast cancer risk for individuals within each methylation quartile compared with individuals within the lowest (reference) methylation quartile, with quartiles specified using controls only.

5.2.2.4 PCR and Pyrosequencing.

Two genes from the top hits were selected for further validation, HEATR2 (cg11027456) and SLFN12L (cg21705506). PCR and pyrosequencing was carried out as described in chapter 2, however, rather than using a universal biotinylated primers, one of the sequence-specific (forward or reverse) primers was biotinylated, depending on the DNA strand being sequenced (primer sequences shown in table 5.2 (B)). For HEATR2, 12 cases and 5 controls failed pyrosequencing quality control, leaving 231 cases and 281 controls for further analysis. For SLFN12L, 9 cases and 7 controls failed leaving 234 cases and 279 controls.
5.3 Results

5.3.1 Differential Methylation Hybridization

Thirty eight genes were found to be significantly differentially methylated between cases and controls with FDR corrected p-values<0.01. Of these, 14 loci showed an increase in methylation whereas the remaining 24 displayed a loss of methylation in cases relative to controls. Of the loci with significant DMH ratios, those with the greatest absolute magnitude of difference (highest DMH ratio and FDR p-value<0.01) were FOXO3 (1.56, p=1.322E-05), PARP1 (1.25, p=2.054E-05), TLE4 (1.10, p=1.32E-05), TGFB2 (1.08, p=2.054E-05), LBH (1.04, p=1.122E-06), PRR5 (1.01, p=3.14E-05), FZD3 (0.93, p=2.05E-05), SMAD1 (0.90, p=7.76E-07), CXCL2 (0.89, p=2.27E-06) and HIF1α (0.79, p=6.121e-06) (Fig 5.2, Table 5.4).

All of these loci were located at gene promoters, and were hypermethylated in cases. Consistently, by aligning all significantly differentially methylated genes at transcription start -sites and mapping methylation variability (standard deviation of methylation in cases and controls separately) relative to genomic location, a focus of increased methylation variability close to TSS, relative to the rest of the gene, may be clearly seen in cases compared to controls (Fig 5.3). DMH revealed a differentially methylated 740bp locus (DMH ratio 1.25) surrounding the Poly (ADP) ribose polymerase 1 (PARP1) gene transcription start site (TSS). This PARP1 locus was prioritised for pyrosequencing-based validation as it had the most significant p-value, second highest absolute magnitude of methylation difference between cases and controls, the well established biological role of PARP1 in breast cancer (De Soto et al., 2006), and the feasibility of validation based on fragment length.
Table 5.4. Top ten DMH fragments most significantly associated with breast cancer case-control status

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>Fragment start site*</th>
<th>Fragment end site**</th>
<th>Fragment length</th>
<th>Magnitude of difference***</th>
<th>P-value (FDR corrected)****</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOXO3</td>
<td>6</td>
<td>108985118</td>
<td>108985861</td>
<td>744</td>
<td>1.562920707</td>
<td>1.32E-05</td>
</tr>
<tr>
<td>PARP1</td>
<td>1</td>
<td>224662084</td>
<td>224662823</td>
<td>740</td>
<td>1.248197134</td>
<td>5.40E-07</td>
</tr>
<tr>
<td>TLE4</td>
<td>9</td>
<td>81377472</td>
<td>81378298</td>
<td>827</td>
<td>1.098937779</td>
<td>1.32E-05</td>
</tr>
<tr>
<td>TGFBR2</td>
<td>3</td>
<td>30622624</td>
<td>30624223</td>
<td>1600</td>
<td>1.084128379</td>
<td>2.05E-05</td>
</tr>
<tr>
<td>LBH</td>
<td>2</td>
<td>30307336</td>
<td>30308069</td>
<td>734</td>
<td>1.0431579</td>
<td>1.12E-06</td>
</tr>
<tr>
<td>PRR5</td>
<td>22</td>
<td>43476820</td>
<td>43477591</td>
<td>772</td>
<td>1.012598372</td>
<td>3.14E-05</td>
</tr>
<tr>
<td>FZD3</td>
<td>8</td>
<td>28407046</td>
<td>28408099</td>
<td>1054</td>
<td>0.933635867</td>
<td>2.05E-05</td>
</tr>
<tr>
<td>SMAD1</td>
<td>4</td>
<td>146622062</td>
<td>146623572</td>
<td>1511</td>
<td>0.900790693</td>
<td>7.76E-07</td>
</tr>
<tr>
<td>CXCL2</td>
<td>4</td>
<td>75183210</td>
<td>75184256</td>
<td>1047</td>
<td>0.890253405</td>
<td>2.27E-06</td>
</tr>
<tr>
<td>HIF1A</td>
<td>14</td>
<td>61231250</td>
<td>61232396</td>
<td>1147</td>
<td>0.791756546</td>
<td>6.12E-06</td>
</tr>
</tbody>
</table>

* fragment start site: genomic location of fragment start (UCSC genome browser build (NCBI136/hg18))
** fragment end site: genomic location of fragment end (UCSC genome browser build (NCBI136/hg18))
*** Magnitude of difference in methylation between case and control sample groups.
**** Wilcoxon-rank sum test p-value for significance of differential methylation (case/control).
Figure 5.2. Heat-map of Differential methylation Hybridisation (DMH) microarray top hits. Relative methylation (Z-scores) at all loci (n=38) that were differentially methylated (p<0.01 and magnitude of DMH ratio>0.25) between breast cancer cases (Red) and healthy controls (green). Unsupervised hierarchical clustering identified two main clusters of gene fragments, including fragments hypermethylated in cases relative to controls (right-arm) and fragments hypomethylated in cases relative to controls (left arm).
Figure 5.3. Methylation variability at all genes analysed by DMH, mapped in relation to genomic locus. Methylation variability (smoothed average of standard deviation of methylation) at all genes relative to distance from transcription start site (0), separated by case (blue) control (red) status of DNA samples.
Validation of the PARP1 DMH fragment was carried out by pyrosequencing of six regions within the DMH region labelled PARP1s1-PARP1s6, each spanning 3-9 CpG sites (fig 5.1) and covering 54 CpG sites across the whole CpG island. Technical validation of PARP1 hypermethylation in cases was carried out by pyrosequencing of the majority of the samples used for the DMH array (n cases=13, controls=24) (Fig 5.4). Four loci were significantly hypermethylated in cases relative to controls, including PARP1s1 (median control 2.67% vs. median case 3.8%, p=0.02) PARP1s2 (median control 2.06% vs. median case 3.96%, p=0.002), PARP1s4 (median control 3.76% vs. median case 5.69%, p=0.01), and PARP1s5 (median control 3% vs. median case 7.64%, p=0.003), whereas one locus was significantly hypomethylated in cases (PARP1s3 (median control 1.64% vs. median case 1.21%, p=0.04). Biological validation within the remainder of the KConFab study sample set (n=199 controls, 158 cases) was confirmed for PARP1s2 (median control 2.5% vs. Median case 2.98%, p=0.0000132), PARP1s4 (median control 3.55% vs. median case 4.34%, p=9.37E-11), PARP1s5 (median control 1.89% vs. median case 4.29%, p=9.52E-12) and PARP1s6 (median control 1.99% vs. median case 2.55%), whereas PARP1s1 and PARP1s3 were not associated with case-control status (Fig 5.4). Linear regression indicated that association of PARP1 methylation with case-control status remained significant after adjusting for experimental batch (plate). Consistent with the finding of methylation hypervariability at gene promoters on the DMH array, ANOVA indicated that PARP1 methylation was hypervariable (within-group methylation variance was significantly higher) in cases compared with controls for PARP1s1 (p=0.04), PARP1s2 (p=0.00013), PARP1s4 (p=0.006) and PARP1s5 (p=0.0007) in the technical validation set, and for PARP1s2 (p=1.32E-5), PARP1s4 (p=5.12E-9), PARP1s5 (p=5.21E-12) and PARP1s6 (p=1.11E-8) in the biological validation set. Whereas magnitudes of methylation difference between cases and controls were small (1-3%), the
consistency of the association of four regions, and between technical and biological validation sample sets using different assays suggests a genuine association.
Figure 5.4. Validation of PARP1 hypermethylation by bisulphite pyrosequencing of three regions of the PARP1 promoter; (i) PARP1s1, (ii) PARP1s2, (iii) PARP1s3. (A) Technical validation i.e. pyrosequencing in the samples used for DMH discovery (n cases=13, n controls=24). (B) Biological validation of the same regions, i.e. Pyrosequencing in an independent sample set from the KConFab familial breast cancer study (n cases=158, n controls=199).
Figure 5.4. Validation of PARP1 hypermethylation by bisulphite pyrosequencing (continued from previous page) of three regions of the PARP1 promoter; (i) PARP1s4, (ii) PARP1s5, (iii) PARP1s6. (A) technical validation i.e. pyrosequencing in the samples used for DMH discovery (n cases=13, n controls=24). (B) Biological validation of the same regions, i.e. Pyrosequencing in an independent sample set from the KConFab familial breast cancer study (n cases=158, n controls=199).
Further validation of hypermethylation of the PARP1s4, PARP1s5 and PARP1s6 regions in an independent sample set was attempted using the BGS sample cohort described in chapter 2 (cases n=258, controls n=259). However, there was no evidence of differential methylation between cases and controls at any of the three loci, S4 (2.7% v 3.0% P=0.07), S5 (2.4% v 2.6%, P=0.43), S6 (1.5% v 1.6%, P=0.2) (Fig 5.5).

We hypothesised that the difference in results between the KConFab and BGS studies may be due to association of PARP1 DNA methylation with subtype or molecular-pathway-specific features that may be more common in familial breast cancer (KConFab) than in sporadic cancer (BGS), especially as the KConFab study was enriched for BRCA mutation carriers. However, there was no evidence of methylation associated with case status at any of these loci within individuals in the BGS study that reported family history of breast cancer (n=66 cases, 51 controls), S4 (2.7% v 2.6%, P=0.6358), S5 (2.3% v 2.4%, P=0.5), S6 (1.5% v 1.5%, P=0.24) (Fig 5.5). However, the criteria for family history of breast cancer for the BGS study (>1 affected relative) was less stringent than for KConFab (all participants belong to families with breast cancer susceptibility mutations or several affected relatives).

Consistently, PARP1 was not differentially methylated in association with family history of breast cancer, irrespective of case-control status, in BGS data (data not shown). Furthermore, within the KConFab study, PARP1 methylation was not clearly associated with BRCA1/BRCA2 germline mutation status, or tumour features including expression of ER, PR and HER2, or triple negativity, for samples in which follow-up tumour pathological feature data was available (Table 5.5).
Figure 5.5. Bisulphite pyrosequencing of three regions of PARP1: (i) PARP1s4, (ii) PARP1s5, (iii) PARP1s6 in prospectively collected WBC DNA samples of (A) 258 invasive breast cancer cases and 259 matched healthy unrelated healthy controls from the Breakthrough Generations Study (BGS), and (B) a subset of BGS cases (n=66) and controls (n=51) for which family history of breast cancer was reported.
Table 5.5. Assessment of association of blood PARP1 DNA methylation with tumour pathological features in the KConFab study.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Pathological feature</th>
<th>N (+)*</th>
<th>N (-)</th>
<th>Mean methylation (+)</th>
<th>Mean methylation (-)</th>
<th>Median methylation (+)</th>
<th>Median methylation (-)</th>
<th>Wilcoxon p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARP1s4</td>
<td>BRCA1/2</td>
<td>47</td>
<td>81</td>
<td>4.179</td>
<td>4.528</td>
<td>3.94</td>
<td>4.05</td>
<td>0.02334</td>
</tr>
<tr>
<td>PARP1s5</td>
<td>BRCA1/2</td>
<td>10</td>
<td>22</td>
<td>3.107</td>
<td>3.696</td>
<td>2.96</td>
<td>3.28</td>
<td>0.4157</td>
</tr>
<tr>
<td>PARP1s6</td>
<td>BRCA1/2</td>
<td>46</td>
<td>83</td>
<td>2.965</td>
<td>3.025</td>
<td>2.32</td>
<td>2.58</td>
<td>0.5455</td>
</tr>
<tr>
<td>PARP1s4</td>
<td>ER</td>
<td>59</td>
<td>34</td>
<td>4.236</td>
<td>4.241</td>
<td>3.86</td>
<td>3.935</td>
<td>0.7466</td>
</tr>
<tr>
<td>PARP1s5</td>
<td>ER</td>
<td>15</td>
<td>9</td>
<td>3.969</td>
<td>3.01</td>
<td>4.15</td>
<td>2.97</td>
<td>0.07319</td>
</tr>
<tr>
<td>PARP1s6</td>
<td>ER</td>
<td>61</td>
<td>30</td>
<td>3.19</td>
<td>2.989</td>
<td>2.67</td>
<td>2.375</td>
<td>0.4009</td>
</tr>
<tr>
<td>PARP1s4</td>
<td>PR</td>
<td>47</td>
<td>46</td>
<td>4.223</td>
<td>4.253</td>
<td>3.84</td>
<td>3.93</td>
<td>0.5933</td>
</tr>
<tr>
<td>PARP1s5</td>
<td>PR</td>
<td>13</td>
<td>11</td>
<td>3.964</td>
<td>3.19</td>
<td>4.15</td>
<td>2.97</td>
<td>0.1674</td>
</tr>
<tr>
<td>PARP1s6</td>
<td>PR</td>
<td>51</td>
<td>39</td>
<td>3.209</td>
<td>3.021</td>
<td>2.67</td>
<td>2.42</td>
<td>0.4039</td>
</tr>
<tr>
<td>PARP1s4</td>
<td>HER2</td>
<td>23</td>
<td>61</td>
<td>4.045</td>
<td>4.32</td>
<td>3.54</td>
<td>3.93</td>
<td>0.2026</td>
</tr>
<tr>
<td>PARP1s5</td>
<td>HER2</td>
<td>12</td>
<td>6</td>
<td>4.327</td>
<td>3.408</td>
<td>4.015</td>
<td>3.355</td>
<td>0.3355</td>
</tr>
<tr>
<td>PARP1s6</td>
<td>HER2</td>
<td>21</td>
<td>57</td>
<td>3.05</td>
<td>3.333</td>
<td>2.67</td>
<td>2.62</td>
<td>0.7825</td>
</tr>
<tr>
<td>PARP1s4</td>
<td>TNBC</td>
<td>23</td>
<td>65</td>
<td>4.197</td>
<td>4.303</td>
<td>3.93</td>
<td>3.91</td>
<td>0.9848</td>
</tr>
<tr>
<td>PARP1s5</td>
<td>TNBC</td>
<td>4</td>
<td>18</td>
<td>3.132</td>
<td>3.831</td>
<td>3.065</td>
<td>3.39</td>
<td>0.342</td>
</tr>
<tr>
<td>PARP1s6</td>
<td>TNBC</td>
<td>15</td>
<td>71</td>
<td>3.169</td>
<td>3.179</td>
<td>2.42</td>
<td>2.67</td>
<td>0.8555</td>
</tr>
</tbody>
</table>

*positive for ER or PR expression, HER2 amplification or triple-negative breast cancer (TNBC)

*Wilcoxon rank sum test p-value
Lastly, we investigated whether DNA methylation at PARP1 regions significantly associated with cancer risk were associated with time to diagnosis (TTD) in KConFab. Interestingly, there was a significant inverse correlation between TTD and DNA methylation of the PARP1s2 \((r^2=0.22, \ p=0.03)\) and PARP1s6 \((r^2=0.2, \ p=0.03)\) regions, though not the PARP1s4 \((r^2=0.06, \ p=0.69)\) and PARP1s5 \((r^2=-0.19, \ p=0.13)\) regions (Fig 5.6), suggesting that blood PARP1 methylation may represent a marker of early preclinical disease.
Figure 5.6. Association of PARP1 DNA methylation with time to diagnosis. Scatter plots showing correlation of DNA methylation (Y-axes) at (A) PARP1s2, (B) PARP1s4, (C) PARP1s5 and (D) PARP1s6 with time in months between blood draw and breast cancer diagnosis (TTD)(X-axis). Red line represents regression line for association between DNA methylation and TTD. Rsq ($r^2$) and p-values shown for Spearman correlation between methylation and TTD.
5.3.2 Illumina 450k beadchip array discovery study

A second discovery study was conducted using the Illumina Infinium HumanMethylation 450k beadchip array, and including 92 matched case-control pairs from the EPIC study (Table, 5.3 (EPIC4)). Conditional logistic regression was used to identify mean methylation differences between cases and controls, with samples matched on age, ethnicity and study centre. Seventy five loci were nominally associated with case-control status at a threshold of $p<10^{-5}$, however, no locus reached genome-wide significance level ($p<10^{-7}$) (Fig 5.7). Of these, 65 displayed an increase in methylation in cases relative to controls, of which 46 loci displayed a $\Delta\beta>0.05$ (5% higher mean methylation in cases), with the largest $\Delta\beta$ of 0.091 at the GABRQ gene. The largest decrease in methylation in cases relative to controls was within the ASB13 gene ($\Delta\beta = 0.04$). Only at one gene (TFDP3), was there more than one locus (cg16829640 and cg13566059 (adjacent loci, 5bp apart) with $\Delta\beta$-values of 0.07 and 0.086, respectively) identified. TFDP3 is on the X chromosome, and these loci display unusually broad methylation ranges consistent with hypermethylation of the inactivated chromosome (as all participants are female), and methylation ranging between 0-100% on the active allele (Fig 5.8). Methylation at the TFDP3 cg16829640 was strongly correlated with methylation of TFDP3 cg13566059 ($r^2=0.85$, $p= 2.2e-16$).
Figure 5.7. Manhattan plot showing statistical significance of associations between DNA methylation and breast cancer risk for all loci represented on the Infinium HumanMethylation 450k beadchip array, arranged by chromosome (coloured bands, X-axis). Statistical significance (conditional logistic regression) represented as –log10p-values (Y-axis). Grey lines represent cut-offs for statistical significance at p<10^{-5} and p<10^{-7} levels.
Figure 5.8. Significant blood hypermethylation of neighbouring CpG sites in TFDP3. (A) TFDP3_cg16829640 and (B) TFDP3_cg13566059 in breast cancer cases (1, n=92) relative to controls (0, n=92) in Illumina 450k beadchip array data (EPIC4 subgroup). Red lines represent median methylation.
Two loci, within the HEATR2 (cg11027456) and SLFN12L (cg21705506) genes were selected for validation, and selection was based on lowest p-values, highest Δβ-values (0.063 and 0.09 respectively), and ability to design working pyrosequencing assays (Fig 5.9, Table 5.6 (A)).

Pyrosequencing of HEATR2 and SLFN12L was carried out in an independent data set of EPIC samples (EPIC3, n=243 cases, 286 controls). SLFN12L methylation was not significantly different in cases and controls, and quartile analysis did not reveal a significant difference in breast cancer risk in higher methylation categories compared with the lowest (reference category) (Fig 5.9, Table 5.6 (B)). Inconsistent with the discovery set, HEATR2 was significantly hypomethylated in cases compared with controls (p=0.005), i.e. the direction of effect was opposite to the discovery set, however, the difference in median methylation between cases (median=86.6%) and controls (median=87.5%), was very small (0.9%). Quartile analysis indicated significantly decreased cancer risk in the highest (OR=0.48 (95% CI 0.29-0.80), p=0.005) and second highest (OR=0.62 (0.38-1.00), p=0.05) methylation quartiles compared with the lowest methylation quartile (Table 5.6 (B)). As with analysis of fasting-associated loci using the 450k array (chapter 4), methylation ranges of HEATR2 and SLFN12L were much higher and narrower in the pyrosequencing validation set data compared with the array-based discovery set data, indicating a consistent unidentified bias in our analysis pipeline.
Figure 5.9. Association of loci within SLFN12L and HEATR2 with breast cancer case-control status. (A) 450k array DNA methylation data for top hit loci within SLFN12L and HEATR2 in controls (n=92) and cases (n=92) in the EPIC study (EPIC4 subgroup). (B) Pyrosequencing in an independent sample set (n=286 controls, 243 cases) from the EPIC study (EPIC3 subgroup).
Table 5.6 (A). Median, mean and inter-quartile range (IQR) of DNA methylation of controls and cases for HEATR2 and SLFN12L (pyrosequencing data from validation sample set (EPIC3)).

<table>
<thead>
<tr>
<th>Locus</th>
<th>Mean control</th>
<th>Mean case</th>
<th>Median control</th>
<th>Median case</th>
<th>IQR control</th>
<th>IQR case</th>
<th>Wilcox p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEATR2</td>
<td>87</td>
<td>86.28</td>
<td>87.51</td>
<td>86.62</td>
<td>4.33</td>
<td>4.64</td>
<td>0.005</td>
</tr>
<tr>
<td>SLFN12L</td>
<td>75.43</td>
<td>75.33</td>
<td>75.83</td>
<td>75.56</td>
<td>8.84</td>
<td>9.27</td>
<td>0.8</td>
</tr>
</tbody>
</table>

*Wilcoxon rank sum test p-value

Table 5.6 (B). Inter-quartile analysis of association between methylation and cancer incidence at HEATR2 and SLFN12L (pyrosequencing data from validation sample set (EPIC3)).

<table>
<thead>
<tr>
<th>Locus</th>
<th>Quartile (meth range)</th>
<th>case (n)</th>
<th>control (n)</th>
<th>OR (95% CI)*</th>
<th>p **</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEATR2</td>
<td>Q4 (89.4-95.0)</td>
<td>41</td>
<td>73</td>
<td>0.48 (0.29-0.80)</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>Q3 (87.5-89.4)</td>
<td>50</td>
<td>70</td>
<td>0.62 (0.38-1.00)</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Q2 (85.1-87.5)</td>
<td>66</td>
<td>70</td>
<td>0.81 (0.51-1.29)</td>
<td>0.385</td>
</tr>
<tr>
<td></td>
<td>Q1 (53.5-85.1)</td>
<td>83</td>
<td>72</td>
<td>reference</td>
<td>_</td>
</tr>
<tr>
<td>SLFN12L</td>
<td>Q4 (80.1-95.8)</td>
<td>73</td>
<td>73</td>
<td>1 (0.63-1.58)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Q3 (75.8-80.1)</td>
<td>61</td>
<td>74</td>
<td>0.82 (0.51-1.32)</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>Q2 (71.3-75.8)</td>
<td>73</td>
<td>69</td>
<td>1.05 (0.66-1.68)</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Q1 (45.8-71.3)</td>
<td>73</td>
<td>73</td>
<td>reference</td>
<td>_</td>
</tr>
</tbody>
</table>

*Odds ratio (95% confidence interval) for breast cancer incidence in highest, compared with the reference (lowest) methylation category.

**p-value for logistic regression
5.3.3 Analysis of DMH candidate loci in EPIC4 450k array data

450k array DNA methylation data was available for several loci located within the fragments associated with cancer risk in the DMH discovery study; enabling potential validation of DMH associated regions in an independent sample set using a different methylation microarray analysis method. Median methylation levels (control and case) and p-values (conditional linear regression, before and after FDR correction) for association with cancer risk are shown for all of the 450k array loci residing within the ten most significant DMH fragments, are shown in Appendix table 2. Of 87 450k array loci within DMH fragments, only 5 were significantly associated with case-control status with the same direction of effect in EPIC4. All of these displayed very small median methylation differences between case and control groups (<1%), and three loci were significantly differentially methylated in the 450k array data with the opposite direction of effect to the DMH study, and with case-control methylation differences of 0.5-2%; therefore, the loci associated with breast cancer risk in both studies may have occurred due to chance and are unlikely to be of biological relevance. Seven 450k loci were within the PARP1 DMH fragment, however, only three were within regions analysed by pyrosequencing in KConFab, and two were within the PARP1s1 region that was not differentially methylated in KConFab. Only one CpG site within the PARP1s6 region cg07113065 was, therefore, comparable with the KConFab data. This individual CpG sites was significantly hypermethylated in KConFab data (p<2.2e-16), combined technical and biological validation sets (data not shown), but not in the 450k array EPIC data set (p=0.42, Table 5.7), consistent with lack of association with sporadic breast cancer risk.
5.4 Discussion

5.4.1 Differential methylation hybridization

Our DMH study identified 38 genes that were differentially methylated in prospectively collected blood samples of familial breast cancer cases compared with healthy unrelated controls, with an enrichment of DNA methylation variability at gene promoters relative to other genomic regions. The top ten most differentially methylated fragments were hypermethylated in cases relative to controls, which is inconsistent with evidence for greater DNA methylation variability at intergenic, than promoter, gene regions (Flanagan and Wild, 2007; Shenker and Flanagan, 2012). Promoter methylation hypervariability within this study may reflect an increased promoter hypermethylation frequency associated with familial breast cancer risk; however, our approach is biased by the selection of gene promoters on the custom oligonucleotide array. Promoter methylation hypervariability is thought to occur due to loss of protection from DNA methylation at CpG islands a frequent event during early tumorigenesis (Baylin et al., 1998).

Technical validation and biological validation within the same sample study were achieved for promoter regions of the PARP1 promoter, directly upstream (PARP1s2) and directly downstream (PARP1s4, PARP1s5 and PARP1s6) of the TSS. Significant hypermethylation and methylation hypervariability in cases was identified, though the increase in percentage methylation was subtle. Consistency of these associations between samples sets (discovery and biological validation sets) and assay types (DMH and bisulphite pyrosequencing (including four pyrosequencing assays within two independent PCR amplicons)), indicates that PARP1 promoter hypermethylation is genuinely associated with breast cancer risk within the KConFab study.
The finding that hypermethylation of PARPs2 and PARPs6 was inversely correlated with Time to diagnosis in KConFab suggests that \textit{PARP1} hypermethylation may represent an early tumorigenic event rather than an innate disease risk factor. This result is similar to the report by Xu et al (Xu et al., 2013b), in which methylation at many loci associated with breast cancer risk was significantly correlated with TTD.

\textit{Poly (ADP) ribose polymerase 1 (PARP1)} is a key sensor in the DNA damage repair pathway, in which most known breast cancer susceptibility genes are involved (Burkle, 2000; van der Groep et al., 2011). \textit{PARP1} is of clinical interest for familial breast cancer, as PARP inhibitors induce synthetic lethality in tumour cells with defective DNA damage repair due \textit{BRCA1} mutations (Berrada et al., 2010; O'Shaughnessy et al., 2011). Furthermore, PARP inhibitors could be used in first line treatment of \textit{BRCA1} mutated or triple-negative breast cancers (breast cancers lacking expression of ER and PR, and lacking amplification of HER2, and therefore associated with poor prognosis due to lack of molecular targets for treatment). \textit{PARP1} is also over-expressed in triple-negative breast cancers (Ossovskaya et al., 2010). Little is known about epigenetic regulation of \textit{PARP1}, however, some evidence of promoter hypermethylation associated with environmental carcinogen exposure \textit{in vitro} has been reported (Gao et al., 2010), and \textit{PARP1} plays a role in regulation of DNA methylation, probably through regulation of DNMT1 (Beneke, 2012; Guastafierro et al., 2008; Reale et al., 2005; Zampieri et al., 2009).

Given that the \textit{PARP1} promoter was almost completely unmethylated in controls in our data, the increase in cases likely represents an aberrant hypermethylation of a small subset of leukocytes. As hypermethylation occurs at the TSS, it is likely to affect transcription, which may lead to aberrant DNA damage repair (DDR) in \textit{PARP1} hypermethylated cells (Bouchard
et al., 2003; Guastafierro et al., 2008), which may contribute to cellular tumourigenicity potential. Although the increase in PARP1 methylation in cases is subtle (0.48-2.4%), this equates to a 1.2-2.26-fold increase in the frequency of PARP1 hypermethylated cells considering that methylation in individual cells is binary, either methylated or unmethylated. Furthermore, given the technical validation of PARP1, promoter hypermethylation of several other breast cancer-related transcription factors in the DMH data, including FOXO3 (Zou et al., 2008), TGFBR2 (Lucke et al., 2001) and HIF1A (Chiavarina et al., 2010) may be reproducible, at least within the KConFab study. This may be indicative of an underlying defect in the factors protecting the unmethylated state of gene promoter CpG islands in familial breast cancer. A ‘CpG methylator phenotype’ is evident in many cancers, whereby hypermethylation of multiple gene-promoters that are unmethylated in normal tissue occurs (Issa, 2004), therefore, increased frequency of gene promoter hypermethylation in blood of familial breast cancer cases may represent a less-pronounced precursor state for this phenomenon. It is important to remember, however, that our candidate-gene approach selected for genes known to be hypermethylated in breast cancer.

PARP1 was not differentially methylated with case-control status in the BGS study, using pyrosequencing, nor was there any evidence of differential methylation of PARP1 promoter loci within the 450k array data in EPIC. Importantly, the BGS and EPIC studies represent nested case control studies, with most cases representing sporadic breast cancer, whereas all cases within the KConFab have strong family history of breast cancer. Therefore, heritable risk factors are likely to account for many KConFab cancer incidences, whereas environmental or stochastic factors may account for sporadic cases in the BGS and EPIC studies. We did not find evidence of PARP1 hypermethylation within a subset of breast cancer cases or controls for which family history of breast cancer was self-reported in BGS;
however, the criteria for familial breast cancer were less stringent for these individual compared with the KConFab study.

There is increasing evidence that breast cancer risk factors can be disease subtype-specific (Tamimi et al., 2012), and genetic loci specifically associated with ER- disease have recently been identified (Garcia-Closas et al., 2013a). Epigenetic risk factors may also be subtype specific, as blood BRCA1 hypermethylation was associated with development of BRCA1-hypermethylated breast cancer in two retrospective studies (Iwamoto et al., 2010; Wong et al., 2010). We investigated whether the association of PARP1 hypermethylation with familial, but not sporadic breast cancer could be due to association of PARP1 methylation with BRCA mutations or tumour markers that may be indicative of subtype-specific tumorigenic pathways, however, we did not find convincing evidence of this.

A considerable proportion (60%) of familial breast cancer cases occur in the absence of any known susceptibility mutation (van der Groep et al., 2011). It is possible, therefore, that epimutations or epigenetic abnormalities, such as a higher propensity for promoter hypermethylation, may account for some familial cancers (Hitchins et al., 2011; Hitchins et al., 2007b). While it is generally accepted that DNA methylation and epimutations are erased and re-established between generations, underlying genetic factors may mediate transgenerational inheritance of pathogenic epigenetic states (Hitchins and Ward, 2007a; Imai and Yamamoto, 2008; Ligtenberg et al., 2012). Genetic polymorphism in DNMT3b may be associated with risk of breast (Cebrian et al., 2006; Montgomery et al., 2004), prostate (Singal et al., 2005) and lung cancer (Lee et al., 2005b), and DNMT polymorphisms may influence blood DNA methylation patterns (Inoue-Choi et al., 2013), therefore, genetic breast cancer risk factors influencing DNA methylation may account for the association PARP1 hypermethylation in familial, but not sporadic breast cancer. Alternatively, an unknown bias within the KConFab study (discovery and validation sets) may account for the association
with case control status. Investigation in an independent familial breast cancer study is warranted, and is required for ‘external’ validation (Garcia-Closas et al., 2013b).

5.4.2 Illumina Infinium HumanMethylation 450k beadchip array

Our discovery study using the Illumina Infinium HumanMethylayion450 BeadChip microarray represents one of the first prospective EWAS studies for breast cancer risk. Unfortunately, no loci reached ‘genome-wide’ statistical significance ($p<10^{-7}$) for association, indicating that it is unlikely that there are any loci represented on the 450k array that display large methylation differences associated with sporadic breast cancer risk, as have been identified for smoking (Shenker et al., 2012). Our EWAS study was not powered to identify subtle methylation variation between cases and controls (only 2% power to detect $d=0.5$ (~5% methylation case-control difference); 80% power to detect $d=0.93$ (~10% difference) at alpha=$1e^{-7}$). Whereas the greater DNA methylation variability in tumours allows investigators to limit analysis to loci displaying $\Delta\beta$-values of >0.2 (20%), which allows identification of differential methylation with 99% confidence (Bibikova et al., 2011; Wilhelm-Benartzi et al., 2013), investigation of blood DNA methylation in population studies requires methods that allow discrimination of subtle methylation variability.

Consistently, a recent prospective study conducted by Xu et al (Xu et al., 2013b) reported associations of 250 loci with breast cancer risk, however, all loci displayed small absolute $\Delta\beta$-values (<5% methylation difference) associated with breast cancer risk. Similarly, no cancer-associated loci identified in retrospective studies (Marsit et al., 2011; Teschendorff et al., 2009) displayed differential methylation of >6% in cases relative to controls.

The associations of HEATR2 and SLFN12L hypermethylation with case-control status likely represent false-positives, as the methylation distributions of both loci in pyrosequencing data
appeared very different from the Illumina 450 array data. Technical validation was not possible in this case, as the samples used for the array were not available for pyrosequencing. It is possible that the detected hypermethylation represented technical artefacts or were introduced during array data processing. Attempts to validate loci associated with both case-control and fasting status have shown a consistent pattern of narrower and higher methylation ranges in pyrosequencing data compared with array data, at some, but not all loci. It is not known whether this bias lies with the array step or the pyrosequencing step, however as bisulphite sequencing represents the gold standard for DNA methylation analysis (Reed et al., 2010), it is likely that artefactual associations are caused either by technical or data-processing steps of the 450k array analysis pipeline.

The EPIC4 Illumina 450k array data used for this case control analysis (following quality control and normalisation steps) has been previously used to identify loci associated with smoking, using similar analytical methods to the ones used to identify case-control associations (Shenker et al., 2012), and these loci were successfully validated by pyrosequencing in other EPIC data-sets. Therefore, technical validation of loci associated with DNA methylation differences of 2-20% in array data was achieved, providing proof-of-principal for investigation of EWAS studies using our analytical approach. Whereas the Δβ-values for some of the smoking-associated loci were lower than those for HEATR2 and SLFN12L associated with case-control status, the smoking associations were much more consistent than the case-control differences identified. However, technical factors may account (to some extent) for the greater reproducibility of smoking-associated loci than breast cancer-associated loci, as the quality and reproducibility of methylation analysis on the 450k array is locus-specific due to probe-specific biases such as DNA-bindings specificity and probe type (Price et al., 2013; Wilhelm-Benartzi et al., 2013).
Lastly, a caveat of using large sample sizes is that very strong statistical associations may occur in the absence of biologically meaningful differences (Burton et al., 2009); therefore, it is important to assess the reproducibility of associations in validation sample sets as well as their statistical significance (Garcia-Closas et al., 2013b).

5.4.3 Comparison of arrays methods

There was no apparent overlap between discovery studies in terms of loci associated with breast cancer risk, as only 5 of the 450k array loci within the DMH fragment regions were significantly associated with cancer risk in the same direction in the EPIC4 data set, and all of these loci displayed methylation differences between cases and controls of <1%. Both technical differences between array methods and biological differences between study populations may account for this.

Coverage of the DMH fragments on the 450k array is limited to 3-18 CpG sites within each fragment; therefore, the differentially methylated regions discovered by DMH may not be covered on the 450k array. For example, of seven loci within the PARP1 DMH fragment represented on the 450k array, only one was within a region that was differentially methylated in the KConFab samples. Different assay types identify differential methylation based on different criteria, as DMH measures methylation of genomic fragments ~100-200bp in length, enabling detection of small, but spatially consistent shifts in methylation frequency between sample groups. The Illumina 450k array measures methylation at single CpG sites; therefore, discovery of statistically significant differential methylation is dependent on highly consistent mean methylation differences at single CpG sites between cases and controls. Large sample sizes will be required to identify breast cancer-risk associated loci with modest Δβ-values and heterogeneity of effects. Meta-analyses of EWAS studies may
also be useful for identifying subtle differential methylation at individual loci across multiple studies.

Regional methylation shifts may be identified using the 450k array as consistent methylation differences across neighbouring probes where neighbouring probes are sufficiently near to each other to detect methylation correlation between neighbouring probes (Jaffe et al., 2012), as identified for two TFDP3 loci 5bp apart, and for haplotype-associated methylation (see chapter 4). Development of arrays with higher probe densities and bisulphite-sequencing-based methods will enable assessment of regional differential methylation patterns in EWAS studies.

The Illumina 450k beadchip array, however, has many practical advantages over DMH, as methylation analysis is more quantitative and is measured at single base resolution. Validation of single CpG sites is far more time and cost-effective than validation of genomic fragments, and is achieved with 100% coverage of the locus of interest; whereas full coverage of DMH fragments with pyrosequencing assays may often be impossible, especially at CpG dense gene promoters.

As previously mentioned, the inconsistency between our DMH and Illumina arrays may be partially due to biological differences between the KConFab and EPIC study sample-sets, as promoter hypermethylation may be more frequent in familial breast cancer compared with sporadic breast cancer. Consistently, Xu et al found a statistically significant enrichment of breast cancer risk-associated differential methylation at CpG islands, in a prospective investigation of breast cancer (Xu et al., 2013b). However, a prospective study investigating blood DNA methylation in association with sporadic breast cancer also reported enrichment of CpG island loci (Widschwendter et al., 2008).
In conclusion, the identification of *PARP1* hypermethylation associated with breast cancer risk supports the idea that prospective discovery studies may be used to identify additional DNA methylation biomarkers for breast cancer risk prediction. However, our inability to identify additional markers within the 450k array discovery study suggests that large sample sizes and greater understanding of potential confounding factors will be required to identify significant loci in EWAS studies.
CHAPTER 6

Overall discussion, limitations and future perspectives
While a small proportion of breast cancers are caused by high penetrance mutations, the majority are likely attributable to combinations of common heritable and environmental risk factors, as well as stochastic somatic mutations. Our research contributes to the wealth of emerging evidence indicating that many of these risk factors, such as age, estrogen exposure, obesity, carcinogen exposure, and common genetic polymorphisms have profound influences on epigenetic patterns, providing insight into the mechanisms by which these factors affect breast cancer risk on a molecular level. While this evidence provides promise for identification of epigenetic patterns associated with and accounting for breast cancer risk, it also indicates that unlike easily quantifiable exposures such as smoking or age, breast cancer risk represents a highly heterogeneous combination of factors that cannot be easily collectively quantified.

Combined with the low variability (relative to tumour tissue), tissue-specificity, population specificity, and susceptibility to alteration by exposures of DNA methylation, cancer risk heterogeneity greatly limits our ability to identify simple associations between DNA methylation and breast cancer risk. Furthermore, it is impossible to identify individuals who are truly at low risk of breast cancer in the absence of lifetime follow-up; therefore, the most appropriate control participants are unavailable. Complicating matters further are the known interactions between many risk factors such as the interaction between MTHFR genotype and serum folate in contributing to cancer risk (Liu et al., 2013), and the interaction between diet and genotype for obesity related genes (Kaklamani et al., 2011), therefore, many breast cancer risk factors will be context-dependent. Identifying key associations within such a complex systems is a major challenge facing breast cancer molecular epidemiology. The research presented in this thesis has identified some interesting associations with potential
implications for breast cancer risk and biomarker development, but also highlights some wider implications for EWAS studies and cancer risk epigenetics

6.2 ATM DNA methylation

From our analysis of ATMmvp2a methylation and breast cancer risk, the OR for increased breast cancer risk within the highest methylation quintile was 1.89 (95%CI 1.4-2.6). While this first candidate gene represents a higher risk-association than any breast cancer GWAS SNP identified to date (Michailidou et al., 2013), a biomarker of this relatively modest risk association would be unlikely to provide the sensitivity or specificity to predict risk independently. However, whether inclusion of this marker into a cancer risk model will improve the model performance is unknown.

Despite consistent association between ATMmvp2a methylation and breast cancer risk within the KConFab and BGS prospective studies, and in the bilateral breast cancer retrospective study (Flanagan et al., 2009b), we did not find a clear association within the EPIC study. This inconsistency may be explained by biological or demographic differences between study populations, such as age or ethnicity, or by technical factors such as sample preparation.

Given that several environmental factors appear to influence ATMmvp2a methylation, ATMmvp2a methylation is more likely to represent an intermediate marker of breast cancer risk-related exposures rather than an innate susceptibility factor. However, the mechanism underpinning the association with case control status remains unknown. A putative explanation is that ATMmvp2a hypermethylation may reflect increased blood glucose levels in individuals at higher breast cancer risk due to metabolic cancer risk factors.
Our analysis was limited by lack of availability of information on factors such as fasting status, kynurenic acid levels and \textit{MTHFR} genotype within the studies in which ATM\textsuperscript{mvp2a} methylation was associated with cancer risk. Further assessment of ATM\textsuperscript{mvp2a} methylation in breast cancer case control studies, ideally familial cancer studies, may help to establish the consistency of the risk association, and the factors accounting for this association.

A biological mechanism for the apparent specific demethylation of the ATM\textsuperscript{mvp2a} region in response to glucose depletion remains lacking. ATM\textsuperscript{mvp2a} may undergo active demethylation upon T cell activation or deactivation, as has been reported for other genes (Bruniquel and Schwartz, 2003; Toker et al., 2013; Zhang et al., 2013), as we found preliminary evidence that ATM methylation variability is restricted to T lymphocytes, and T cell activation is influenced by glucose metabolism (Finlay, 2013; Jacobs et al., 2008).

However, methylation analysis of ATM\textsuperscript{mvp2a} within specific blood cellular subsets, in association with fasting and glucose restriction, is required to investigate this, and to rule out the alternative hypothesis that the apparent ATM demethylation occurs due to proliferation or selective survival of T cells. Further research will also be required to determine the mechanism by which ATM becomes demethylated upon glucose depletion, and to investigate the possibility that active demethylation occurs. Such research is highly warranted, as our research suggests that T cell-specific ATM demethylation may represent one the first reported examples of environmentally mediated active demethylation \textit{in vivo}.

Furthermore, this research provides novel evidence of interactions between DNA methylation and metabolism of glucose and kynurenines, both of which relate to cancer risk (Cairns et al., 2011; Mandi and Vecsei, 2012; Seyfried et al., 2011; Stone et al., 2013). This is consistent with emerging evidence of widespread interactions between the epigenome and the metabolome (Donohoe and Bultman, 2012; Petersen et al., 2013). As breast cancer
is associated with distinct metabolic profiles detectable in serum (Oakman et al., 2011),
research into interactions in blood may uncover the epigenetic mechanisms by which
cancer-associated metabolic profiles influence gene expression.

6.3 Tissue-specific DNA methylation

Tissue-specificity of DNA methylation presents two key caveats for investigation of blood
DNA methylation biomarkers for cancer risk. Firstly, it is likely that whole blood DNA
methylation does not directly reflect methylation within the cancer affected tissue. This,
however, does not exclude the possibility that blood DNA contributes to cancer risk, as many
blood DNA methylation patterns may contribute to breast cancer through roles in immune
surveillance, inflammation, energy homeostasis, or other blood-related processes.
Assessment of the relevance of blood DNA methylation for breast cancer risk may be aided
by the increasing availability of DNA methylation data for normal breast epithelial tissue.

Secondly, cell-type-specific methylation within whole blood, coupled with inter-individual
variability in cell-type proportions within blood represent a common confounding factor for
association of DNA methylation with different exposures (Koestler et al., 2012; Reinius et al.,
2012), especially as immune profiles are altered by innate inter-individual differences,
exogenous exposures, and disease states (Koestler et al., 2012; Langevin and Kelsey,
2013). This issue is currently being addressed by cataloguing of tissue-specificity of DNA
methylation for all probes on the Illumina 450k array (Koestler et al., 2013a; Reinius et al.,
2012), and by development of statistical methods to adjust for cell count, which may be
measured prior to analysis (Koestler et al., 2013b).

Alternatively, methylation analysis may be carried out within specific blood cell populations
separately (Zhang et al., 2013). As it becomes increasingly apparent that blood methylation
variability relates strongly to immune functions (Koestler et al., 2012; Zhang et al., 2013),
investigation of methylation in specific immune cell-types may identify epigenetic patterns
associated with immune-related breast cancer risk factors such as inflammation and obesity
(Grivennikov et al., 2010).

6.4 Mechanisms of DNA methylation

A common assumption of epigenetics association studies is that DNA methylation variability
has mechanistic implication for gene transcription and genome function, yet the role of DNA
methylation across the genome has not been fully elucidated. DNA methylation appears to
have functions apart from transcriptional regulation, as several genome-scale studies have
indicated that DNA methylation in blood is not associated with gene expression at the
majority of loci (Eckhardt et al., 2006; Fraser et al., 2012; Illingworth et al., 2008; Lam et al.,
2012). Whereas the role of intragenic methylation remains unclear, the identification of
intragenic, and even intergenic loci, with smoking (Shenker et al., 2012), and of breast
cancer GWAS SNPs in distal enhancers (Ahmadiyeh et al., 2010), implicates intragenic
regions in cancer risk. ATMvp2a may represent an interesting model for further
investigation of the role of intragenic DNA methylation, especially with regard to potential
roles in alternative splicing. Currently, mechanistic studies and investigating causality of
differential methylation are limited by lack of methods to specifically alter epigenetic marks,
as can be done for genetics studies using targeted mutagenesis. Understanding of the role
of DNA methylation in cancer risk may benefit from investigation of interactions with other
epigenetic mechanisms such as histone acetylation or nucleosome positioning.
6.5 Epigenetic biomarkers based on response to cancer related interventions

ATMmvp2a methylation may represent an interesting marker of epigenetic response to environmental exposures, as there was evidence of inter-individual variability in response to fasting of methylation at the locus (albeit in only 6 individuals), and as markers of inter-individual variability in response to fasting have been postulated as potential biomarkers for metabolic health (Oliver et al., 2013).

Investigation of inter-individual variability in epigenetic response to cancer risk-related exposures and risk-modifying interventions, such as alcohol consumption, hormone-replacement therapy, and treatment with anticarcinogenic chemotherapies such as tamoxifen or metformin, may represent a more relevant approach to investigate epigenetic breast cancer risk markers, compared with single time-point 'snapshot' measures of methylation. Furthermore, such studies would be more easily controlled due to the measurement of quantifiable risk exposures, and due to the elimination of bias associated with inter-individual variability in ‘baseline’ methylation levels and factors such as age, genetics and exposures. Measurement of response to transient exposures, such as fasting or chemo-preventative therapy may be easily controlled for batch effect due to the short duration between collection of ‘before and after’ samples. Inclusion of intermediate markers, such as circulating estrogen levels, as well as standardised questionnaires with regard to cancer risk exposures, and follow-up for cancer incidence would also improve such studies.

A causal role for cancer risk exposure on the outcome (DNA methylation), often lacking in epidemiological studies, could be easily established by measurement at further time-points, such as before and after consecutive rounds of preventative therapy. Ideally, such studies would be incorporated into intervention studies for which the reduced risk associated with intervention has been quantified (Cuzick et al., 2002; Visvanathan et al., 2013).
6.6 genome-wide hypomethylation

Overall evidence suggests that genomic 5meC levels are lower in cancer patients compared with healthy individuals; however, the inconsistent findings between studies, and failure to validate this association in prospective studies (Huang et al., 2012) may suggest that the finding is not generally applicable, or is a consequence of, rather than a risk factor for, cancer. This is consistent with perhaps the most plausible putative cause of genome-wide hypomethylation; that supply of dietary methyl donors is affected by the presence of rapidly proliferating tumour cells within an individual (Liu et al., 2013). Whole-methylome analysis will likely improve our understanding of genomic methylation landscapes and the widespread and focal changes associated with cancer. Meanwhile, cost-effective mass-spectrometry-based methods for 5meC analysis requiring smaller amounts of DNA will allow high-throughput investigation of the utility of genomic 5meC levels for biomarker development.

Despite the well-cited role of DNA methylation in repressing repetitive element transcription, potentially protecting against retrotransposition-mediated mutagenesis, genomic instability (Belshaw et al., 2010; Friso et al., 2013; Soares et al., 1999; Wilson et al., 2007), and spurious RNA transcription from within repetitive elements, DNA methylation of individual genomically distinct repetitive elements remains virtually unexplored. Novel methods using specific methods should enable the investigation into epigenetic regulation of specific ‘functional’ L1 elements, such as those capable of retrotransposition (Cordaux and Batzer, 2009; Wolff et al., 2010), or driving ectopic expression of neighbouring genes (Faulkner et al., 2009; Wolff et al., 2010).
6.7 *PARP1* hypermethylation.

Further investigation of the potential association of *PARP1* hypermethylation with familial breast cancer risk, and the association of methylation with Time to diagnosis (TTD), is highly warranted, as there is a current requirement for biomarkers to improve sensitivity and specificity of breast cancer diagnosis at early stage (Garcia-Closas et al., 2013b). Association of *PARP1* hypermethylation with familial, but not sporadic breast cancer is consistent with a stronger association of ATMmvp2a methylation with cancer risk in the KConFab study compared with other studies. Whereas this may point to unidentified confounding factors within the KConFab study underpinning both associations, it may also indicate heterogeneity between breast cancer types in terms of epigenetic etiology, and of general epigenetic dysregulation associated with familial breast cancer risk. This is somewhat paradoxical, as epigenetic marks are not thought to be heritable, therefore, should not contribute to heritable disease risk. Despite this, twin studies (Gordon et al., 2012; Heijmans et al., 2007), and other population studies (Bjornsson et al., 2008; Dite et al., 2012; Kile et al., 2010; Mirabello et al., 2010), have shown that heritable factors do contribute considerably to DNA methylation genome-wide. Mounting evidence that SNPs in epigenetic modifier enzymes affect DNA methylation in *trans* (Inoue-Choi et al., 2013; Shukla et al., 2010; Tajuddin et al., 2013), and are associated with increased cancer risk (Cebrian et al., 2006; Kullmann et al., 2013; Lee et al., 2005b; Montgomery et al., 2004; Shukla et al., 2010), indicate that common heritable factors may associate with DNA methylation variability contributing to cancer risk. Furthermore, the potential role of cancer susceptibility genes such as BRCA1 in regulating DNA methylation (Shukla et al., 2010), suggests that potential interaction between epigenetic mechanisms and genetic risk factors for breast cancer.
6.8 Additional candidate markers identified by DMH

Requirement for coverage of ~100bp fragments using multiple pyrosequencing assays was a technical limitation that has impeded validation of other DMH fragments, however, given technical validation of PARP1 within the KConFab study; attempt to validate other ‘top hit’ fragments displaying cancer-associated hypermethylation may be worthwhile. It may be that rather than representing a locus-specific event, PARP1 hypermethylation may be indicative of the beginnings of a general loss of protection from DNA at gene promoters referred to as the ‘CpG island methylator phenotype’ that occurs frequently in tumors (Issa, 2004).

Despite the semi-quantitative nature of methylation analysis of DMH (Dai et al., 2008), this method had shown impressive sensitivity to detect subtle PARP1 hypermethylation within a relatively small set of blood samples, highlighting the potential utility of methylation-sensitive-restriction enzyme-based assays as complementary methods of methylation analysis in addition to more quantitative bisulphite-sequencing-based approaches.

6.9 Haplotype-associated DNA methylation.

We observed a trimodal distribution of DNA methylation at an LGR6 promoter region, potentially indicative of haplotype-associated methylation (chapter 4). This is interesting, as a recent GWAS study has reported that the LGR6 minor allele is associated with increased risk of ER-breast cancer (Garcia-Closas et al., 2013a), suggesting that the affect of the reported risk-associated SNP (rs6678914) may be mediated through haplotype-specific hypermethylation. The LGR6 DMR spans the transcription start site of the LGR6 NM_021636 transcript, a POLR2A binding site, and conserved binding sites for GATA and EVL transcription factors, suggesting a potential effect of methylation on LGR6 transcription.
The trimodal distribution of a CpG site within the LGR6 DMR (Fig 6.1 labelled LGR6-DMR-CpG3) was confirmed by pyrosequencing in EPIC3 (n=559, Fig 6.1). Surprisingly, the trimodal methylation pattern of LGR6 cg05044291 was not evident in normal breast epithelium DNA samples (n=98, Fig 6.1), or breast tumour (n=382, data not shown) within publicly available Illumina Infinium HumanMethylation450 BeadChip array data from The Cancer Genome Atlas (TCGA) study. Similarly, haplotype-associated DNA methylation at the promoter of the FGFR2 gene has recently been reported (Heyn et al., 2013), and GWAS studies have reproducibly shown that the FGFR2 minor allele, displaying increased FGFR2 expression, is associated with increased breast cancer risk (Easton et al., 2007; Garcia-Closas et al., 2013a). Consistently, trimodal methylation of these FGFR2 CpG sites was apparent in our data (data not shown). The possibility that haplotype-associated promoter hypermethylation may represent a general silencing mechanism associated with GWAS SNPs has been raised before (Bell et al., 2012; Leung et al., 2012; Ogino et al., 2013), and DNA methylation pattern associated with a SNP in the obesity-associated FTO gene, potentially regulating activity of a long-range enhancer, has been identified (Bell et al., 2010). The importance of interactions between DNA methylation and genotype are currently being uncovered, as genetic motifs strongly influencing DNA methylation patterns in cis have been identified (Lienert et al., 2011). Tissue-specificity of the haplotype-associated methylation pattern of LGR6 is consistent with a recent study that reported tissue-specificity of ~40% of 479 locus-specific haplotype-associated DNA methylation patterns (Heyn et al., 2013), and the observation that haplotype-associated FGFR2 expression is tissue-specific (Huijts et al., 2011). This combined with the finding that haplotype-specific methylation patterns are enriched for H3k27ac and CTCF binding indicates that epigenetic factors must interact with haplotypic genetic differences to control promoter DNA methylation state. Therefore, investigation of tissue-specific haplotype-associated promoter hypermethylation may help
elucidate the mechanisms by which aberrant promoter hypermethylation occurs. Lastly, epigenetic mediation of the effects of disease-associated genetic variants may offer potential for therapeutic intervention, due to the reversible nature of epigenetic marks.
Figure 6.1. Trimodal DNA methylation of CpG sites within LGR6 (A) Methylation distributions (strip-charts) of all LGR6 loci represented on the Illumina Infinium HumanMethylation450 BeadChip array arranged by genomic location in 228 blood DNA samples from the EPIC study (EPIC5/EPIC6). A trimodal distribution of DNA methylation at four adjacent CpG sites within LGR6 promoter 2 (P2) is evident, compared with constitutively unmethylated loci within promoters 1 (P1) and 3 (P3), and constitutively methylated loci in P4. (B) The LGR6 locus selected for pyrosequencing validation is labelled CpG3. The only locus outside the DMR displaying trimodal methylation distribution is a polymorphic CpG site (PM CpG). rs6678914 marks the approximate location of a SNP associated with ER- breast cancer risk in a recent GWAS study (Garcia-Closas et al, 2013). Map information indicating distance (bp) between CpG sites is shown on the X-axis. (B) Individual strip-chart for LGR6 CpG3 within the same 450k array DNA methylation data as A (C) Validation of trimodal methylation distribution by pyrosequencing in an independent sample set of 559 blood DNA samples (EPIC3 subgroup) (D) Publicly available 450k array data for LGR6 CpG3 methylation in normal breast epithelial samples (n=98), from The Cancer Genome Atlas (TCGA) study.

A.

LGR6 methylation distribution by genomic location

B.

LGR6cg05044291 methylation (EPIC5)

C.

LGR6cg05044291 methylation (EPIC5)

D.

LGR6cg05044291 methylation (TCGA-NRE2)
6.10 EWAS studies

6.10.1 EWAS in general

EWAS studies represent a potential way of identifying DNA methylation patterns associated with both breast cancer risk and breast cancer related exposures in a relatively unbiased, high-throughput way, and may therefore be useful for identification of epigenetic risk biomarkers. Methylation-dependent single-base sequence alteration by bisulphite conversion means that in practical terms, genome-wide methylation analysis is similar to measurement of SNPs genome-wide.

It is becoming increasingly clear, however, that instability and tissue-specificity of DNA methylation at many loci make EWAS studies far more complicated than GWAS studies, and that great consideration for study design and exclusion of potential biases must be taken in order to ensure that EWAS studies can be appropriately controlled.

6.10.2 DNA methylation stability

Temporal stability of molecular markers is a prerequisite for their development as disease risk prediction biomarkers, as population screening would, in practicality, only allow testing every few years (Hanash et al., 2011). Furthermore, investigation of DNA methylation biomarkers within prospective studies using stored tissue samples requires methylation stability. Thirdly, contribution to long-term cancer risk, either as an innate factor or through marking epigenetic memory of early-life exposures, would also require long-term stability of DNA methylation (Jefferson et al., 2013; Talens et al., 2010). Methylation stability at many
loci has been shown in longitudinal studies (Talens et al., 2010; Woodfine et al., 2011), and
association of methylation marks with past exposures to famine (Heijmans et al., 2008;
Shenker et al., 2013) and smoking demonstrate epigenetic memory. However, the finding
of alteration of ATMmvp2a methylation with fasting and other environmental exposures, and
\textit{PARP1} hypermethylation associated with TTD, as well as reported alteration of DNA
methylation by immune activation (Lam et al., 2012; Zhang et al., 2013) metabolic profiles
(Petersen et al., 2013), diet (Lampe et al., 2013), and age (Lam et al., 2012; Rakyan et al.,
2010), indicate that DNA methylation at many loci may undergo dynamic alteration, and that
single time-point measures may reflect current environmental conditions rather than long
term cancer risk. Despite this, DNA methylation likely displays strong stability at many loci,
relative to other molecular measures such as RNA expression, protein, and metabolic
profiles, and other epigenetic marks such as histone modifications and miRNA expression
(Hanash et al., 2011).

6.10.3 450k array

The 450k array represents the most suitable platform for current EWAS investigations, as it
provides highly quantitative, single-base resolution methylation data, is a high-throughput
system and relatively low cost, and is less biased than many other methods in terms of
genome coverage (Langevin and Kelsey, 2013; Rakyan et al., 2011).

However, technical biases caused by probe cross-hybridisation, different probe types and
batch effects can be problematic, and methylation biases caused by polymorphic CpG sites
or haplotype-specific methylation coupled with differential allele frequencies between
experimental groups can confound analyses.
Despite widely reported technical validation of associations identified using the Illumina 450k array (Dedeurwaerder et al., 2011), we were unsuccessful in identifying differentially methylated loci that were reproducible in pyrosequencing-based assays, with the exception of haplotype-specific DNA methylation at LGR6. Though we did not carry out technical validation of the Illumina 450k array, it appears likely that technical artefacts within array data, rather than biological differences between test and validation sample sets account for failure to validate associations, as the distributions of methylation at many loci was inconsistent between array and pyrosequencing data.

Perhaps the greatest challenge to identification of differential methylation in blood using this platform is the low variability in blood DNA methylation relative to tumour DNA methylation. For both EWAS analyses of fasting status and case-control status, almost all significantly differentially methylated loci displayed Δβ-values less than 10%, whereas a general consensus among researchers is that significant probes displaying Δβ-values less than 10-20% should be excluded from analysis, due to high variability between technical replicates using the 450k array (Dedeurwaerder et al., 2013; Wilhelm-Benartzi et al., 2013). While pyrosequencing-based validation of smoking-associated loci with small Δβ values (~2%) using the same data-set and analytical methods as used for our case-control analysis has been published (Shenker et al., 2012), these associations were highly consistent. Such consistency of associations with breast cancer risk may be unlikely, due to the highly heterogeneous and multi-factorial nature of breast cancer risk compared with smoking. In the absence of large Δβ values or highly consistent associations, very large sample-sets will be required to identify significant associations in future studies. Power analysis to determine sample sizes required, as well as stratified and multivariate analyses to detect associations for specific disease subgroups may be useful in future studies, as will consortium-based
meta-analyses of multiple studies, which have been useful for identification of GWAS associations in breast cancer (Fanale et al., 2012).

Lastly, it is possible that inter-individual DNA methylation variability may be greater within individual blood cell types than within whole blood, and that such variability may be masked in whole blood by methylation variability between cell types. Therefore, EWAS studies using DNA samples derived from specific blood cell types may have more power to detect reproducible associations compared with studies using whole blood.

6.10.4 Potential rarity of epigenetic events associated with cancer risk

The statistical approach we have used for EWAS studies selects for large, consistent differences in DNA methylation between experimental groups, as such loci are easier to detect and validate using available methods, and may offer better discriminatory value as biomarkers. However, it is important to consider that this approach may filter out regions of low DNA methylation variability that may be of higher biological significance, and regions that are strongly associated with the exposure of interest, but only within a subset of cases. Prioritisation of loci with high Δβ-values tends to select for genomic regions of high DNA methylation variability, consistent with the finding that differential methylation associated with cancer and age is enriched at gene bodies (Teschendorff et al., 2009). However, it is possible that selection of high methylation variability enriches for regions at which DNA methylation is under low selective pressure, and is dispensable for normal cellular function. Furthermore, regions of high DNA methylation variability often display intermediate DNA methylation, suggesting that neither the unmethylated or methylated state is ‘abnormal’. In contrast, regions with low DNA methylation variability, such as CpG islands and
transposable elements are under tight epigenetic control in normal tissue, and display strong conservation between individuals and tissues, suggesting an important functional role of methylation at these regions (Jones, 2012). Therefore, the rarity of aberrant epigenetic event on a cellular level may be directly related to the functional importance of methylation at those regions, which would prevent us from detecting them in EWAS studies. The promoter regions of BRCA1 (Hansmann et al., 2012) and PARP1 represent examples of loci identified through candidate gene approaches, at which aberrant DNA methylation appears to occur, yet hypermethylation is rare on a cellular level, and is therefore just within the detection sensitivity of pyrosequencing. While the functional consequences of hypermethylation of these regions for individual blood cells is not known, given that BRCA1 promoter hypermethylation is associated with transcriptional repression and impaired DDR in breast tumors (Catteau et al., 1999), it is highly plausible that BRCA1 hypermethylation in normal cells may increase propensity for tumourigenicity. Therefore, in order to identify epigenetic events contributing to cancer susceptibility, it may be more worthwhile investigating subtle methylation changes at regions that are normally constitutively unmethylated or methylated rather than regions displaying intermediate or highly variable methylation. Assays with improved sensitivity to detect subtle aberrant DNA methylation, such as single-cell epigenetic technologies or methods that enrich for cells based on methylation state may enable investigation of the cellular consequences of rare epigenetic states in future studies.

Another bias is that the current methodology selects for common DNA methylation changes potentially associated with disease, as rare events of high penetrance, would not significantly change median DNA methylation across populations. While we focused on median shifts in DNA methylation across populations, an alternative approach would be to assess the relative frequency of statistical outliers i.e. individuals with DNA methylation
outside the normal methylation range (Hansmann et al., 2012; Herman, 1998). The latter approach is limited in that vast numbers of samples are required to confirm association of rare events with disease risk, and that it ignores common epigenetic variability, therefore a combination of both approaches may be required.

6.11 Conclusions

Despite the challenges facing EWAS studies, the availability of large prospective study sample cohorts provides a unique opportunity to carry out appropriately powered molecular epidemiological studies for breast cancer. Furthermore, a range of different study cohort types, such as case-control studies, longitudinal studies and twin studies are available, allowing interrogation of different biological questions. The complexities of tissue-specific methylation, dynamic alteration of DNA methylation, and interactions of the epigenome with the genome, metabolome, environment and immune system highlight the requirement for a holistic, ‘systems biology’ approaches to investigating molecular mechanisms of cancer risk, with integration of different data-types, and sophisticated bioinformatic and statistical approaches that will allow modelling of complex interactions.

6.12 Summary

Our research highlights the potential value of blood DNA methylation as a rich source of information for investigation of breast cancer risk factors and development of biomarker for both risk prediction and early detection using blood samples from population cohorts.
This research provides novel insights into interactions between the epigenome and factors affecting breast cancer risk, including both heritable and environmental factors.

While there are many challenges facing EWAS studies for breast cancer risk, and direct associations with breast cancer incidence of DNA methylation patterns remain elusive, investigation of interactions between breast cancer risk factors and DNA methylation may provide insight into the mechanisms by which risk factors increase risk at the molecular level. Furthermore, investigation of inter-individual variability in epigenetic response to risk factors and risk altering interventions will likely provide insight into the factors affecting breast cancer risk. Additionally, we provide evidence that *PARP1* hypermethylation in blood may represent an early marker of preclinical disease, supporting the idea that dysregulation of DNA methylation represents a very early event in disease initiation. Lastly, the potential role of DNA methylation in mediating effects of breast cancer associated genetic variability is supported by our findings. Further research into each of these findings will likely provide valuable and clinically relevant information regarding the contribution of DNA methylation to breast cancer susceptibility.
References


(2013). Sex hormones and risk of breast cancer in premenopausal women: a collaborative reanalysis of individual participant data from seven prospective studies. The lancet oncology.


The histologic phenotypes of breast carcinoma occurring before age 40 years in women with and without BRCA1 or BRCA2 germline mutations: a population-based study. Lancet 349, 1505 - 1510.


Friso, S., Choi, S. W., Girelli, D., Mason, J. B., Dolnikowski, G. G., Bagley, P. J., Olivieri, O., Jacques, P. F., Rosenberg, I. H., Corrocher, R., and Selhub, J. (2002). A common mutation in the 5,10-


of ABCB1, FOXC1, PPP2R2B and PTEN in ductal carcinoma in situ and early invasive breast cancer. Breast cancer research: BCR 12, R3.


251


Singer, H., Walier, M., Nusgen, N., Meesters, C., Schreiner, F., Woelfle, J., Fimmers, R., Wienker, T., Kalscheuer, V. M., Becker, T., et al. (2012). Methylation of L1Hs promoters is lower on the inactive X, has a tendency of being higher on autosomes in smaller genomes and shows inter-individual variability at some loci. Human molecular genetics 21, 219-235.


methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome research 20, 440-446.


Appendix
Appendix table 1. Studies included in meta-analyses of genome-wide DNA methylation and cancer risk

<table>
<thead>
<tr>
<th>Ref (figures)</th>
<th>Pubmed ref</th>
<th>Author</th>
<th>Publication year</th>
<th>Cancer type</th>
<th>measure</th>
<th>assay</th>
<th>Sex</th>
<th>Place of study</th>
<th>N (case/ctrl)</th>
<th>Prospective or retrospective</th>
<th>Newly diagnosed</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Meta-analysis 01/01/2000-08/06/2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>20179218</td>
<td>Wilhelm</td>
<td>01/03/2010</td>
<td>bladder</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M&amp;F*</td>
<td>U.S.</td>
<td>285/465</td>
<td>retrospective</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>20478068</td>
<td>Mirabello</td>
<td>17/03/2010</td>
<td>testicular</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M</td>
<td>U.S.</td>
<td>114/151</td>
<td>retrospective</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>20099281</td>
<td>Hou</td>
<td>15/10/2010</td>
<td>gastric</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M&amp;F</td>
<td>Poland</td>
<td>300/419</td>
<td>retrospective</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>21188491</td>
<td>Zhu</td>
<td>22/03/2011</td>
<td>all (combined)</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M</td>
<td>Boston</td>
<td>304/87</td>
<td>retrospective</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>21188491</td>
<td>Zhu</td>
<td>22/03/2011</td>
<td>all (combined)</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M</td>
<td>Boston</td>
<td>204/478</td>
<td>retrospective</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>22126203</td>
<td>Rusiecki</td>
<td>03/04/2011</td>
<td>gastric</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M&amp;F</td>
<td>Oman</td>
<td>110/104</td>
<td>retrospective</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>22076155</td>
<td>Liao</td>
<td>11/06/2011</td>
<td>renal cell</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M&amp;F</td>
<td>Europe</td>
<td>328/654</td>
<td>retrospective</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>21960343</td>
<td>Di</td>
<td>12/10/2011</td>
<td>hepatocellular</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M&amp;F</td>
<td>China</td>
<td>315/356</td>
<td>retrospective</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>22173668</td>
<td>Gao</td>
<td>31/01/2012</td>
<td>gastric</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>F</td>
<td>Shanghai</td>
<td>192/382</td>
<td>prospective</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>22371529</td>
<td>Xu</td>
<td>27/02/2012</td>
<td>breast</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>F</td>
<td>US</td>
<td>1064/1100</td>
<td>retrospective</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>21445976</td>
<td>Cash</td>
<td>03/01/2012</td>
<td>bladder</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M&amp;F</td>
<td>China</td>
<td>510/528</td>
<td>retrospective</td>
<td>no</td>
</tr>
<tr>
<td>12</td>
<td>22374981</td>
<td>Brennan</td>
<td>01/05/2012</td>
<td>breast</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>F</td>
<td>UK</td>
<td>252/255</td>
<td>prospective</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>22374961</td>
<td>Brennan</td>
<td>01/05/2012</td>
<td>breast</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>F</td>
<td>Europe</td>
<td>263/257</td>
<td>prospective</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>22581841</td>
<td>Wu</td>
<td>07/05/2012</td>
<td>hepatocellular</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M&amp;F</td>
<td>Taiwan</td>
<td>305/1254</td>
<td>prospective</td>
<td>yes</td>
</tr>
<tr>
<td>15</td>
<td>22678115</td>
<td>Wu</td>
<td>07/06/2012</td>
<td>breast</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>F</td>
<td>US/Australia</td>
<td>279/340</td>
<td>retrospective</td>
<td>no</td>
</tr>
<tr>
<td>16</td>
<td>17203338</td>
<td>Hsiung</td>
<td>16/01/2007</td>
<td>head and neck</td>
<td>LRE1</td>
<td>COBRA**</td>
<td>M&amp;F</td>
<td>Boston</td>
<td>278/526</td>
<td>retrospective</td>
<td>yes</td>
</tr>
<tr>
<td>17</td>
<td>20099281</td>
<td>Hou</td>
<td>15/10/2010</td>
<td>gastric</td>
<td>Alu</td>
<td>pyrosequencing</td>
<td>M&amp;F</td>
<td>Poland</td>
<td>302/421</td>
<td>retrospective</td>
<td>yes</td>
</tr>
<tr>
<td>18</td>
<td>21188491</td>
<td>Zhu</td>
<td>22/03/2011</td>
<td>all (combined)</td>
<td>Alu</td>
<td>pyrosequencing</td>
<td>M</td>
<td>Boston</td>
<td>205/486</td>
<td>retrospective</td>
<td>no</td>
</tr>
<tr>
<td>19</td>
<td>22173668</td>
<td>Gao</td>
<td>31/01/2012</td>
<td>gastric</td>
<td>Alu</td>
<td>pyrosequencing</td>
<td>F</td>
<td>Shanghai</td>
<td>192/381</td>
<td>prospective</td>
<td>yes</td>
</tr>
<tr>
<td>Study ID</td>
<td>Authors</td>
<td>Date</td>
<td>Tissue</td>
<td>Genotype</td>
<td>Treatment</td>
<td>Gender</td>
<td>Country</td>
<td>Cases</td>
<td>Controls</td>
<td>Study Design</td>
<td>Prospective</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>------</td>
<td>--------</td>
<td>----------</td>
<td>------------</td>
<td>--------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>20</td>
<td>Wu</td>
<td>07/06/2012</td>
<td>Breast</td>
<td>Alu</td>
<td>MethyLight</td>
<td>F</td>
<td>US/Australia</td>
<td>266/334</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Wu</td>
<td>12/05/2012</td>
<td>Hepatocellular</td>
<td>Sat2</td>
<td>MethyLight</td>
<td>M&amp;F</td>
<td>Taiwan</td>
<td>305/1254</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Wu</td>
<td>07/06/2012</td>
<td>Breast</td>
<td>Sat2</td>
<td>MethyLight</td>
<td>F</td>
<td>US/Australia</td>
<td>266/333</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pufulete</td>
<td>02/05/2003</td>
<td>Colorectal</td>
<td>5meC</td>
<td>Methyl acceptance</td>
<td>M&amp;F</td>
<td>UK</td>
<td>28/76</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Pufulete</td>
<td>02/05/2003</td>
<td>Colorectal adenoma</td>
<td>5meC</td>
<td>Methyl acceptance</td>
<td>M&amp;F</td>
<td>UK</td>
<td>35/76</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Lim</td>
<td>01/01/2008</td>
<td>Colorectal</td>
<td>5meC</td>
<td>HPLC and MS***</td>
<td>F</td>
<td>U.S.</td>
<td>115/115</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Moore</td>
<td>09/04/2008</td>
<td>Bladder</td>
<td>5meC</td>
<td>HPCE and hpaII****</td>
<td>M&amp;F</td>
<td>Spain</td>
<td>775/397</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Choi</td>
<td>07/07/2009</td>
<td>Breast</td>
<td>5meC</td>
<td>Mass spectrometry</td>
<td>F</td>
<td>U.S.</td>
<td>176/173</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Xu</td>
<td>27/02/2012</td>
<td>Breast</td>
<td>CCGG******</td>
<td>LUMA****</td>
<td>F</td>
<td>US</td>
<td>1055/1101</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Studies excluded from meta-analysis (01/01/2000-08/06/2012)**

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Authors</th>
<th>Date</th>
<th>Tissue</th>
<th>Genotype</th>
<th>Treatment</th>
<th>Gender</th>
<th>Country</th>
<th>Cases</th>
<th>Controls</th>
<th>Study Design</th>
<th>Prospective</th>
<th>Retrospective</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>Choi</td>
<td>07/07/2009</td>
<td>Breast</td>
<td>LINE1</td>
<td>Pyrosequencing</td>
<td>F</td>
<td>U.S.</td>
<td>19/18</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Cho</td>
<td>30/12/2010</td>
<td>Breast</td>
<td>LINE1</td>
<td>MethyLight</td>
<td>F</td>
<td>Turkey</td>
<td>40/40</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Cho</td>
<td>30/12/2010</td>
<td>Breast</td>
<td>Sat2</td>
<td>MethyLight</td>
<td>F</td>
<td>Turkey</td>
<td>40/40</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Cho</td>
<td>30/12/2010</td>
<td>Breast</td>
<td>Alu</td>
<td>MethyLight</td>
<td>F</td>
<td>Turkey</td>
<td>40/40</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Pyathilake</td>
<td>27/05/2011</td>
<td>Breast</td>
<td>CIN*****</td>
<td>LINE1</td>
<td>Pyrosequencing</td>
<td>F</td>
<td>U.S.</td>
<td>103/273</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Wiendschendter</td>
<td>16/07/2008</td>
<td>Breast</td>
<td>Alu</td>
<td>MethyLight</td>
<td>F</td>
<td>Germany</td>
<td>169/180</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Patchsung</td>
<td>15/05/2012</td>
<td>Bladder</td>
<td>LINE1</td>
<td>COBRA</td>
<td>M&amp;F</td>
<td>Thailand</td>
<td>61/45</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>pobook</td>
<td>12/1/2010</td>
<td>Oral</td>
<td>LINE1</td>
<td>COBRA</td>
<td>M&amp;F</td>
<td>Thailand</td>
<td>34/43</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**(B) Meta-analysis 08/06/2012-05/09/2013**

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Authors</th>
<th>Date</th>
<th>Tissue</th>
<th>Genotype</th>
<th>Treatment</th>
<th>Gender</th>
<th>Country</th>
<th>Cases</th>
<th>Controls</th>
<th>Study Design</th>
<th>Prospective</th>
<th>Retrospective</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>Delgado-Cruzata</td>
<td>07/08/2012</td>
<td>Breast</td>
<td>5meC</td>
<td>Methyl-acceptance</td>
<td>F</td>
<td>US</td>
<td>273/335</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>Huang</td>
<td>21/11/2012</td>
<td>Colorectal</td>
<td>5meC</td>
<td>HPLC and TMS</td>
<td>M&amp;F</td>
<td>US</td>
<td>370/493</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Walters</td>
<td>17/06/2013</td>
<td>Colorectal</td>
<td>ALU</td>
<td>MethyLight</td>
<td>M&amp;F</td>
<td>Australia</td>
<td>539/242</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>40</td>
<td>22705975</td>
<td>Delgado-Cruzata</td>
<td>07/08/2012</td>
<td>Breast</td>
<td>CCGG</td>
<td>LUMA</td>
<td>F</td>
<td>US</td>
<td>273/335</td>
<td>retrospective</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>23804018</td>
<td>Walters</td>
<td>17/06/2013</td>
<td>colorectal</td>
<td>LINE1</td>
<td>MethyLight</td>
<td>M&amp;F</td>
<td>Australia</td>
<td>539/242</td>
<td>retrospective</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>23222549</td>
<td>Hyland</td>
<td>23/02/2013</td>
<td>Melanoma</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M&amp;F</td>
<td>US</td>
<td>115/123</td>
<td>retrospective</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>23804018</td>
<td>Walters</td>
<td>17/06/2013</td>
<td>colorectal</td>
<td>Sat2</td>
<td>MethyLight</td>
<td>M&amp;F</td>
<td>Australia</td>
<td>539/242</td>
<td>retrospective</td>
<td>no</td>
<td></td>
</tr>
</tbody>
</table>

**Studies excluded from meta-analysis (08/06/2012-05/09/2013)**

<p>| | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>23300023</td>
<td>Friso</td>
<td>22/03/2013</td>
<td>Mixed</td>
<td>5meC</td>
<td>HPLC</td>
<td>M&amp;F</td>
<td>Italy</td>
<td>68/68</td>
<td>prospective</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>23300023</td>
<td>Friso</td>
<td>22/03/2013</td>
<td>Mixed</td>
<td>5meC</td>
<td>HPLC</td>
<td>M&amp;F</td>
<td>Italy</td>
<td>68/68</td>
<td>retrospective</td>
<td>no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>23943784</td>
<td>Tahara</td>
<td>13/08/2013</td>
<td>Gastric</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M&amp;F</td>
<td>Japan</td>
<td>72/67</td>
<td>retrospective</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>22997064</td>
<td>Li</td>
<td>01/01/2013</td>
<td>Mixed</td>
<td>LINE1</td>
<td>pyrosequencing</td>
<td>M</td>
<td>Norway</td>
<td>49/98</td>
<td>prospective</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations**

*M&F: Samples from both sexes included in study*

**COBRA: Combined bisulphite restriction analysis (fluorescent based)**

***HPLC and MS: High Performance Liquid Chromatography and Mass Spectrometry***

****HPCE and hpaII: High Performance Capillary Electrophoresis and HpaII restriction digestion***

*****LUMA: LUminometric methylation assay***

******CCGG: HpaII restriction site***

*******CIN: Cervical Intraepithelial Neoplasia***
Appendix table 2. Assessment of association with case-control status of 450k array loci (EPIC4) within the regions identified as significantly differentially methylated in the differential methylation hybridisation array discovery study.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Genomic location</th>
<th>control median</th>
<th>case median</th>
<th>$\Delta \beta$ (case-control)*</th>
<th>cond_glm $p$**</th>
<th>cond_glm $p(fdr)$*</th>
<th>gene region</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOXO3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.09E+08</td>
<td>0.699</td>
<td>0.671</td>
<td>-0.028</td>
<td>0.000757</td>
<td>0.021879</td>
<td>N_Shore</td>
</tr>
<tr>
<td>6</td>
<td>1.09E+08</td>
<td>0.057</td>
<td>0.052</td>
<td>-0.005</td>
<td>0.009338</td>
<td>0.060223</td>
<td>Island</td>
</tr>
<tr>
<td>6</td>
<td>1.09E+08</td>
<td>0.024</td>
<td>0.028</td>
<td>0.004</td>
<td>0.269833</td>
<td>0.436705</td>
<td>Island</td>
</tr>
<tr>
<td>PARP1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.27E+08</td>
<td>0.013</td>
<td>0.017</td>
<td>0.004</td>
<td>0.252119</td>
<td>0.41783</td>
<td>Island</td>
</tr>
<tr>
<td>1</td>
<td>2.27E+08</td>
<td>0.052</td>
<td>0.047</td>
<td>-0.005</td>
<td>0.113018</td>
<td>0.248244</td>
<td>Island</td>
</tr>
<tr>
<td>1</td>
<td>2.27E+08</td>
<td>0.005</td>
<td>0.005</td>
<td>0</td>
<td>0.770416</td>
<td>0.853954</td>
<td>Island</td>
</tr>
<tr>
<td>1</td>
<td>2.27E+08</td>
<td>0.018</td>
<td>0.018</td>
<td>0</td>
<td>0.844013</td>
<td>0.90362</td>
<td>Island</td>
</tr>
<tr>
<td>1</td>
<td>2.27E+08</td>
<td>0.053</td>
<td>0.05</td>
<td>-0.003</td>
<td>0.081674</td>
<td>0.202221</td>
<td>Island</td>
</tr>
<tr>
<td>1</td>
<td>2.27E+08</td>
<td>0.065</td>
<td>0.06</td>
<td>-0.005</td>
<td>0.009269</td>
<td>0.060007</td>
<td>S_Shore</td>
</tr>
<tr>
<td>1</td>
<td>2.27E+08</td>
<td>0.166</td>
<td>0.182</td>
<td>0.016</td>
<td>0.013828</td>
<td>0.073302</td>
<td>S_Shore</td>
</tr>
<tr>
<td>TLE4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>82187670</td>
<td>0.015</td>
<td>0.015</td>
<td>0</td>
<td>0.408164</td>
<td>0.570606</td>
<td>Island</td>
</tr>
<tr>
<td>9</td>
<td>82187855</td>
<td>0.008</td>
<td>0.01</td>
<td>0.002</td>
<td>0.093041</td>
<td>0.219343</td>
<td>Island</td>
</tr>
<tr>
<td>9</td>
<td>82187938</td>
<td>0.082</td>
<td>0.074</td>
<td>-0.008</td>
<td>0.007547</td>
<td>0.054282</td>
<td>Island</td>
</tr>
<tr>
<td>9</td>
<td>82187997</td>
<td>0.035</td>
<td>0.03</td>
<td>-0.005</td>
<td>0.014938</td>
<td>0.076342</td>
<td>Island</td>
</tr>
<tr>
<td>9</td>
<td>82188393</td>
<td>0.011</td>
<td>0.008</td>
<td>-0.003</td>
<td>0.523045</td>
<td>0.667799</td>
<td>Island</td>
</tr>
<tr>
<td>TGFB2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30647540</td>
<td>0.329</td>
<td>0.301</td>
<td>-0.028</td>
<td>0.001444</td>
<td>0.027167</td>
<td>N_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30647670</td>
<td>0.244</td>
<td>0.235</td>
<td>-0.009</td>
<td>0.213956</td>
<td>0.375037</td>
<td>N_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30647672</td>
<td>0.253</td>
<td>0.245</td>
<td>-0.008</td>
<td>0.214524</td>
<td>0.375705</td>
<td>N_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30647725</td>
<td>0.071</td>
<td>0.065</td>
<td>-0.006</td>
<td>0.895693</td>
<td>0.937007</td>
<td>N_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30647773</td>
<td>0.126</td>
<td>0.129</td>
<td>0.003</td>
<td>0.999556</td>
<td>0.999751</td>
<td>N_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30647802</td>
<td>0.048</td>
<td>0.045</td>
<td>-0.003</td>
<td>0.034671</td>
<td>0.121154</td>
<td>N_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30647808</td>
<td>0.035</td>
<td>0.04</td>
<td>0.005</td>
<td>0.005964</td>
<td>0.04852</td>
<td>N_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30647919</td>
<td>0.035</td>
<td>0.039</td>
<td>0.004</td>
<td>0.22666</td>
<td>0.389613</td>
<td>N_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30647968</td>
<td>0.05</td>
<td>0.055</td>
<td>0.005</td>
<td>0.545773</td>
<td>0.686039</td>
<td>N_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30647988</td>
<td>0.033</td>
<td>0.043</td>
<td>0.01</td>
<td>0.03862</td>
<td>0.128916</td>
<td>Island</td>
</tr>
<tr>
<td>3</td>
<td>30648096</td>
<td>0.051</td>
<td>0.048</td>
<td>-0.003</td>
<td>0.312261</td>
<td>0.480468</td>
<td>Island</td>
</tr>
<tr>
<td>3</td>
<td>30648098</td>
<td>0.052</td>
<td>0.049</td>
<td>-0.003</td>
<td>0.153762</td>
<td>0.30235</td>
<td>Island</td>
</tr>
<tr>
<td>3</td>
<td>30648325</td>
<td>0.029</td>
<td>0.028</td>
<td>-0.001</td>
<td>0.322848</td>
<td>0.49074</td>
<td>Island</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30648407</td>
<td>0.048</td>
<td>0.044</td>
<td>-0.004</td>
<td>0.01037</td>
<td>0.063464</td>
<td>Island</td>
</tr>
<tr>
<td>3</td>
<td>30648598</td>
<td>0.033</td>
<td>0.031</td>
<td>-0.002</td>
<td>0.521665</td>
<td>0.666665</td>
<td>Island</td>
</tr>
<tr>
<td>3</td>
<td>30648819</td>
<td>0.049</td>
<td>0.048</td>
<td>-0.001</td>
<td>0.125032</td>
<td>0.264708</td>
<td>Island</td>
</tr>
<tr>
<td>3</td>
<td>30648949</td>
<td>0.021</td>
<td>0.031</td>
<td>0.01</td>
<td>0.095446</td>
<td>0.222978</td>
<td>S_Shore</td>
</tr>
<tr>
<td>3</td>
<td>30649017</td>
<td>0.023</td>
<td>0.02</td>
<td>-0.003</td>
<td>0.881921</td>
<td>0.928128</td>
<td>S_Shore</td>
</tr>
</tbody>
</table>

**LBH**

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>30453758</td>
<td>0.055</td>
<td>0.053</td>
<td>-0.002</td>
<td>0.169786</td>
<td>0.322414</td>
</tr>
<tr>
<td>2</td>
<td>30454146</td>
<td>0.02</td>
<td>0.03</td>
<td>0.01</td>
<td>0.004687</td>
<td>0.043545</td>
</tr>
<tr>
<td>2</td>
<td>30454275</td>
<td>0.016</td>
<td>0.016</td>
<td>0</td>
<td>0.121003</td>
<td>0.25924</td>
</tr>
<tr>
<td>2</td>
<td>30454279</td>
<td>0.055</td>
<td>0.05</td>
<td>-0.005</td>
<td>0.007253</td>
<td>0.053263</td>
</tr>
<tr>
<td>2</td>
<td>30454334</td>
<td>0.069</td>
<td>0.076</td>
<td>0.007</td>
<td>0.134526</td>
<td>0.277395</td>
</tr>
<tr>
<td>2</td>
<td>30454344</td>
<td>0.027</td>
<td>0.031</td>
<td>0.004</td>
<td>0.003297</td>
<td>0.037427</td>
</tr>
<tr>
<td>2</td>
<td>30454363</td>
<td>0.037</td>
<td>0.036</td>
<td>-0.001</td>
<td>0.390592</td>
<td>0.554877</td>
</tr>
</tbody>
</table>

**PRR5**

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>45098335</td>
<td>0.012</td>
<td>0.015</td>
<td>0.003</td>
<td>0.016752</td>
<td>0.081177</td>
</tr>
<tr>
<td>22</td>
<td>45098346</td>
<td>0.072</td>
<td>0.069</td>
<td>-0.003</td>
<td>0.090926</td>
<td>0.216224</td>
</tr>
<tr>
<td>22</td>
<td>45098367</td>
<td>0.057</td>
<td>0.054</td>
<td>-0.003</td>
<td>0.797592</td>
<td>0.872574</td>
</tr>
<tr>
<td>22</td>
<td>45098800</td>
<td>0.039</td>
<td>0.036</td>
<td>-0.003</td>
<td>0.009792</td>
<td>0.06166</td>
</tr>
<tr>
<td>22</td>
<td>45098907</td>
<td>0.007</td>
<td>0.009</td>
<td>0.002</td>
<td>0.068959</td>
<td>0.1821</td>
</tr>
</tbody>
</table>

**FZD3**

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>28351217</td>
<td>0.058</td>
<td>0.056</td>
<td>-0.002</td>
<td>0.443144</td>
<td>0.601361</td>
</tr>
<tr>
<td>8</td>
<td>28351457</td>
<td>0.056</td>
<td>0.053</td>
<td>-0.003</td>
<td>0.042412</td>
<td>0.136066</td>
</tr>
<tr>
<td>8</td>
<td>28351501</td>
<td>0.046</td>
<td>0.046</td>
<td>0</td>
<td>0.029758</td>
<td>0.111002</td>
</tr>
<tr>
<td>8</td>
<td>28351567</td>
<td>0.033</td>
<td>0.039</td>
<td>0.006</td>
<td>0.006221</td>
<td>0.049528</td>
</tr>
<tr>
<td>8</td>
<td>28351601</td>
<td>0.064</td>
<td>0.063</td>
<td>-0.001</td>
<td>0.024936</td>
<td>0.100587</td>
</tr>
<tr>
<td>8</td>
<td>28352084</td>
<td>0.035</td>
<td>0.035</td>
<td>0</td>
<td>0.027404</td>
<td>0.105964</td>
</tr>
<tr>
<td>8</td>
<td>28352396</td>
<td>0.064</td>
<td>0.059</td>
<td>-0.005</td>
<td>0.796057</td>
<td>0.871462</td>
</tr>
</tbody>
</table>

**SMAD1**

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.02</td>
<td>0.021</td>
<td>0.001</td>
<td>0.017973</td>
<td>0.084275</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.018</td>
<td>0.019</td>
<td>0.001</td>
<td>0.462039</td>
<td>0.617553</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.026</td>
<td>0.032</td>
<td>0.006</td>
<td>0.002112</td>
<td>0.031294</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.046</td>
<td>0.042</td>
<td>-0.004</td>
<td>0.375589</td>
<td>0.540959</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.071</td>
<td>0.064</td>
<td>-0.007</td>
<td>0.115489</td>
<td>0.251706</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.045</td>
<td>0.049</td>
<td>0.004</td>
<td>0.926792</td>
<td>0.956258</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.089</td>
<td>0.084</td>
<td>-0.005</td>
<td>0.023567</td>
<td>0.097479</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.019</td>
<td>0.021</td>
<td>0.002</td>
<td>0.082892</td>
<td>0.20409</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.016</td>
<td>0.017</td>
<td>0.001</td>
<td>0.008662</td>
<td>0.058049</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.026</td>
<td>0.03</td>
<td>0.004</td>
<td>0.067288</td>
<td>0.179395</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.021</td>
<td>0.025</td>
<td>0.004</td>
<td>0.260197</td>
<td>0.426468</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.011</td>
<td>0.013</td>
<td>0.002</td>
<td>0.949973</td>
<td>0.970512</td>
</tr>
<tr>
<td>4</td>
<td>1.46E+08</td>
<td>0.043</td>
<td>0.039</td>
<td>-0.004</td>
<td>0.358502</td>
<td>0.524863</td>
</tr>
</tbody>
</table>

**CXCL2**

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>7496260</td>
<td>0.197</td>
<td>0.194</td>
<td>-0.003</td>
<td>0.535802</td>
<td>0.678067</td>
</tr>
<tr>
<td>4</td>
<td>7496493</td>
<td>0.096</td>
<td>0.092</td>
<td>-0.004</td>
<td>0.907881</td>
<td>0.944651</td>
</tr>
<tr>
<td>4</td>
<td>74964856</td>
<td>0.029</td>
<td>0.031</td>
<td>0.002</td>
<td>0.176206</td>
<td>0.330308</td>
</tr>
<tr>
<td>4</td>
<td>74964920</td>
<td>0.094</td>
<td>0.09</td>
<td>-0.004</td>
<td>0.165471</td>
<td>0.317061</td>
</tr>
<tr>
<td>4</td>
<td>74965068</td>
<td>0.032</td>
<td>0.036</td>
<td>0.004</td>
<td>0.021065</td>
<td>0.091874</td>
</tr>
<tr>
<td>4</td>
<td>74965075</td>
<td>0.053</td>
<td>0.047</td>
<td>-0.006</td>
<td>0.048382</td>
<td>0.146922</td>
</tr>
<tr>
<td>4</td>
<td>74965079</td>
<td>0.06</td>
<td>0.055</td>
<td>-0.005</td>
<td>0.287643</td>
<td>0.455425</td>
</tr>
<tr>
<td>4</td>
<td>74965135</td>
<td>0.059</td>
<td>0.067</td>
<td>0.008</td>
<td>0.014618</td>
<td>0.075483</td>
</tr>
<tr>
<td>4</td>
<td>74965154</td>
<td>0.053</td>
<td>0.05</td>
<td>-0.003</td>
<td>0.193753</td>
<td>0.351502</td>
</tr>
<tr>
<td>4</td>
<td>74965162</td>
<td>0.032</td>
<td>0.022</td>
<td>-0.01</td>
<td>0.57804</td>
<td>0.711429</td>
</tr>
<tr>
<td>4</td>
<td>74965226</td>
<td>0.013</td>
<td>0.012</td>
<td>-0.001</td>
<td>0.922298</td>
<td>0.953536</td>
</tr>
<tr>
<td>4</td>
<td>74965262</td>
<td>0.153</td>
<td>0.146</td>
<td>-0.007</td>
<td>0.507926</td>
<td>0.655305</td>
</tr>
<tr>
<td>4</td>
<td>74965278</td>
<td>0.082</td>
<td>0.08</td>
<td>-0.002</td>
<td>0.124559</td>
<td>0.264099</td>
</tr>
</tbody>
</table>

**HIF1A**

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>62161583</td>
<td>0.15</td>
<td>0.151</td>
<td>0.001</td>
<td>0.745886</td>
<td>0.836718</td>
</tr>
<tr>
<td>14</td>
<td>62161848</td>
<td>0.068</td>
<td>0.062</td>
<td>-0.006</td>
<td>0.001149</td>
<td>0.025102</td>
</tr>
<tr>
<td>14</td>
<td>62161878</td>
<td>0.039</td>
<td>0.039</td>
<td>0</td>
<td>0.451068</td>
<td>0.608173</td>
</tr>
<tr>
<td>14</td>
<td>62161958</td>
<td>0.005</td>
<td>0.008</td>
<td>0.003</td>
<td>0.014684</td>
<td>0.075682</td>
</tr>
<tr>
<td>14</td>
<td>62161964</td>
<td>0.007</td>
<td>0.009</td>
<td>0.002</td>
<td>0.198307</td>
<td>0.35693</td>
</tr>
<tr>
<td>14</td>
<td>62162064</td>
<td>0.011</td>
<td>0.013</td>
<td>0.002</td>
<td>0.834167</td>
<td>0.89724</td>
</tr>
<tr>
<td>14</td>
<td>62162178</td>
<td>0.009</td>
<td>0.011</td>
<td>0.002</td>
<td>0.061346</td>
<td>0.169453</td>
</tr>
<tr>
<td>14</td>
<td>62162214</td>
<td>0.038</td>
<td>0.037</td>
<td>-0.001</td>
<td>0.173315</td>
<td>0.326743</td>
</tr>
<tr>
<td>14</td>
<td>62162460</td>
<td>0.043</td>
<td>0.04</td>
<td>-0.003</td>
<td>0.007494</td>
<td>0.054124</td>
</tr>
</tbody>
</table>

*Δβ (case-control): Difference in median methylation between control and case groups

**cond_glm-p: Conditional linear regression p-value for differential methylation between cases and controls

***(fdr): FRD-corrected p-values
R script

Reading in and sub-setting data

# The data frame is called dat#

dat<-read.table("table.txt", header=TRUE, sep="\t")

# to retrieve basic summary statistics, including mean, median, interquartile range, min and max, for a linear variable e.g. age#

summary(dat$age)

# Standard deviation for age#

sd(dat$age)

# to retrieve numbers within each group for a categorical variable e.g. case_control status#

summary(dat$case_control)

# making index for categorical data i.e. subsetting data (e.g. case-control status)#

case<-which(dat$case_control==1)

control<-which(dat$case_control==0)

# Subsetting data groups based on multiple criteria (e.g. case_control & fasting_status)#

dat$groups[dat$case_control==0 & dat$fasting_status==0]=1

dat$groups[dat$case_control==0 & dat$fasting_status==1]=2

dat$groups[dat$case_control==1 & dat$fasting_status==0]=1

dat$groups[dat$case_control==1 & dat$fasting_status==1]=2
#To make this a factor (column in dat)#

dat$groups<-as.factor(dat$groups)

#Splitting data for a continuous variable into quantiles (in this case quintiles of age)#

#package gregmisc#

age_quintiles<-quantcut(dat$age, q=c(0, 1/5, 2/5, 3/5, 4/5, 5/5), na.rm=TRUE)

#To look at these quantiles#

summary(age_quintiles)

#To change which quantile is used as the reference quantile, e.g. the highest quintile (75,102)#

age_quintiles<- relevel(age_quintiles, "(75,102]")

#To make z-scores (for example, for methylation)#

#Z score=(x-mean)/sd

#first calculate mean and standard deviation of methylation for the population#

mean(dat$methylation, na.rm=TRUE)

sd(dat$methylation, na.rm=TRUE)

#then make a column of z-scores (z) by applying (x-mean)/sd to the methylation data#

dat$z<-((dat$methylation-mean(dat$methylation)))/(sd(dat$methylation))

**Statistical tests**

#student's T-test (example association of age (exposure) with methylation (outcome). Default is unmatched, 2-tailed T-test#

t.test(dat$methylation~dat$age)
# Wilcoxon rank sum test for same variables#

```r
wilcox.test(dat$methylation~dat$age)
```

# when I want to test an association within a sample subset e.g. controls#

```r
wilcox.test(dat$methylation[control]~dat$age[control])
```

# To test for significant differences in DNA methylation between different genotypes (in this case MTHFR genotype 'GG' vs. combined 'CG' and 'CC' genotypes#

# first make factor#

```r
dat$MTHFR[dat$MTHFR_genotype!="GG"]<-
```

```r
dat$MTHFR[dat$MTHFR_genotype=="GG"]<-
```

```r
dat$MTHFR<-as.factor(dat$MTHFR)
```

```r
wilcox.test(dat$methylation~dat$MTHFR)
```

# Chi-squared test (example association of age (exposure) with methylation (outcome))#

```r
chisq.test(table(dat$age,dat$methylation))
```

# ANOVA to test for difference in within group log(DNA methylation) variability between categories (e.g. case_control groups)#

```r
summary(aov(log(dat$methylation)~dat$case_control , data=dat))
```

# To find spearman correlation r-squared and p-value for two continuous variables (paired samples), e.g. correlation between methylation and time to diagnosis for each sample [cases]#

```r
cor.test(dat$methylation[case],dat$TTD[case], method=c("spearman"))
```

# Linear regression model with two linear variables (e.g. age (exposure) and methylation (outcome)#
# First see if data is normally distributed for each variable by plotting histograms#

hist(dat$methylation)

# If not, log data, and again use histogram to see if data is more normally distributed#

dat$log_methylation <- log(dat$methylation)

Methage <- glm(dat$log_methylation ~ dat$age)

summary(Methage)

# To adjust for other variables such as case-control status and smoking status and for interactions between these#

Methage_adjusted <- glm(dat$log_methylation ~ dat$age + dat$case_control + dat$smoking_status + dat$case_control:dat$smoking_status)

# Logistic regression with categorical variables, e.g. DNA methylation quartiles (outcome) and case_control status (exposure)#

# first make methylation quartiles (package gregmisc)#

meth_quartiles <- quantcut(dat$methylation, q=c(0, 1/4, 2/4, 3/4, 4/4), na.rm=TRUE)

then run model

summary(glm(meth_quartiles ~ dat$case_control))

# To get odds ratio for each quartile, exponentiate the estimates#

exp(estimate)

# Conditional logistic regression using a matching variable (indicating which samples represent matched pairs) (survival package)#

# This time treating methylation as the exposure and case_control status as the outcome#

summary(clogit(dat$case_control ~ dat$methylation + strata(dat$matching_variable)))
plotting

# boxplot e.g. methylation stratified by case_control status#
boxplot(dat$methylation~dat$case_control)

# or #
boxplot(dat$meth[case],dat$meth[control]) # Usually used if multiple boxes are required#

# Strip-chart (e.g. methylation stratified by case_control status) with jitter distribution and vertical strips#
stripchart(dat$methylation~dat$case_control, vertical=T, method='jitter')

# scatterplot for methylation vs. age, with line of best fit#
# first make a line of best fit for methylation vs. age#
line<-lm(dat$methylation~dat$age)

# Make plot with labels#
plot(dat$methylation~dat$age, ylab="methylation", xlab="age", main="methylation Vs. age", cex.axis=1.5, cex.lab=1.5, cex.main=1.5)
abline(line, col="red")

# Add text at position 110,5#
text(110,5, labels="rsq=-0.2, p=0.026")

Meta-analyses

# make a data frame (meta) including study names, estimates (e.g. OR), standard error for the estimates and sample n#

# To conduct a meta_analysis with random effects model, weighted by study size (n) (package rmeta)#
meta_analysis=meta.summaries(meta$OR,meta$SE,method="random",names=meta$study, weights=meta$n)

# To plot this model as a forest plot with title, adjusted axis label sizes and a vertical line at 0#
plot(meta_analysis, main="meta-analysis", cex.lab=1.5, cex.axis=1.5)
abline(v=1)

**450k array analysis**

#Reading in 450k array data#

mval<-read.table("M_values.txt", header=T, sep="t")
annot<-read.table("annotation_file.txt", header=T, sep="t")
pd2<-read.table("patient_information.txt", header=T, sep="t")

mval<-as.matrix(mval)
annotmat<-as.matrix(annot)
pd2<-as.data.frame(pd2)
fasting<-as.factor(pd2$fasting)

#logistic regression model to identify loci associated with fasting or case-control status#

#function code (provided by Dr. Charlotte Wilhelm-Benartzi)

ggGLMByLocus <- function(ExData,model,LList=NULL,family=gaussian(),data=NULL,...){
  if(is.null(data)) Xframe <- model.frame(model,na.action = na.pass)
  else Xframe <- model.frame(model,data=data,na.action = na.pass)
  #Xmiss <- apply(is.na(as.matrix(Xframe)),1,any)
  X <- model.matrix(model,data=Xframe)
  #ExData <- ExData[Xmiss,]

  if(is.null(LList)) {
    #rest of the code here
  }
}
LList <- list(matrix(0,nrow=1,ncol=dim(X)[2]))

LList[[1]][2] <- 1

names(LList) <- colnames(X)[2]

nLoci <- dim(ExData)[2]
nLmat <- length(LList)

PVAL <- COEF <- matrix(NA,nLmat,nLoci)

rownames(PVAL) <- rownames(COEF) <- names(LList)

colnames(PVAL) <- colnames(COEF) <- colnames(ExData)

for(j in 1:nLoci){
  fit <- glm(ExData[,j]~X-1, family=family)

  for(l in 1:nLmat){
    stat <- LList[[l]] %*% fit$coef
    if(length(stat)==1) COEF[l,j] <- stat
    else COEF[l,j] <- sum(stat*stat)

    V <- LList[[l]] %*% vcov(fit) %*% t(LList[[l]])

    xstat <- t(stat) %*% solve(V, stat)

    PVAL[l,j] <- 1-pchisq(xstat,dim(LList[[l]])[1])
  }
}

out <- list(coef=COEF, pv=PVAL)
class(out) <- "ggGLMByLocus"

out
# Running model. Output is a matrix consisting of two columns including logistic regression estimates and p-values for association of DNA methylation at all 450k probes with a categorical variable with two levels (in this case fasting --status) #

```r
glm_B2ck<- ggGLMByLocus(mval, ~fasting)
```

#converting list of coefficients and p-values for each probe to a data frame#

```r
coefficient=glm_B2ck$coef
pval=glm_B2ck$pv
coeff<-as.vector(coefficient)
pval<-as.vector(pval)
results=cbind(coef, pval)
rownames(results)=rownames(mval)
```

#Making volcano plots where -log10 (p-value) is on Y axis, and estimate is on the X axis#

```r
plot(coef, -log10(pvals), main="Volcano plot 450K data Breast II by significant p values", xlab="Linear model coefficient for Fasting status effect on methylation", ylab="-Log P value")
```

#Adding FDR-corrected p-value (q) column to 'results' data frame (model estimates and p-values for each locus)#

```r
results$q=p.adjust(results$pval, method="fdr")
```

#restricting hits to those with q-value <0.01#

```r
hits<-which(results$q<0.01)
resultstop=results[hits,]
```

#Sorting results on q-value (from smallest to largest)#

```r
hits_sort<-results[order(results$q),]
```
#Merging hits_sort file with annotation file for array#

annohits<-merge(annot,_hits_sort, by="row.names")
write.table(annohits, file = "annohits.txt", sep="\t")

#Making median non-fasting and fasting values, and delta beta values for all significant hits#

non_fasting<-which(pd2$fasting==0)
fasting<-which(pd2$fasting==1)
med_non_fasting<-apply(bval[non_fasting],1,median,na.rm=T)
med_fasting<-apply(bval[fasting],1,median,na.rm=T)
nonfast<-as.matrix(med_non_fasting)
fast2<-as.matrix(med_fasting)
fast<-data.frame(nonfast,fast2, Row.names=rownames(nonfast))
annohitsck<-merge(annohits,fast, by="Row.names")
fast$diff<-(fast$nonfast-fast$fast2)

#Adding data from Price et al re-annotation file#

reannot<-read.table("Price_450k_Reannotation.txt", sep="\t", header=TRUE)
row.names(reannot)<-reannot$ID

#Making large data frame with all data#

methdiffs<-cbind(results, medianunfast4, medianfast4)
methdiffs$diff<-methdiffs$medianunfast4-methdiffs$medianfast4)
row.names(reannot)<-reannot$ID
methdiffsreannot<-merge(methdiffs, reannot, by="row.names")

methresults_annot<-merge(annot,methdiffs, by="row.names")

meth_annot_reannot<-merge(methresults_annot,reannot, by="Row.names")

methresults_sort<-meth_annot_reannot2[order(meth_annot_reannot2$q),]

save(methresults_sort,file="methresults_070513.Rdata")

##Restricting methresults_sort to q<0.05##

hits2<-which(methresults_sort$q<0.05)

methHITS<-methresults_sort[hit2,]

write.table(methHITS, file = "methHITS.txt", sep="\t")

#volcano plot for figure 4.12#

plot(methHITS$diff1,-log(methHITS$q), main="450k loci significantly associated with fasting status", ylab="-log(q-val)", xlab="methylation difference (fasting-non-fasting)")

abline(v=0)

abline(v=0.5,lty=2)

abline(v=0.05,lty=2)

abline(v=-0.05,lty=2)

#screening out polymorphic CpG sites#

xx<-which(methHITS$Target.CpG.SNP=="")

methhitsxx<-methHITS[xx,]

#making data-frame restricted to loci with median beta-value different >5 between fasting and non-fasting groups#
greatersdiff<-which(methresults_sort$diff1>0.05)

b2diff<-methresults_sort[greatersdiffb2,]

#Writing this to a table#
write.table(b2diff, file = "b2diff.txt", sep="\t")

#Finding p-value for Wilcoxon rank sum test for fasting analysis for all loci##
FAST<-which(pd2$fasting==1)
NONFAST<-which(pd2$fasting==0)

#Applying Wilcoxon rank sum test across all probes on array to find probes significantly differentially methylated by fasting status#
wilc<-rep(NA, nrow(mval))
for(i in 1:nrow(mval))
{
  wilc[i]<-wilcox.test(bvalb3[i,FAST],bvalb3[i,NONFAST])$p.value
}

#FDR adjusting Wilcox p-values#
wilc_fdr<-p.adjust(wilc, method="fdr")
fdr_hits<-which(wilc_fdr<0.05)
FDR_HITS<-wilc[fdr_hits,]

#Make data frame with new hits (wilcoxon rank sum test) merged with annotation and re-annotation files#
fdr_hits_annot<-merge(annot,fdr_hits, reannot, by="row.names")
# Removing polymorphic CpG sites #

fdr_hits_annot$noSNP[fdr_hits_annot$Target.CpG.SNP=="""]=0
fdr_hits_annot$noSNP[fdr_hits_annot$Target.CpG.SNP!=""]=1
fdr_hits_annot$noSNP<-as.factor(fdr_hits_annot$noSNP)
nosingnucpol<-which(fdr_hits_annot$noSNP==0)
br2c2b1_hits_noSNP<-fdr_hits_annot[nosingnucpol,]
br2c2b1_hits_noSNP$unique[br2c2b1_hits_noSNP$AlleleA_Hits!=1]=0
br2c2b1_hits_noSNP$unique[br2c2b1_hits_noSNP$AlleleA_Hits==1]=1
promisc<-which(br2c2b1_hits_noSNP$unique==0)
uni<-which(br2c2b1_hits_noSNP$unique==1)
br2c2b1_hits<-br2c2b1_hits_noSNP[uni,]

# Volcano plot for these hits #

plot(br2c2b1_hits$diffbreast2C2, -log10(br2c2b1_hits$breast2C2_fdr), col=br2c2b1_hits$noSNP, main="450K methylation diff by significant p values", xlab="methylation difference", ylab="-Log10 P value")

# Writing to a data frame #

write.table(br2c2b1_hits, file = "br2c2b1_hits.txt", sep="\t")

# Plotting difference in methylation between fasting and non-fasting against gene locus for an individual gene (e.g. LGR6) (fig 4.15) #

# First make index for gene 

LGR6<-which(annot$UCSC_RefGene_Name="LGR6")
png(file="LGR6.png", width=960, height=480)
plot(annot$MAPINFO[LGR6], as.numeric(br2c2b1_hits$med_non_fasting[LGR6])-as.numeric(br2c2b1_hits$med_fasting[LGR6]))

title("Plot Delta Beta Value- LGR6 by location")

dev.off()