Exploring the transcriptional and translational implications of DNA methylation in breast cancer

Dr Natalie Susannah Shenker

The Epigenetics Unit, Division of Cancer, Department of

Surgery and Cancer, Imperial College London

A thesis submitted for the degree of Doctor of Philosophy

September 2015

ABSTRACT

The significance of DNA methylation beyond the gene promoter is little understood. High levels of intragenic methylation (IGM) correlate with active gene expression in lower organisms. Pyrosequencing assays across oestrogen receptor alpha (*ESR1*) in ESR1-positive and negative breast cancer cell lines showed higher IGM levels in ESR1-positive cells, with reduced *ESR1* expression after demethylation with decitabine. However, different *ESR1* IGM patterns were determined in breast cancer biopsies, probably because of cell heterogeneity.

DNA methylation can reflect cellular environmental exposures. A methylation signature of smoking was determined in a European Prospective Investigation into Cancer and Nutrition epigenome-wide association study (EWAS, 450K array), in case-control pairs of breast (n=92) and colorectal cancer (n=95). Differentially methylated loci ($p=1x10^{7}$) were validated independently by pyrosequencing in peripheral mononuclear blood cells from 102 never, 45 former and 33 current smokers (n=180). Validated loci included intragenic loci in *AHRR*, involved in dioxin metabolism, and an intergenic region, 2q37, that was associated with breast cancer risk (p=0.003). Four loci comprised an index that outperformed the gold standard test for former smoking, cotinine.

DNA methylation signatures and exposure levels vary between cell types, but few studies have exclusively investigated the target for breast carcinogenesis, the breast epithelial cell. Therefore, a proof-of-principle study into cells from breast milk samples (n=150 milk bank donors) characterised >97% as epithelial (EMA-positive) and >80% as ESR1-positive by flow cytometry. Extracted DNA and RNA underwent methylation, genotyping and expression analyses. *ESR1* IGM patterns matched

ESR1-positive tumour cell lines. Furthermore, 2q37 was hypomethylated in 19 former smokers compared to 69 non-smokers ($p=0.026$).

To conclude, differentially methylated non-promoter loci may be functionally important, and useful as biomarkers, but tissue selection for prospective epigenetic studies should be carefully considered. Future large-scale breast milk sample collections will inform population-level studies into environmental exposures, breast cancer risk, and mammary biology.

TABLE OF CONTENTS

ABSTRACT 1
LIST OF TABLES 7
LIST OF FIGURES 9
ABBREVIATIONS 13
PUBLICATIONS ARISING FROM THIS THESIS 14
DECLARATION OF ORIGINALITY 15
COPYRIGHT DECLARATION 15
ACKNOWLEDGEMENTS 16
CHAPTER 1 19
INTRODUCTION: BREAST CANCER AND THE ROLE OF EPIGENETIC RESEARCH 19
1.1 Breast cancer 20
1.1.1 Overview 20
1.1.2 The normal breast 20
1.1.3 Breast cancer 23
1.1.4 Incidence, mortality and prognosis 25
1.1.5 Aetiology and risk factors for breast cancer 29
1.1.6 Signs and symptoms 35
1.1.7 Diagnosis and screening 36
1.1.8 Genetic basis of breast cancer 36
1.2 Epigenetics 39
1.2.1 Overview 39
1.2.2 DNA methylation 39
1.2.3 Intragenic methylation 42
1.2.4 DNA methylation and cancer 44
1.2.5 DNA methylation in breast cancer 46
1.3 Doctoral hypotheses and aims 48
1.3.1 Hypotheses 48
1.3.2 Aims 48
CHAPTER 2 49
THE TRANSCRIPTIONAL AND TRANSLATIONAL IMPLICATIONS OF INTRAGENIC DNA METHYLATION IN BREAST CANCER 49
2.1 INTRODUCTION 50
2.1.1 Intragenic methylation 50

2.1.2 Oestrogen receptor alpha (ESR1) and its role in breast cancer 51
2.1.3 The epigenetics and function of ESR1: what was known at the outset of this thesis 54
2.1.4 Aim of this study 56
2.2 METHODS 58
2.2.1 Cell lines 58
2.2.2 Decitabine (DAC) treatment 58
2.2.3 Methylation analyses 59
2.2.4 RNA isolation, cDNA synthesis and RT-qPCR 62
2.2.5 Breast tumour samples 63
2.2.6 Cell cycle analysis 65
2.2.7 Antisense siRNA transfection studies 66
2.2.8 Statistical analysis 66
2.3 RESULTS 68
2.3.1 Cell growth 68
2.3.2 Differences in* ESR1 *IGM patterns between ER-pos and ER-neg cell lines 68
2.3.3 In vitro* ESR1 *methylation changes after DAC treatment 69
2.3.4 Expression levels of *ESR1* mRNA 74
2.3.5* ESR1 *IGM levels in DNA from breast tumour biopsies 77
2.3.6 Cell cycle analyses 78
2.3.7 Cell proliferation 81
2.3.8 Antisense mRNA levels of *ESR1* 81
2.4 DISCUSSION 84
2.4.1 Limitations 89
2.4.2 Conclusions 91
CHAPTER 3 93
INVESTIGATING THE DNA METHYLATION SIGNATURE OF SMOKING AND ITS ASSOCIATION WITH BREAST CANCER RISK 93
3.1 INTRODUCTION 94
3.2 METHODS 99
3.2.1 Study participants 99
3.2.2 Microarray hybridisation (conducted by Dr Silvia Polidoro at the HuGeF Foundation, Turin, Italy) 100
3.2.3 Validation of microarray data findings 100
3.2.4 Human lung analyses 101
3.2.5 Murine model of smoking 102
3.2.6 Development of the methylation index to predict smoking status 103

3.2.7 Data analysis (conducted by myself and Dr. James Flanagan) **105**
3.3 RESULTS **107**
3.3.1 Epigenome-wide association studies **107**
3.3.2 Validation of microarray loci of interest by pyrosequencing **111**
3.3.3 Human lung tissue methylation and expression analyses **118**
3.3.4 Expression analyses in a mouse model of smoking **118**
3.3.5 Development of the methylation model for smoking **119**
3.4 DISCUSSION **124**
3.4.1 General discussion 124
3.4.2 Methylation index of former smoking 130
3.4.3 Longevity of methylation changes 132
3.4.4 Limitations 133
3.4.5 Conclusions 134
CHAPTER 4 137
ASSESSING THE FEASIBILITY OF THE USE OF EPITHELIAL CELLS EXTRACTED FROM BREAST MILK FOR EPIGENOMICS 137
4.1 INTRODUCTION 138
4.1.1 Breast epithelial cells from milk for breast cancer research 140
4.1.2 Aims 141
4.2 METHODS 143
4.2.1 Breast milk collection 143
4.2.2 Cell pellet generation 144
4.2.3 Cell characterisation 145
4.2.4 DNA extraction and bisulphite conversion 147
4.2.5 Genotype analyses 149
4.2.6 RNA extraction and analysis 151
4.2.7 Statistical analysis 153
4.3 RESULTS 155
4.3.1 Establishment of the QCCH Milk Bank collaboration 155
4.3.2 Cell pellet generation and cell characterisation 156
4.3.3 DNA extraction and quality control 158
4.3.4 450K analysis and pyrosequencing 161
4.3.5 RNA analysis 164
4.2.6 Genotyping results 167
4.3.7 Combining meQTL data into the polygenic risk score for breast cancer 170
4.3.8 The methylation signature of former smoking in donor samples 171

4.4 DISCUSSION ... 173
4.4.1 Breast epithelial cells: their role in breast cancer ... 173
4.4.2 Initial sample characterisation, handling and troubleshooting ... 176
4.4.3 Data generation ... 179
4.4.4 Limitations ... 181
4.4.5 Conclusions ... 183
CHAPTER 5 ... 184
DISCUSSION, AVENUES FOR FUTURE RESEARCH AND CONCLUSIONS ... 184
5.1 GENERAL DISCUSSION ... 185
5.2 FUTURE RESEARCH DIRECTIONS ... 196
5.2.1 Breast feeding and cancer risk ... 196
5.2.2 The Breast Milk Epigenetic Cohort Study (BECS) ... 200
5.2.3 The UK Milk Bank Network: purpose and scope ... 201
5.2.4 HAMLET ... 205
5.3 LIMITATIONS AND STRENGTHS ... 207
5.4 CONCLUSIONS ... 210
6. REFERENCES ... 211
7. APPENDIXES ... 242

LIST OF TABLES

Table 1.1: Principal stages of embryonic breast development (Knight and Peaker, 1982).

Table 1.2: Expression profiles of breast cancer molecular subtypes, adapted from (Schnitt, 2010).

Table 1.3: The 1-, 5-, 10- and 20-year survival rates for women aged 15-99 years in England and Wales.

Table 2.1: Cell counts (number of cells per ml x 10^5) after the completion of the DAC incubation experiments.

Table 2.2: Ki67 scores of stained tumour sections.

Table 3.1: Organs at risk of carcinogenesis as a result of carcinogens in tobacco smoke, as assessed by the International Agency for Research on Cancer Working Groups.

Table 3.2: Characteristics of the three cohorts investigated in this study.

Table 3.3: Leading differentially methylated genomic loci between smokers and former or non-smokers in a breast and colonic case-control cohort.

Table 3.4: Relationship to case-control status adjusted for smoking status, and for oestrogen receptor status in breast and colon cancer.

Table 3.5: Relationship between methylation and smoking intensity.

Table 3.6: Mean methylation percentages for the four cytosine-guanine nucleotides identified from the microarray study as differentially methylated in current smokers vs. former and non-smokers, in addition to the mean cotinine values (ng/ml) in each group of individuals.

Table 3.7: Cotinine levels (ng/ml) in the test sample set (n = 81).

Table 4.1: Primer sequences used for the five SNP PCR assays and genotyping reactions.

Table 4.2: DNA concentrations for 49 samples from Picogreen assay compared with Nanodrop concentrations.

Table 4.3: Results of triplicate repeat pyrosequencing assays across *ESR1* in bisulphite-converted DNA samples from donated EBM (n = variable, stated in table).
Table 4.4: RNA Integrity Scores (RIN) for the 11 RNA samples from fresh EBM samples and one frozen sample.

Table 4.5: Fold change results from the RT-qPCR assay for *ESR1* mRNA in 11 EBM samples, MCF7 and MDA-MB-231, relative to GAPDH.

Table 4.6: Genotype frequencies of the five breast cancer risk-associated SNPs selected from the OncoArray data in 80 EBM samples (minor allele underlined).

Table 4.7: Validation of OncoArray SNP data at five breast cancer risk-associated SNPs with genotyping assays.

LIST OF FIGURES

Figure 1.1: Normal female breast anatomy, courtesy of *Breast Cancer Treatment (PDQ)®,* published by the National Cancer Institute of the NIH, USA (NCI, 2014).

Figure 1.2: Current trends in incidence and mortality rates of all cancers and breast cancer in the UK.

Figure 1.3: Relative risk of developing breast cancer according to age at first live birth (data derived from Brinton et al. (Brinton et al., 1983).

Figure 1.4: The addition of a methyl group to the 5-carbon of a cytosine base, enzymatically catalysed by a member of the DNA methyltransferase (Dnmt) family.

Figure 1.5: Schematic of the results of Zilberman et al. showing that genes with high expression levels have higher levels of intragenic methylation (Zilberman et al., 2007, Shenker and Flanagan, 2012).

Figure 2.1: A schematic structure of the ESR1 protein. Abbreviations: AF1: activation function-1 domain; DBD: DNA-binding domain; H: hinge; LBD: ligand binding domain. NH2 and COOH indicate the 5' and 3' end of the protein, correlated to the coding mRNA. Adapted from Robinson et al., 2013.

Figure 2.2: A schematic showing the position of pyrosequencing assays across the *ESR1* gene.

Figure 2.3: Results of the MTT assays conducted in all six breast cancer cell lines to assess the cytotoxicity of decitabine (DAC).

Figure 2.4: The DNA methylation levels across the *ESR1* gene in three ER-pos cell lines (blue) and three ER-neg cell lines (red).

Figure 2.5: The effect of decitabine on DNA methylation in breast cancer cell lines. A) Composite graphs showing the DNA methylation levels across the ESR1 gene in three ER-neg cell lines treated with DAC (yellow) compared to DMSO-treated controls (red). Data was collected in triplicate in each cell line at each locus, and the standard error of the mean was then calculated (error bars); * $p < 0.05$, Bonferroni corrected t-test. B) DNA methylation across ESR1 gene in ER-pos cell lines treated with DAC (purple) compared to untreated (blue). The loci assessed can be mapped to specific regions of the ESR1 gene using the schematic shown in Figure 2.1.

Figure 2.6: RT-qPCR gene expression in individual cell lines treated with DAC in A) three ER-neg cell lines, and B) three ER-pos cell lines.

Figure 2.7: This heat map shows the relative protein expression of very highly expressed genes in MCF7 cells (n = 40 genes) that are downregulated after DAC treatment, compared to DAC-treated MDA-MB-231 cells from two publically available datasets (gse10613 and gse13733).

Figure 2.8: A) Pyrosequencing-based methylation data across the ESR1 gene in fresh frozen (FF) human breast tumour samples (B) In ER-positive FF tumours, methylation levels were very similar between grades.

Figure 2.9: 450K Beadchip methylation analysis of ESR1 methylation from TCGA breast tumours.

Figure 2.10: The number of MCF7 cells in the G0/G1 phase of the cell cycle increased significantly with decreasing concentrations of serum in the medium.

Figure 2.11: A, B) Expression changes for antisense across ESR1 in untreated (blue) and DAC-treated MCF7 and MDA-MB-231 (red). C, D) Antisense expression changes in all six cell lines pre- and post-DAC treatment at position p3 (2nd intron of *ESR1*);

Figure 3.1: Manhattan plot and quantile-quantile plot for EWAS results for smoking status in two case-control studies.

Figure 3.2: Box-plot representation of pyrosequencing-derived methylation percentages for the first locus identified from AHRR (cg23576855) in the validation control population (n = 180).

Figure 3.3: Association between smoking status and AHRR.

Figure 3.4: Association between smoking status and 2q37.

Figure 3.5: Smoking intensity is directly correlated with hypomethylation at *AHRR* and *2q37.1* in smokers.

Figure 3.6: Box plots of distribution for methylation at four genomic loci for never, former, and current smokers.

Figure 3.7: Smoking induces methylation and expression changes in human and mouse lung tissues.

Figure 3.8: ROC for prediction of former smoking status based on MI.

Figure 3.9: A) MI compared with duration of smoking in former smokers. B) MI versus time since quitting in former smokers.

Figure 3.10: UCSC genome browser image of genomic locus 2q37, which is demethylated in current and former smokers. This region maps to a cluster of DNaseI hypersensitivity regions (shown in grey shading), as well as a CpG island (http://genome.ucsc.edu/).

Figure 3.11: The AHR pathway for the metabolism of dioxins and dioxin-like compounds, found in cigarette smoke.

Figure 4.1: Flow diagram of the flow cytometry protocol, showing the controls and antibody steps used.

Figure 4.2: A) Number of donors recruited monthly throughout the pilot study (August 2014 – August 2015) from a single Milk Bank; B) number of samples (20-25 ml) collected throughout the sample period.

Figure 4.3: Flow cytometry results for epithelial membrane antigen and ESR1.

Figure 4.4: A) Flow cytometry for epithelial membrane antigen (EMA) from a single donor (#21) compared to negative control unstained cells; B) cells from the same donor stained with anti-ESR1 primary antibody + PE secondary antibody vs. PE-stained control cells.

Figure 4.5: Scattergram to show that there was no correlation between cell counts and DNA concentration).

Figure 4.6: A, B) Standard curve generated from two Picogreen assays of DNA standards (conc. 0 – 20 ng/µl), with averages generated from triplicate repeats. A and B relate to the two different batches (A = donor IDs 1-4, B = donor IDs 25-49).

Figure 4.7: COMBAT analysis for the three chips of EBM samples (n = 36 in total, 12 in each chip) showing a high level of similarity in the bimodal peaks of the three colours for each.

Figure 4.8: A) Methylation values from 450K analysis of 36 DNA samples extracted from breast epithelial cells from expressed breast milk. B) Validation of the methylation levels using pyrosequencing in a separate set of 53 EBM samples (black solid line).

Figure 4.9: Quality control data from the Bioanalyser assays.

Figure 4.10: RT-qPCR analysis of 11 EBM epithelial cell RNA samples showing high levels of ESR1 expression normalised against ER-pos MCF7 and ER-neg MDA-MB-231 cell lines.

Figure 4.11: Heat map to show correlation between genotype and methylation levels across the *ESR1* gene in 12 DNA samples from EBM. X-axis probes, Y-axis, samples.

Figure 4.12: Differential methylation depending on genotype at a single CpG site, cg12219469.

Figure 4.13: Results of pyrosequencing for the top marker previously associated with former smoking in white blood cell DNA, showing hypomethylation in former smokers (n = 19) compared to never smokers (n = 69).

Figure 4.14: The cellular structure of human breast milk ducts and surrounding stroma, reproduced from (Visvader and Stingl, 2014).

Figure 5.1: Prevalence of exclusive breastfeeding in the UK at ages up to 6 months in 2005 and 2010 (reproduced from the Infant Feeding Survey, http://www.hscic.gov.uk/catalogue/PUB08694/ifs-uk-2010-chap2-inc-prev-dur.pdf).

Figure 5.2: ChIP-PCR data for *ESR1* in a single EBM sample.

ABBREVIATIONS

BCC	Breast Cancer Campaign (renamed Breast Cancer Now in April 2005)
CpG	Cytosine-guanine dinucleotide
DAC	Decitabine
DCIS	Ductal carcinoma in situ
DMSO	Dimethyl sulphoxide
EBM	Expressed breast milk
ERE	Estrogen response element
FBS	Fetal bovine serum
IGM	Intragenic methylation
IHC	Immunohistochemistry
MIQE	Minimum Information for Publication of Quantitative Real-Time PCR Experiments
MSP	Methylation specific PCR
MVR	Methylation variable region
PBS	Phosphate buffered saline
RPFNA	random periareolar fine needle aspiration
SNP	Single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
CGI	CpG islands
DMEM	Dulbecco's modified Eagle's medium
ER/ESR1	Oestrogen receptor-alpha
FCS	Fetal calf serum
FFPE	Formalin-fixed paraffin embedded
LINE	Long-interspersed nuclear elements
PCR	Polymerase chain reaction
PI	Propidium iodide
SD	Standard deviation
STR	Short tandem repeat
TSS	Transcription start site

PUBLICATIONS ARISING FROM THIS THESIS

1. **Shenker N**, Flanagan JM. Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research. *Br J Cancer*. 2012 Jan 17; 106 (2): 248-53. Epub 2011 Dec 13. Review.

2. **Shenker NS,** Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, Belvisi MG, Brown R, Vineis P, Flanagan JM. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Hum Mol Genet.* 2013 Mar 1;22(5):843-51.

3. **Shenker NS**, Ueland PM, Polidoro S, van Veldhoven K, Ricceri F, Brown R, Flanagan JM, Vineis P. DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology*. 2013 Sep; 24 (5):712-6.

4. **Shenker NS**, Dai W, Flower KJ, Bell E, Gore E, El Bahrawy M, Brown R and Flanagan JM. Transcriptional implications of intragenic DNA methylation in the oestrogen receptor alpha gene in breast cancer. BMC Cancer. 2015 May 1;15(1):337.

5. Flower KJ, **Shenker NS**, El-Bahrawy M, Goldgar DE, Parsons MT, KConFab Investigators, AFFECT study group, Spurdle AB, Morris JR, Brown R and Flanagan JM.DNA methylation profiling to assess pathogenicity of BRCA1 unclassified variants in breast cancer. *Epigenetics*

DECLARATION OF ORIGINALITY

I, Natalie Shenker, declare that this thesis is solely my own work. I have consulted all of the references myself during the preparation of this manuscript. Except where stated, the work presented in this thesis was performed by myself.

COPYRIGHT DECLARATION

The copyright of 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 the license terms of this work clear to others.

ACKNOWLEDGEMENTS

I am deeply appreciative of both of my supervisors, Dr James Flanagan and Professor Bob Brown. James took two chances on me, firstly by letting me come into his group with no previous experience in, or much knowledge of, epigenetics. He took another chance in letting me explore the use of breast milk for epigenetic research in the final year of my thesis, and I hope not to have let him down on either front. He has been a constant, even daily, source of support over my doctoral years, enabling me to work flexibly, and expressing only patience when things did not quite go to plan. My eldest daughter calls him Magic James, and I can only agree.

I thank Professor Brown for his time, words of wisdom, and interest in my work. He has created a group based on mutual support and friendship, which drives productivity. Most importantly, he has given me the confidence to start asking my own questions, and guidance in terms of how to go about answering them.

A great number of colleagues have taught me techniques and taken time to train me in data analysis. To list everyone would take more space than this thesis allows, but I would be remiss not to mention Dr Kirsty Flower, Michelle Metcalf, Angela Wilson, Dr Jane Borley, Elham Shamshaei, Dr Wei Dai, Dr Constanze Zeller, Dr Ed Curry, Dr Charlotte Wilhelm-Benartzi and Dr Kevin Brennan.

I was funded by the Medical Research Council through the DTP scheme at Imperial, and consumables were funded by the Breast Cancer Campaign Scientific Fellowship received by Dr James Flanagan. Funding for the BECS pilot study was awarded in the form of a Pilot Grant Award from Breast Cancer Campaign (£20,000), which I co-wrote with Dr James Flanagan and on which I am a co-investigator.

The work in this thesis would not have possible without a number of collaborations. Professor Maria Belvisi's group at the National Lung Institute of Imperial College provided samples of lung tissue and blood from human patients and mouse models that enabled us to examine the methylation profile of smoke-exposed lung tissue. Professor Paolo Vineis of Imperial College and Human Genome Foundation (HuGeF, Turin) provided 450K data and DNA samples from the EPIC study (European Prospective Investigation into Cancer and Nutrition into the effect of human nutrition on cancer risk) that enabled me to validate a smoking exposure methylation signature and develop a methylation index that was a powerful predictive tool for former smokers compared to current gold standards.

The breast milk collection component of the project was hugely enabled by the cooperation and problem solving capabilities of Gillian Weaver, Milk Bank Manager, and her staff, particularly Jennifer Sangalang, at the Queen Charlotte and Chelsea Hospital Milk Bank. Moreover, I thank the donors to the Milk Bank who provided consent for their milk to be used in research. Our work has also been supported by the United Kingdom Association of Milk Banking, a charity of which I am now a trustee with particular responsibility for planning a research strategy for donated human milk.

Beyond the laboratory, I simply wouldn't have been able to commence this PhD without the support of a great number of people, including Dr Mark Sullivan, Dr Nisha Shah, Una Hunt, Maureen ni Fiann, Dr Colin Ferrie, and Dr Clare Gerada.

My greatest support, inspiration and ability to focus was provided unconditionally by my family, particularly my husband Andrew, my mother Susan and father Louis, and my daughters, Sophia and Arielle.

For Hilda, Angela and Myrna.

CHAPTER 1

INTRODUCTION: BREAST CANCER AND THE ROLE OF EPIGENETIC RESEARCH

1.1 Breast cancer

1.1.1 Overview

Breast cancer is one of the leading causes of cancer worldwide, alongside cancers of the lung, colorectum, liver and stomach (WHO, 2009). It is the foremost cause of death in females, causing 521,000 deaths worldwide in 2012, of which almost 99% were women (WHO Report, 2008). The incidence of breast cancer also continues to increase, particularly in developed countries and in developing countries that are increasingly adopting a Western lifestyle. During the course of this doctorate alone, the lifetime risk of the disease in the UK has increased from 1 in 9 women to 1 in 8 (Sasieni et al., 2011, CRUK Report, 2012). Much of this increased incidence can be explained by improved screening techniques and nationwide awareness programmes, which have led to improved detection rates (Berry and Ravdin, 2007), but increased longevity and lifestyle changes (e.g., obesity, alcohol, delayed child bearing, reduced breastfeeding rates) have also increased risk. However, the morbidity and cost implications of such a large disease burden elevate breast cancer prevention programmes to the top of policy decision-making and public awareness.

1.1.2 The normal breast

Human breasts are bilateral mammary glands with the principal function of producing and secreting milk to provide nutrition to infants. The mammary gland derives its name from mamma, the Latin term for breast. Mammary glands are present in both males and females but are unique as human organs in that they are undeveloped at birth, but undergo further developmental stages in females at menarche (the onset of menstruation and full female puberty). Breast development is predominantly under

the control of progesterone (Hilton et al., 2012), and occurs over several stages during life. The main stages of embryonic development are described in Table 1.1.

Table 1.1: Principal stages of embryonic breast development (Knight and Peaker, 1982).

Embryonic week	Breast development
4th week	The embryo is 2.5 mm long, and the milk streak develops on the anterior surface to both sides of the midline. Ectoderm invaginates into the underlying mesodermal tissue and can be seen as a thickened ridge.
5th week	Milk line or ridge starts to develop. The underlying mesoderm is compressed into concentric layers, which will become the glandular stroma.
6th week	Mammary glands start to develop. The ectodermal cells become 'pear-shaped' and continue to grow into the mesoderm in a branching network that will become the ducts of the future breast, comprising lobes, lobules and alveoli.
16th week	Approximately 15-25 ectodermal strips coalesce in the subcutaneous tissue. These will form the breast lobules.
32nd - 40th week	Primary milk ducts have appeared and are visible histologically. The secondary mammary structures develop, including the hair follicles and sebaceous glands.

Terminal differentiation of the breast takes place during pregnancy to enable lactation, before involution after the menopause. The mammary gland is a tubuloalveolar gland with multiple branches that is embedded in a stroma of irregular connective tissue (Figure 1.1). Upon the commencement of lactation, milk is synthesized from the distal differentiated acini of the gland and conveyed proximally through a series of ducts to the nipple. From distal to proximal, the glands are arranged as clustered alveoli, which are connected to intralobular ducts that flow into

lobules. The mammary ducts are lined by two types of differentiated epithelial cell: the inner layer of luminal epithelial cells produce milk during lactation, and the outer myoepithelial cell layer lies on a basement membrane. These cells contract to propel the milk through the ductal network to the teat cistern.

The majority of breast cancers arise from cells within the ductal network (NCI, 2014), and changes in the normal composition of breast ducts constitute an important early stage in carcinogenesis (Hilton et al., 2013). Examination of the normal histology of breast tissue is therefore important to understand the normal functioning of cells that will become cancers (Hassiotou and Geddes, 2013). There are three main types of cell in the mammary ductal system: stem cells (Hassiotou et al., 2012), lineage progenitor cells and terminally differentiated cells (luminal or myoepithelial cells). The proportion of each type of cell is dependent on the stage of lactation (Hassiotou et al., 2013). During lactation, the glandular part of the human breast undergoes a number of changes.

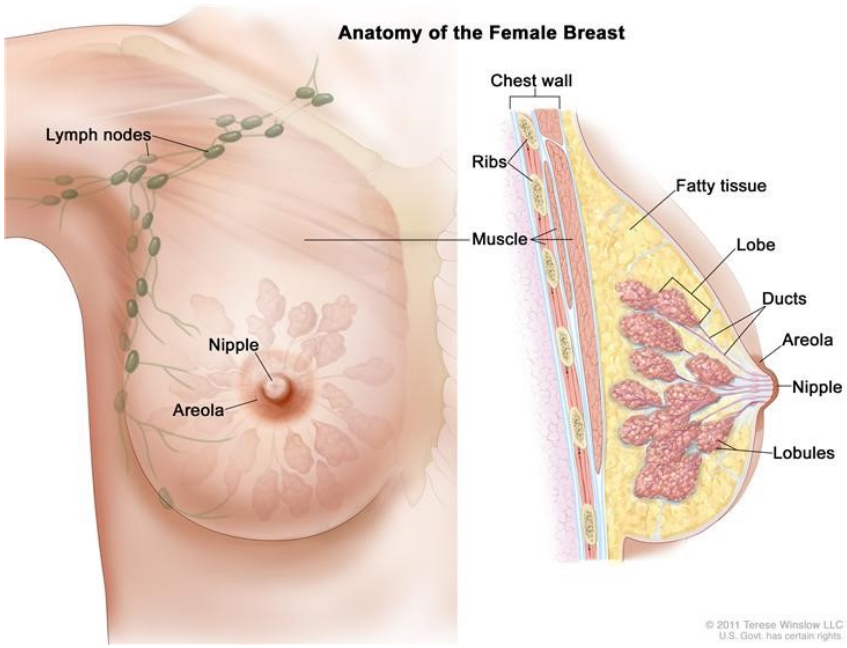


Figure 1.1: Normal female breast anatomy, courtesy of *Breast Cancer Treatment (PDQ)®,* published by the National Cancer Institute of the NIH, USA (NCI, 2014).

1.1.3 Breast cancer

Breast cancer is characterised by a set of traits that comprise the 'hallmarks' of cancer (Hanahan and Weinberg, 2000). Cancerous cells are able to grow independently and resist external growth suppression signals; they can evade apoptosis, or programmed cell death; they have limitless potential for self-replication and can stimulate blood vessels to grow, thereby ensuring a constant supply of nutrients to support their growth; and lastly, cancers invade into local healthy tissues, and spread to more distant sites via the lymphatic or vascular systems.

Although these traits are found in all cases, breast cancer is a clinically and biologically heterogeneous disease (Jacquemier et al., 2005). Indeed, the large body of available data has enabled the development of technology that enables patients to undergo genetic 'typing' of their tumours (Amos et al., 2013), which has led to breast cancer management becoming one of the first examples of truly personalised treatment approaches (Sorlie et al., 2006, Sorlie et al., 2001). Several classification systems exist for the diagnosis of breast cancer, which broadly determine whether the tumour has arisen from breast lobules or epithelial cells that line the ducts (Fig. 1.1 for normal breast anatomy), and therefore whether the cancer is lobular or ductal carcinoma. The majority of cancers in the breast arise from the ductal epithelial cells that line the milk ducts (Visvader and Stingl, 2014, Santagata and Ince, 2014, Ince et al., 2007). Tumours are classified by their grade (level of cell differentiation, whereby more differentiated tumours are referred to as low grade and closer to normal tissue in appearance, or high grade for poorly differentiated tumours, indicating a greater degree of malignancy.

Breast cancer can be divided into multiple pathological subtypes depending on the expression of cell surface receptors by tumour cells. The primary diagnostic step

is to assess whether the tumour expresses oestrogen receptor (i.e., ER-positive vs.ER-negative tumours), progesterone receptor and HER2 (Receptor tyrosine-protein kinase erbB-2 or ERBB2), an oncogenic growth factor receptor. These proteins are crucial for predicting disease course and outcome, with so-called 'triple-negative' tumours (negative for ER, PR and HER2) carrying the worst prognosis and the most challenging to treat (Liedtke and Rody, 2015). Moreover, these receptors are major treatment targets, as will be discussed below.

Until very recently, prognosis was dependent on an analysis of patient age, breast cancer histology (ER/PR/HER2 status), tumour grade and lymph node or distant spread. The recent advent of high throughput sequencing technology and microarray platforms, such as the Oncotype DX commercially available array (Bombard et al., 2015), has enabled the identification of not only risk-related mutations, but also single nucleotide polymorphisms that confer risk in specific populations and individuals (Wacholder et al., 2010). Assessments of copy number variation in breast tumours has enabled further molecular subtypes of breast cancer to be defined (Jönsson et al., 2010), and additional categorisation has been possible using epigenetic analyses (Holm et al., 2010). These studies have shifted our understanding away from the broader histopathological categorisation of the disease to a more individualised diagnosis, with exciting potential for personalised treatment strategies (Schnitt, 2010). The broadest categories include luminal A and B, HER2 and basal-like tumours (Sorlie et al., 2001). Table 1.2 shows the breast cancer subtypes that have been identified by expression profiling.

Table 1.2: Expression profiles of breast cancer molecular subtypes, adapted from (Schnitt, 2010).

Molecular subtype	Expression profile
Luminal A	ER+ +/- PR+, HER2-, low Ki67 (<14%)
Luminal B	
Luminal HER2	ER+ +/- PR+ and HER2+
Pure luminal B	ER+ +/- PR+, HER2- and high Ki67 (>14%)
HER2	ER-, PR- and HER2+
Basal-like	ER-, PR-, HER2-, CK5/6 +/- EGFR+
Claudin-low	ER-, PR-, HER2-, low cell-cell junction proteins (Perou, 2011)

1.1.4 Incidence, mortality and prognosis

Although breast cancer is a disease that occurred in antiquity (Sakorafas and Safioleas, 2009, Ekmektzoglou et al., 2009), breast cancer was still a relatively uncommon disease until the 19th century. This was largely because women did not generally survive long enough to develop the disease, and lifestyle factors that are common today and predispose to breast cancer risk were not as common (Wilson et al., 2004). In the United Kingdom, as with developed countries worldwide, the incidence of breast cancer is still rising. The lifetime incidence of breast cancer has risen to 1 in 8 over the last few years (Figure 1.2A). Worldwide, rates of new diagnoses are also increasing, although mortality rates are generally lower and longer survival durations are recorded for individuals with breast cancer (Ferlay et al., 2013).

Breast cancer incidence rises in parallel with the socioeconomic development of individual countries, but the reasons for this remain unknown. This observation is likely to reflect a number of interacting environmental, societal and cultural changes that interact to increase breast cancer risk. An analysis of the underlying changing risk factors in these countries has provided insights into the aetiology of breast cancer. Such factors include lower rates of breastfeeding (Cancer Research UK

Report, 2002), earlier menarche (Gail et al., 1989, Magnusson et al., 1999), later age of childbearing (Brinton et al., 1983) and lower levels of parity (Velie et al., 2005). The role of ethnicity has not been implicated in variable breast cancer risk, but geographical variability in breast cancer rates is more likely to result from cultural differences that modify individual breast cancer risk factors (Gathani et al., 2014). Individual countries face unique problems that may exacerbate the increasing incidence of breast cancer, such as the issue of the one-child policy in China (Fan et al., 2006). Arab countries face issues of low rates of awareness and self-examination as a taboo subject, meaning that presentation is delayed (El Saghir et al., 2007, El Saghir, 2008). Interestingly, the rates of breast cancer in women from black and south-east Asian ethnic backgrounds are lower than the white Caucasian populations in the UK. An analysis of the Million Women Study cohort indicated that this finding principally reflected the lower rates of known risk factors for breast cancer in these ethnic groups (Gathani et al., 2014). Once adjustments were made for these risk factors, the incidences between the ethnic groups were relatively similar.

Despite a generally lower mortality rate and longer survival durations for individuals with breast cancer, rates of new diagnoses are increasing worldwide. Interestingly, ER-positive tumours have been shown to be rising in incidence over and above those of ER-negative breast cancers (Hou and Huo, 2013), which may reflect a pattern that is underpinned by ethnic diversity or age at onset. A recent Danish study concluded that current trends could be projected to a 13% rise in ER-positive tumours over the next 4 years, with a fall in rates of ER-negative disease by 15%, and an overall rise of 7% in breast cancer incidence overall, which is in

accordance with the trends to rising incidence in developed countries worldwide (Anderson et al., 2013).

Breast cancer is the third most common cause of death from cancer in the UK, and the second leading cause in women (Figure 1.2B). However, as shown in Figures 1C and D, mortality rates from breast cancer have been decreasing over that previous 40 years, largely due to the implementation of screening programmes (Marmot et al., 2013), early detection, and more effective treatment strategies (Lyratzopoulos et al., 2011, Mukhtar et al., 2013). Breast cancer survival rates vary depending on the histological subtype of the cancer. Taken together, the most recent overall survival data derived from independent research by CR-UK and the Office for National Statistics Statistical Bulletin is presented in Table 1.3. One-year relative survival rates have been steadily increasing in the UK since the 1970s from 82% to 96%, which reflects both early diagnosis and treatment successes. Five-year survival rates were only 52% in England and Wales between 1971-1975, but are now over 85% (Statistics, 2011).

Table 1.3: The 1-, 5-, 10- and 20-year survival rates for women aged 15-99 years in England and Wales.

	Relative Survival (%)			
	1 Year	**5 Year**	**10 Year**	**20 Year**
Sex	**2005-2009**	**2005-2009**	**2007**	**2001-2003**
Female	95.8	85.1	77	64.5

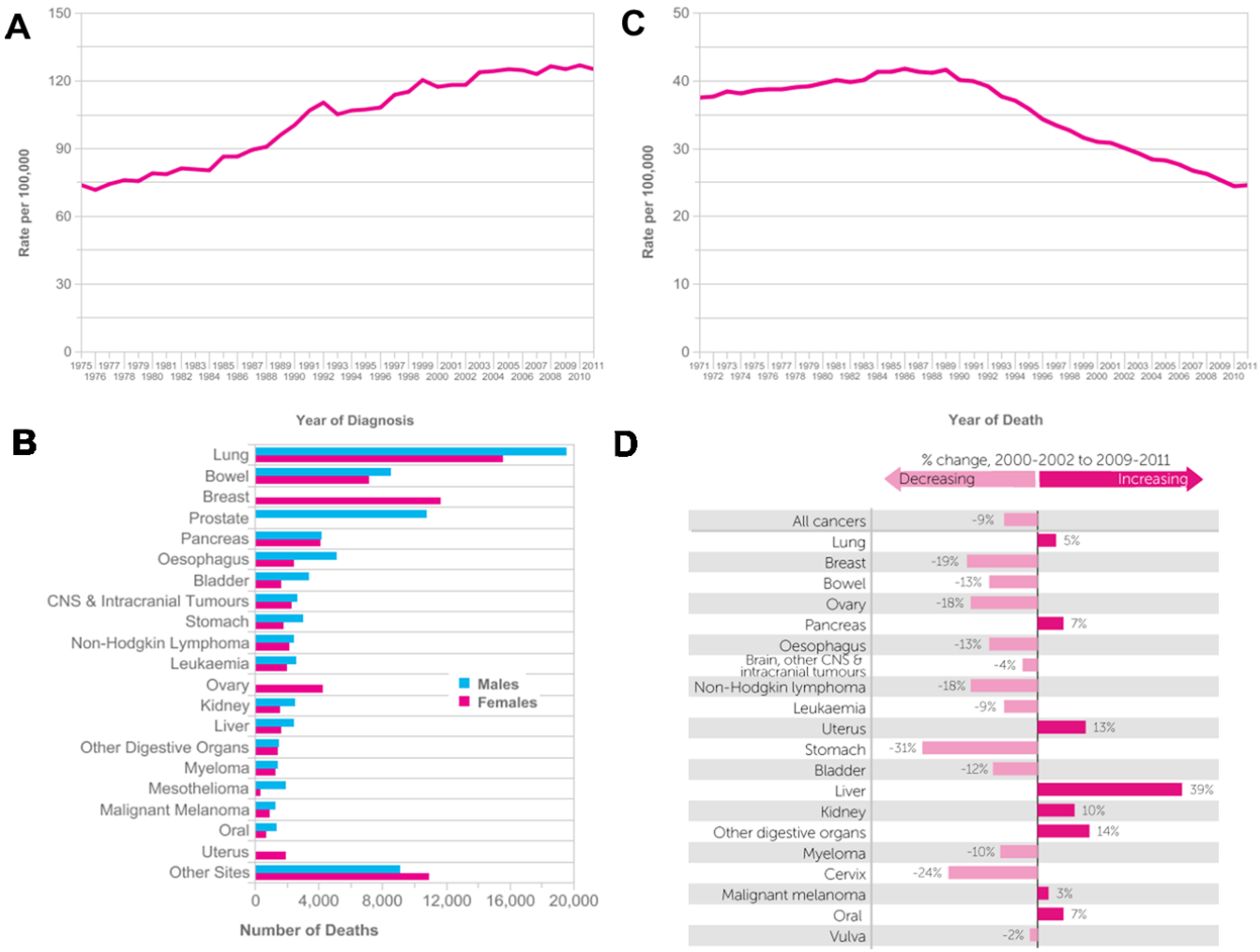


Figure 1.2: Current trends in incidence and mortality rates of all cancers and breast cancer in the UK. A) Age-standardised incidences of breast cancer in women in Great Britain per 100,000 population; B) most common causes of cancer mortality in the UK in both genders; C) rate of breast cancer death in females in the UK over the last 40 year period; and D) trends in cancer mortality (ref: http://info.cancerresearchuk.org/cancerstats/).

There are major differences in survival outcomes for breast cancer in different socioeconomic and ethnic groups in the UK, which reflect underlying differences in presentation times and available treatments (Aizer et al., 2014). African-American women have a higher risk of death from breast cancer than their white Caucasian counterparts, and although there is an overall lower incidence of breast cancer across this ethnic group, cases occur in younger women and a higher rate of the more aggressive triple-negative form of breast cancer occurs (Akinyemiju et al, 2015). Given that the same features of the disease are found in women from West Africa, from where the majority of African-American women originated, it is likely that this biological disease behaviour reflects an underlying genetic background (Newman, 2014), or even a different epigenetic background (Zhu et al., 2003, Mehrotra et al., 2004). However, rates of breastfeeding are significantly lower in African-American women than other ethnic groups, in addition to higher rates of obesity and diet, and the effects on these factors on the increased risk of death needs to be further explored (Brawley, 2013).

1.1.5 Aetiology and risk factors for breast cancer

Despite a vast body of research and a broad understanding of the risk factors for breast cancer, its precise causes are not fully understood; therefore, the ability to predict who will develop the disease is limited (Amir et al., 2010). As a greater understanding of genetic and environmental risk factors has been gained, more accurate predictive tools will be developed. We currently know that genetic mutations account for the pathogenesis of approximately 20-30% of all cases of breast cancer (Ford et al., 1998), and an array of prediction tools factor in the effect of relevant mutations (Antoniou et al., 2008, Jacobi et al., 2009). However, the causes of the

majority of sporadic cases remain unclear. The following description of the risk factors for the disease and the development of the hormone hypothesis of breast cancer aetiology gives an overview of potential aetiological agents and events.

The greatest risk factor for breast cancer is increasing age, but the influence of reproductive hormones and related factors such as age of menarche, parity, breastfeeding history and use of peri- and post-menopausal hormone therapy are all highly significant. In 1896, George Beatson published the report of two cases of oophorectomy in the Lancet, and described the regression of the breast cancer in both cases, but without a clear understanding of the mechanism (Beatson, 1896). This insight led to the investigation of the effect of oestrogen, a steroid hormone that is principally produced by the ovaries in women of reproductive age.

The link between oestrogen levels, reproductive variables and breast cancer risk were first formally summarized by Cole and MacMahon in the Lancet in 1969 (Cole and MacMahon, 1969), but the hard data to support the hypothesis that reproductive factors were associated with breast cancer risk came from a large population cohort study of 280,000 women in the USA in the early 1980s (Brinton et al., 1983). These women were followed-up for 5 years, and individuals who developed breast cancer during that period were matched with regards to age with healthy women in the same cohort. Their results showed that the relative risk (RR) of developing breast cancer increased substantially with increasing age of their first live birth (Figure 1.3). First-trimester abortion and miscarriages were excluded as contributors to risk. Multiparity, particularly women with five or more children, inferred a slighted reduced risk (RR: 0.80), even for women whose first live birth occurred after they were 30 years old.

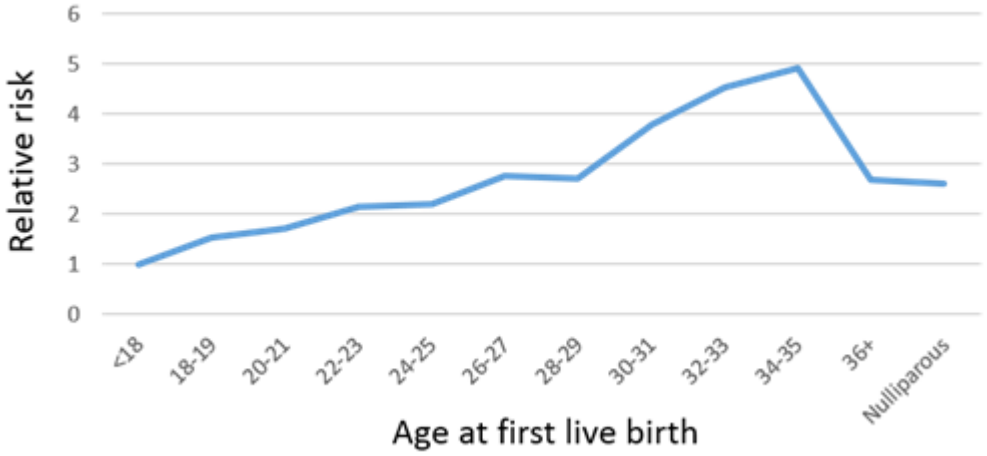


Figure 1.3: Relative risk of developing breast cancer according to age at first live birth (data derived from Brinton et al. (Brinton et al., 1983).

This study by Brinton et al. also gave the first indication that breastfeeding was protective, with the RR decreasing from 1.00 in women who had never breastfed to 0.90 in women who had breastfed all of their children (n = 444 vs. 335; 95% CI: 0.7-1.1). Interestingly, this study also indicated that being breastfed as an infant was protective for the women in later life (RR 1.00 vs. 0.86 not-breastfed [n = 188] vs. ever breastfed [n = 1004]; 95%CI, 0.7-1.1), which has since been confirmed by later large studies and meta-analyses (Cancer Research UK Report, 2002, Freudenheim et al., 1994, Anothaisintawee et al., 2013). This aspect of breast cancer risk will be discussed in-depth later in this thesis.

Later studies replicated the results of Brinton et al., and further showed that the reported association of early-onset menarche (<12 years of age) was associated with increased breast cancer risk (Magnusson et al., 1999). Additionally, various anthropometric factors have been associated with an increased breast cancer risk, including increased body mass index (Gonzalez and Riboli, 2010), adult height (Nilsen and Vatten, 2001), and greater birth weight (Velie et al., 2005). The links

between causation and correlation though are difficult to establish in these highly confounded parameters, and associated studies may also be subject to various degrees of recall bias. Breast density as based on mammographic assessment criteria has recently been included in the NICE guidelines for breast cancer risk assessment (Brand et al., 2014, Lindstrom et al., 2014).

A positive family history of breast cancer is associated with a higher chance of developing the disease, subject to a range of variables such as the age of the affected relative and the genetic closeness of the relationship. The National Institute for Health and Care Excellence (NICE) have published an updated list of guidelines in 2013 that enable clinicians to assess the risk of an individual of developing breast cancer based on their family history. An individual should be referred for genetic counselling and further investigation if they have any of the following criteria (adapted from https://www.nice.org.uk/guidance/cg164/chapter/Key-priorities-for-implementation, accessed 05.03.2015):

- One first degree female relative diagnosed with breast cancer aged younger than 40 (a first degree relative is your parent, brother or sister, or your child)
- One first degree male relative diagnosed with breast cancer at any age
- One first degree relative with cancer in both breasts where the first cancer was diagnosed aged younger than 50
- Two first degree relatives, or one first degree and one second degree relative, diagnosed with breast cancer at any age (second degree relatives are aunts, uncles, nephews, nieces, grandparents, and grandchildren).
- One first degree or second degree relative diagnosed with breast cancer at any age and one first degree or second degree relative diagnosed with ovarian cancer at any age (one of these should be a first degree relative).

- Three first degree or second degree relatives diagnosed with breast cancer at any age.

Individuals also have a higher chance of developing breast cancer if their relatives had bilateral breast cancer, the affected relative was male, they have Jewish ancestry, and multiple other cancers were diagnosed when the individual was young (https://www.nice.org.uk/guidance/cg164/resources, accessed 05.03.2015).

The use of post-menopausal hormone replacement therapy (HRT) has been strongly implicated in promoting breast cancer incidence with a lag time after treatment of approximately 5 years, although this varies depending on the population and methodology used (Antoine et al., 2014). Indeed, the decrease in breast cancer incidence by the Million Women Study in the UK in 2003 (Beral, 2003) was attributed to the discovery of this link and consequent reduction in the prescribing of HRT by physicians (Berry and Ravdin, 2007, Ravdin, 2008, Glass et al., 2007). Although the increase in relative risk caused by the use of HRT is relatively small, post-menopausal breast cancer is common, and therefore the numbers of women affected and the consequent absolute risk is quite high (Vessey, 1997).

Independent from reproductive factors and body habitus, several other environmental factors increase the lifetime risk of breast cancer. Leading these is alcohol, with even light drinking being associated with an increased risk (Seitz et al., 2012, Bagnardi et al., 2015). For example, Nichols et al. showed an odds ratio (OR) for developing breast cancer of 1.85 (95% CI, 1.32-2.61) in Asian women who consumed any alcohol (beer or rice vodka) compared to women who did not drink (Nichols et al., 2005), and the degree of risk is not thought to be modified by the type of alcohol consumed (Newcomb et al., 2009), but by the volume consumed (Breslow et al., 2011).

The evidence for smoking as a risk factor for breast cancer is complex, and highly dependent on the age of smoking initiation (Hamajima et al., 2002, Reynolds, 2013, Welp et al., 1998). A large (>100,000 women) cohort study showed that women who started smoking as teenagers (<15 years old) and smoked for at least 20 years had a relative risk of developing breast cancer of 1.48; interestingly, women that started to smoke at an older age but before their first birth had a RR of 1.27, compared to no increased RR in age-matched women who started to smoke after their first birth (Gram et al., 2005). Of note, women who begin to smoke after the menopause may be protected from breast cancer by the induction of the cytochrome p450 system within the liver, which metabolises toxins and hormones, including oestrogen (Kendall et al., 2007). Reduced circulating levels of oestrogen would theoretically therefore be protective against driving breast carcinogenesis. The effect on women may be more individualised, depending on their genotype in specific genes involved in the metabolism of cigarette toxins, specifically N-acetyltransferase 1 and 2 (NAT1 and NAT2) (Millikan et al., 1998, Morabia et al., 2000, Alberg et al., 2004, Lissowska et al., 2006, Ochs-Balcom et al., 2007). Smoking as a risk factor for breast cancer will be explored in more detail in Chapter 3 of this thesis.

Finally, and more speculatively, a viral basis for breast cancer has been proposed, in a similar manner to the widely accepted theory of human papilloma virus as the primary aetiological agent underlying cervical cancer. Pogo et al. reviewed the literature in 1997 to illustrate the links between mouse mammary tumour virus (MMTV), which causes mammary tumours in mice (Pogo and Holland, 1997). Stewart et al. proposed that breast cancer rates were highest within the distribution of *Mus domesticus*, the house mouse, which carry MMTV (Stewart et al., 2000). However, their study did not adjust for reproductive factors or mouse

population density, and was widely refuted (Khuder and Mutgi, 2000). However, this theory has gained a degree of traction, with MMTV DNA sequences found in 30-40% of human breast cancers compared to 4% of normal tissue (Cotterchio et al., 2002, Lawson et al., 2006b). Lawson et al. demonstrated that MMTV promotes oncogenesis by the insertion of Wnt oncogenes into the human genome of mammary tissue, and that other viruses that include human papillomavirus and Epstein-Barr virus may also play a role in breast oncogenesis (Lawson et al., 2006a, Lawson et al., 2010).

1.1.6 Signs and symptoms

The most common symptom of breast cancer is a lump within the breast, but the majority of all breast lumps that present to medical services are benign. Benign lumps are due to normal cyclical variability in the breast tissue, fibrous glandular tissue (fibroadenomas), and fluid-filled cysts. Signs that the breast lump is more likely to be malignant are overlying dimpling of the skin, termed peau d'orange as it resembles the skin of an orange peel. Malignancy is also more likely if the lump itself distorts the normal shape or the breast, or if the lump is hard. Nipple signs can also be suggestive of malignancy, including nipple distortion, bloody nipple discharge or a peri-areolar rash. Breast cancer commonly metastasizes to the lymph nodes (axillary, supraclavicular, infraclavicular and internal mammary nodes), bones, lung, liver and brain. Symptoms of distant spread include masses within the axilla on the same side, lumps in the contralateral breast (which may be de novo or spread from a unilateral primary across the internal mammary lymph nodes)

1.1.7 Diagnosis and screening

The evidence base for the diagnosis of breast cancer is strongly established and enshrined in a National Institute of Clinical Excellence (NICE) guideline (http://www.nice.org.uk/guidance/cg80). In brief, the current guidelines advocate triple assessment for women with suspicious lumps or breast appearance within 2 weeks of presentation. Triple assessment includes: (i) clinical examination; (ii) mammogram and/or ultrasound (USS) in women aged <35 years or with dense breast tissue ; and (iii) histological assessment by either fine needle aspiration (FNA) or core biopsy. Tumours in women diagnosed with invasive breast cancer or recurrent disease should be assessed for ER and HER2 status within 2 weeks of the diagnosis, in order that appropriate planning can be conducted.

In the UK, the breast cancer screening programme was established in 1988, and offers mammography to all women aged 50-70 years or more every 3 years (http://www.cancerscreening.nhs.uk/breastscreen/, accessed 28.01.2015). However, given that women that attend screening have a 20% reduction in mortality (equivalent to 1300 deaths prevented under the current guidelines (Marmot et al., 2013, Screening, 2012), but one in four cases of breast cancer occur in women aged below 50 years of age, calls are growing to extend the screening programme to women at higher risk according to their family history and genotype, as explained in Section 1.1.8 (Hall and Easton, 2013).

1.1.8 Genetic basis of breast cancer

As mentioned above, approximately 10% of breast cancers arise from germline mutations, whereby a specific mutation leads directly to the pathogenesis of breast

cancer. These cases usually present as familial clusters, where mutations are inherited between generations. By far the most common breast cancer-associated mutations are those that affect the *BRCA1* (breast cancer 1, early onset) and *BRCA2* (breast cancer 2, early onset) genes (Lancaster et al., 1996, Miki et al., 1994). BRCA-associated breast cancer usually presents in younger women and generally has an aggressive triple-negative phenotype (Foulkes and Shuen, 2013, De Summa et al., 2013). Women that carry a mutation in either gene have high lifetime risks of breast (46%-71%) and ovarian cancer (41%-46% and 17%-23% for *BRCA1* and *BRCA2* mutations, respectively) (Nelson et al., 2014); carriers are also more susceptible to numerous other cancer types (Narod and Foulkes, 2004), and these genes are inherited in an autosomal dominant manner, which gives children of carriers a 50% chance of inheritance. Other germline mutations carry a high level of risk, such as those in *PTEN*, *TP53* and *STK11* (Walsh et al., 2006), while others have lower levels of risk, while the inheritance of mutations in *CHEK2* and *ATM* confer a lower lifetime risk (Mavaddat et al., 2010, Shiovitz and Korde, 2015). These genes account for a further 5%-10% of inherited breast cancer. The majority of cases of familial cancer are therefore unaccounted for by single gene mutations, and may arise as a result of a range of combinations of genetic and epigenetic susceptibility factors (Aloraifi et al., 2015), or by variants of unknown significance in established risk-associated genes.

Recent advances in array-based and genomic sequencing technologies has enabled case-control studies to tease out variants with even lower levels of penetrance. These genome-wide association studies (GWAS) have identified a host of susceptibility loci for breast cancer by simultaneously analysing genomes for thousands of single nucleotide polymorphisms (SNPs). SNPs are assessed in place

of the more impractical sequencing of the entire genome, but SNPs in both inter- and intragenic regions are assessed, as the latter may indicate disruption of regulatory regions. GWAS studies that have assessed thousands of matched breast cancer cases and healthy individuals have to date identified 77 susceptibility loci (Michailidou et al., 2015, Michailidou et al., 2013).

Low penetrance genetic factors (i.e., risk-associated SNPs) do not currently form part of the standard assessment of breast cancer risk in women at a high risk of the disease because of a significant family history. Testing in these women is limited to specific genes for which a high index of suspicion exists, alongside a detailed clinical and family history. However, these genetic risk factors have been amalgamated into several array-based assays that enable risk to be quantified for an individual as a polygenic risk score (Shiovitz and Korde, 2015, Mavaddat et al., 2015). Larger GWAS studies are currently being conducted as a consequence of the formation of consortia, such as the Breast Cancer Consortium (BCAC), which enable higher powered studies that can identify rarer risk SNPs.

Interestingly, some of the identified risk-associated SNPs may act via epigenetic mechanisms. For example, GATA3 is a transcription factor that is frequently mutated in breast cancer. Mutations in GATA3 have been reported to increase the stability of its binding to chromatin at oestrogen binding sites, although the functional significance of this effect remains unknown (Adomas et al., 2014). Other SNPs can disrupt CpG sites, leading to the loss of methylation at these loci; again, the functional significance of this, such as whether enhancer or other regulatory regions are affected, remain broadly unknown. Further intensive work is ongoing to determine the functional effect of these SNPs, their effects on transcription and downstream signalling, and their interaction with epigenetic mechanisms.

1.2 Epigenetics

1.2.1 Overview

The definition of epigenetics continues to evolve, but can be most broadly described as the study of all mechanisms that regulate gene transcription and genome stability that are maintained between cell divisions but do not include the code of the DNA sequence. This information can then influence whether a gene is expressed or not, physical changes in the compaction of chromatin, around which the DNA is packaged, and enhance genomic integrity. The mechanisms or 'marks' that enable this transfer of information include DNA methylation, histone modifications, and microRNAs (miRNAs).

The term epigenetics was first used by Conrad H. Waddington in 1942, who combined the words genetics and epigenesis. Waddington's interests ranged across genetics, paleontology and embryology, which led him to lay the basis for the concept of systems biology (Robertson, 1977). Waddington's intention in coining the term epigenetics was to describe an over-arching mechanism by which genes, which had not yet been characterised, could be influenced by their environment to alter the phenotype of the organism (Goldberg et al., 2007).

1.2.2 DNA methylation

DNA methylation is perhaps the most intensively investigated epigenetic modification in the human genome, largely because of the recent advances in technology that have enabled whole-genome analyses to be performed more easily and cost-effectively. In addition, the mark is relatively simple to analyse compared to other epigenetic mechanisms such as histone modifications. Methylated DNA describes

the stable and heritable chemical modification of the 5-carbon of the cytosine carbon ring by the enzymatic addition of a methyl (CH_3) group. Cytosines are most commonly methylated when they are immediately upstream of a guanine base, in a so-called CpG dinucleotide (where the p stands for a phosphorylated group). The methylation is achieved in two ways: methylation is passed down between generations via the action of maintenance methyltransferases that add CH_3 (methyl) groups to hemimethylated DNA during replication (Figure 1.4); alternatively, de novo methyltransferases add new methyl groups to CpG sites after DNA replication is complete (Groth et al., 2007). The latter represents interactions between the genotype and phenotype according to the influence of environmental exposures (Goldberg et al., 2007).

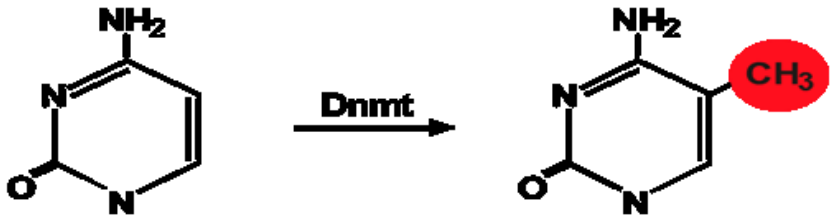


Figure 1.4: The addition of a methyl group to the 5-carbon of a cytosine base, enzymatically catalysed by a member of the DNA methyltransferase (Dnmt) family.

In adults, >70% of CpG sites in the genome are methylated, and the majority cluster in regions of high density CpG islands (CGIs). CpG sites are found at relatively lower rates than would be expected in the human genome, which is likely to be because of the high mutability of these sites (Li et al., 2012).

This year, 2015, marks the 40th anniversary of the first publications that suggested DNA methylation was a heritable epigenetic mark that was implicated in gene regulation developmental processes (Holliday and Pugh, 1975, Riggs, 1975).

Their paper gave the first insight that the four accepted nucleotide bases in the human genome, cytosine, guanine, adenosine and thymine, were not exclusive. Lister et al. (2009) demonstrated using novel sequencing technology (MethylC-Seq) that two cell types from the mammalian genome, human embryonic stem cells and fetal fibroblasts, had profoundly different patterns of DNA methylation (Lister et al., 2009). Not only did this paper confirm the dynamic nature of this modification at enhancer and promoter regions, but also showed that almost 25% of methylation in pluripotential stem cells occurred at non-CpG sites. Therefore, DNA methylation in a developmental context may act via different mechanisms to those in differentiated cells.

Gene transcription is initiated at the promoters of genes, which are DNA regions that encompass the transcription start site (TSS) and, usually, a sequence upstream of the TSS. Promoters can vary in length from 10-1000 base pairs, and are highly diverse in eukaryotes. They have been defined in humans by sequence-specific binding sites for RNA polymerase II, which mediates the transcription of messenger RNA (mRNA), and transcription factor binding sites that are specific to the function of that gene. Promoters can also be defined by the combination of epigenetic marks at that region. CpG sites tend to cluster in higher densities at promoter regions than throughout the rest of the genome, which has led to the definition of the term, CpG island (CGI) (Bird, 1986). Promoter-associated CGIs tend to be unmethylated (Takai and Jones, 2002), while up to 80% of the total number of CpG sites in a genome are methylated. CpGs are found at significantly higher densities in gene-rich areas of human chromosomes compared to gene-poor regions (Weber et al., 2005).

The hypermethylation of CGIs within the promoters of genes, which are present in 60% of human gene promoters (Gardiner-Garden and Frommer, 1987), has been defined as a repressive mark (Doerfler, 2008). Interactions have been shown to occur between methylated DNA and another repressive epigenetic mark, the polycomb group of histone proteins (Viré et al., 2006). Gene expression has been proposed to be repressed by the methylated CGIs altering the conformation of DNA in addition to local histone structures (Cedar and Bergman, 2009, Ng and Bird, 1999). However, this mechanism has only been thought to account for the transcriptional control in approximately half of all genes, and this figure may indeed be as low as a subset of 200 genes in any given cell type (Weber et al., 2005). More recent studies have suggested that transcriptional repression occurs independently from promoter-associated methylation states (Lou et al., 2014). DNA methylation at non-promoter regions may have additional roles in the transcriptional control of genes, by maintaining genomic stability and repressing the transcription of repetitive elements.

1.2.3 Intragenic methylation

The potentially important role of intragenic methylation (IGM) was first suggested in a study into parasites in 1997 (Yoder et al., 1997). The first genome-wide analysis of DNA methylation in an organism was published in 2006 by Zilberman et al. in a study of *Arabadopsis thaliana* (Zilberman et al., 2007). In this study, it was demonstrated that only 5% of gene promoters in this plant had hypermethylated promoter regions, but over a third of genes had highly methylated intragenic regions. Most crucially though from a potential functional point of view, Zilberman et al. demonstrated that there was a correlation between levels of DNA methylation and gene transcription

rates (Figure 1.5). Genes that had levels of expression in the top three deciles had higher levels of IGM than genes that were expressed at low levels, and low levels of methylation were observed in the flanking regions of all genes in *Arabadopsis*. It has since been discovered that these patterns of IGM are conserved between the plant and animal kingdoms (Zemach et al., 2010b, Feng et al., 2010).

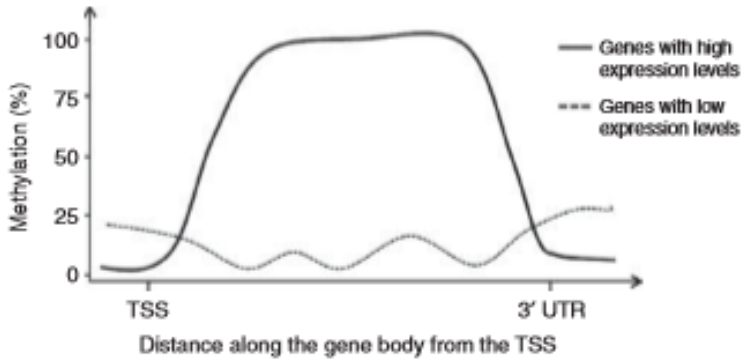


Figure 1.5: Schematic of the results of Zilberman et al. showing that genes with high expression levels have higher levels of intragenic methylation (Zilberman et al., 2007, Shenker and Flanagan, 2012).

Several reports have now established that there is a similar relationship between IGM and gene expression in humans (Hellman and Chess, 2007, Aran et al., 2011, Ball et al., 2009, Rauch et al., 2009). Hellman and Chess used an array-based method to illustrate methylation patterns in the X chromosome relevant to X-linked gene silencing (Hellman and Chess, 2007). Unexpectedly, the active X-allele had higher methylation levels than the silenced X-allele, of which 49% of methylated sites were in intragenic coding regions. Interestingly, there was no bias towards the methylation of repetitive elements within genes, which contrasted with the hypothesis postulated initially by Zilberman et al. that the role of IGM was to suppress the transcription of repetitive elements (Zilberman et al., 2007). Rauch et al. confirmed that methylated intragenic regions correlated with higher levels of gene transcription in B cells (Rauch et al., 2009), and proposed that IGM could in fact regulate gene

transcription from alternative promoters. This hypothesis was supported by a study that examined DNA methylation in a specific genomic region in human brain tissues (Maunakea et al., 2010). IGM may also play a role in the regulation of alternative splicing of mRNA transcripts to produce different isoforms from the same gene (Maunakea et al., 2013).

The rate of cell division differs between human tissue type, which appears to influence the level of IGM in genes that are actively expressed and inactive genes (Aran et al., 2011). Both expressed and non-expressed genes in tissue types with low rates of cell division, such as the lung or brain, tended to have unmethylated promoters and highly methylated intragenic regions; in contrast, in rapidly dividing tissues, such as bone marrow lymphocyte populations and the placenta, IGM levels were found to be reduced in inactive genes. The cause of this differential methylation profile is yet to be established.

1.2.4 DNA methylation and cancer

In cancers, a substantial proportion of genes with promoter-associated CpG islands become hypermethylated, alongside global genome-wide hypomethylation (Jones and Baylin, 2002, Jones and Baylin, 2007). Hypomethylation of promoters can result in the expression of oncogenes, while promoter hypermethylation frequently silences tumour suppressor genes in cancers. The consequences of genomic hypomethylation, whether intergenic or intragenic, are also unclear, but may lead to the reactivation of repressed repetitive elements, and further genomic instability. However, hypomethylation in tumours has been shown to occur in blocks, which are broadly unrelated to repetitive elements, and therefore this is likely to only represent one of the many potential underlying causes (Hansen et al., 2011). The causal

mechanisms for genomic hypomethylation are still unknown, but these are unlikely to be due to the dysregulation in DNA repair mechanisms or members of the DNA methyltransferase (DNMT) enzyme family (Wild and Flanagan, 2010). However, the profiling of specific sites or panels of sites that are aberrantly methylated within tumour cells are under investigation as biomarkers of early prediction and prognostication where cancer-associated methylation can be detected in samples such as tumour biopsies, cell-free serum, urine, and peritoneal fluid (Schulz and Goering, 2011). Furthermore, DNA methylation profiling has been used to define novel tumour subgroups (Flanagan, 2011).

Further understanding of the epigenetic mechanisms that regulate gene transcription in pathological states such as cancer will be crucial in the development of novel and innovative therapies for these diseases. Already, promoter hypermethylation can be reversed by non-specific demethylating agents, such as decitabine (DAC) and 5-azacytidine (5-aza) which has been approved in the therapy of several haematological malignancies (Fandy et al., 2007). Few clinical trials of demethylating agents in solid tumours have been performed. To date, the largest phase 1 study in solid tumours suggested that the doses required to produce the same effects in humans as in mice were tolerable (Appleton et al., 2007), but as yet demethylating therapies for solid tumours such as breast cancer have not been successful in humans. The first published phase IIa trial suggested that azacytidine could help to reverse platinum-based drug resistance in patients with ovarian cancer (Fu et al., 2011). However, a further recent phase II trial in ovarian cancer suggested that the use of decitabine in combination with platinum-based therapies could actually reduce the response rate to therapy in women with ovarian cancer (Glasspool et al., 2014). While the intention of demethylating agents is to re-activate

promoters that are silenced by promoter hypermethylation, the entire genome could become demethylated, including intragenic sequences. Recent investigations have also suggested that 5-aza has broader effects than purely as a demethylating agent, including altering activating histone modifications to repressive marks, and expression changes independent of promoter methylation (Komashko et al., 2010). However, this study only examined promoter regions and further insights may be gained by investigating the intragenic effects of demethylating agents on both histone and DNA methylation. The temporal nature of these events also requires further investigation. Therefore, understanding the role of intragenic methylation in transcription and in carcinogenesis is critical to understand the full consequences of chemically-induced demethylation as a therapy.

1.2.5 DNA methylation in breast cancer

Data on the role and pattern of IGM in specific cancer types is limited. In breast cancer, the distribution of hypomethylated CpG sites in primary tumours and cell lines has been shown to be clustered in gene-poor chromosomal regions, at the 5' region of frequently expressed genes (including the promoter, first exon and first intron), and large intragenic regions of hypomethylation at chromosomal breakpoints, large genes and tissue-specific gene clusters (Shann et al., 2008). One of the caveats to this field of research though is the lack of comparison with normal breast epithelial cell types, as a consequence of the difficulty of obtaining such cells and their lack of growth potential in culture (Heijmans and Mill, 2012). The significance of methylation changes in cancer cells is therefore yet to be fully established. Significant variability has previously been shown in peripheral blood DNA methylation in intragenic sequences in patients with bilateral breast cancer

compared to matched healthy controls using a tiled microarray (Flanagan et al., 2010). Most variability was found in intragenic repetitive elements, and one repetitive element within the *ATM* gene was associated with a 3-fold increased risk of breast cancer in two large case-control studies from our group (Flanagan et al., 2009, Brennan et al., 2012). Although we do not fully understand the role of IGM changes in the aetiology of breast cancers of different types, these data highlight that potential DNA methylation biomarkers can be determined within intragenic sequences. However, as yet, a quantifiable signature of epigenetic risk with regards to breast cancer has not been ascertained. If one could be developed and validated, it would enhance the already available polygenic risk score and lead to a truly personalised risk prediction tool.

To summarise, aberrations in DNA methylation in cancer tissues have been well described, and work is underway to understand how this happens, at what point prior to or during carcinogenesis epigenetic mechanisms are disrupted, and what are the pathological implications of disrupted DNA methylation. However, most of the studies to date have been conducted on the genome-wide level. The implications of gene body DNA methylation changes at the individual gene level have not been studied, and no study has to date attempted to understand the role of gene body methylation in a single gene.

1.3 Doctoral hypotheses and aims

1.3.1 Hypotheses

- Intragenic methylation is functionally important for gene transcription on an individual gene level.
- Specifically, intragenic methylation regulates *ESR1* gene expression.
- Epigenetic variability is driven by environmental exposures throughout life that leads to altered DNA methylation that may associate with increased cancer risk.
- DNA methylation variability may represent a predictive biomarker of breast cancer risk, potentially as a surrogate of cancer risk exposures such as smoking.

1.3.2 Aims

- To assess intragenic methylation levels of the *ESR1* gene in the context of gene transcription in breast cancer cell lines, tumour biopsies and healthy breast epithelial cells using an array of epigenetic techniques.
- To perform an epigenome-wide association study to investigate methylation levels associated with smoking status as an example of a cancer-risk exposure.
- To develop novel resources for the further investigation of epigenetic changes that can identify risk prospectively. Specifically, with specific reference to the prediction of individuals who are at a higher risk of developing breast cancer, by developing a cohort study collection of breast epithelial cells from donated breast milk samples.

CHAPTER 2

THE TRANSCRIPTIONAL AND TRANSLATIONAL IMPLICATIONS OF INTRAGENIC DNA METHYLATION IN BREAST CANCER

2.1 INTRODUCTION

2.1.1 Intragenic methylation

As described in Chapter 1, strong evidence exists to support the role of aberrant epigenetic mechanisms in tumorigenesis, of which the most well studied type is altered patterns of DNA methylation (Jones and Baylin, 2007, Baylin et al., 2001, Baylin et al., 1998, Doerfler, 2008). DNA methylation is likely to be evolutionarily the oldest mechanism of epigenetic transcriptional regulation (Zemach et al., 2010b, Zemach and Zilberman, 2010), whereby a methyl group is covalently added to the 5-carbon of cytosine bases in a cytosine-guanine dinucleotide (CpG site). CpG sites tend to cluster into non-random CpG islands (CGIs) around the transcription start sites (TSS) of approximately 60% of genes (Bird, 1986). Methylation of the promoter region-associated CGIs leads to conformational changes in the DNA strand and regional chromatin, enabling the DNA to bend by altering the electrostatic composition of the DNA strand (Nathan and Crothers, 2002, Cedar and Bergman, 2009). This bending potentially inhibits the initiation of the transcriptional machinery and prevents the recruitment of RNA polymerase II (Appanah et al., 2007). Therefore, current dogma states that if the CGI is unmethylated, the gene should be actively transcribed (Goll and Bestor, 2005).

During neoplastic change, focal hypermethylation of CGIs affects the promoter regions of multiple genes (Jones and Baylin, 2002, Baylin et al., 2001, Feinberg and Tycko, 2004). Furthermore, concurrent genome-wide hypomethylation occurs in tumour cells in a variety of cancers, including breast cancer (Kanai, 2010, Veeck and Esteller, 2010, Hansen et al., 2011). The pathological mechanisms that underlie genome-wide demethylation in cancer, as well as their full implications, are not

understood, but this event may result in the reactivation of repetitive elements that are usually hypermethylated with consequent genomic instability (Wild and Flanagan, 2010). These changes may follow a distinct order during carcinogenesis and provide biomarkers of disease risk in healthy women (Mulero-Navarro and Esteller, 2008, Michels, 2010). Such biomarkers may be either specific loci or broad signatures of methylation change (Wang and Srivastava, 2010, Martens et al., 2009, Tost, 2010).

The transcriptional regulation of genes during the development of breast neoplasia and throughout the disease course also remains poorly understood. Oestrogen receptor alpha (*ESR1*), whose function will be discussed below, has important diagnostic and prognostic implications in breast cancer (Cleator et al., 2009, Sorlie et al., 2006, Carey, 2010), whereby the primary pathological aim is to determine the intracellular expression of the ESR1 protein (Weigel and deConinck, 1993). Tumours determined to be ESR1-positive (ER-pos) have a markedly different prognosis than ESR1-negative (ER-neg) tumours due to the availability of anti-hormone treatments. Causative mutations, single nucleotide polymorphisms, and loss of heterozygosity have only very rarely been implicated in the control of ESR1 transcription (Keen and Davidson, 2003). Moreover, the importance of epigenetic processes in ESR1 expression and function is becoming increasingly recognised.

2.1.2 Oestrogen receptor alpha (ESR1) and its role in breast cancer

The role of sex hormones in the pathogenesis of breast cancer has been postulated for just over 100 years. During the 1800s, the surgeon to Queen Victoria, Astley Cooper, observed that the size of tumours in premenopausal breast cancer patients fluctuated depending on the timing of the menstrual cycle. Breast cancer was first

identified to be a largely endocrine-dependent disease after the chance discovery in 1880 by Thomas Nunn that menopause bilateral oophorectomy led to the regression of a breast tumour (Love and Philips, 2002). In 1896, George Beatson explored whether observations in cattle and sheep, where oopherectomy altered udder size and lactation, were relevant in human breast cancer; he showed that oopherectomy in premenopausal women reduced breast tumour size (Beatson, 1896, Leake, 1996). This became rapidly adopted throughout the UK as a treatment modality (Boyd, 1897), but a review of a series of patients 2 years later demonstrated that only two thirds of women showed any improvement after oopherectomy (Boyd, 1897, Boyd, 1899). Further hormone-based treatments were developed, including ovarian irradiation, adrenalectomy and hypophysectomy, but the same response rate of two thirds was reported by all studies (Jensen and Jordan, 2003). It was only with the development of modern molecular pathology tools that staining for ESR1 became possible that it became clear that only ≈70% of breast cancers expressed high levels of oestrogen receptor, and would therefore be amenable to hormone-targeted treatment (Carey, 2010).

The discovery of the first oestrogen receptor in 1962 demonstrated the first piece of the complex jigsaw of oestrogenic functions (Jensen, 1962, Heldring et al., 2007). ESR1 is a member of the nuclear receptor superfamily, which comprises 48 members and are activated by a range of ligands that include steroid hormones, thyroid hormones and vitamins (Evans and Mangelsdorf, 2014). The human *ESR1* gene was first cloned and sequenced in 1986 (Green et al., 1986), and shown to be inducible once bound to oestrodiol with high levels of affinity. More recent studies have identified how oestrogens bind to ESR1 (Toft and Gorski, 1966, Jensen et al.,

1968, Brzozowski et al., 1997). ESR1 is composed of regions with distinct functional roles, as illustrated in Figure 2.1.

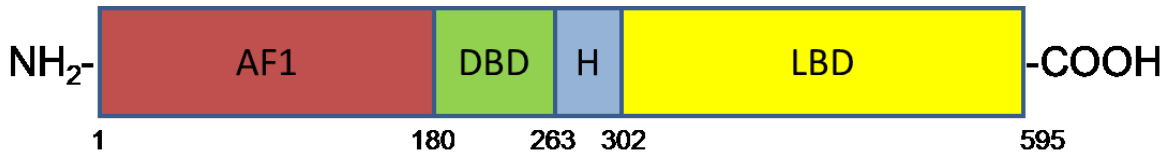


Figure 2.1: A schematic structure of the ESR1 protein. Abbreviations: AF1: activation function-1 domain; DBD: DNA-binding domain; H: hinge; LBD: ligand binding domain. NH_2 and COOH indicate the 5' and 3' end of the protein, correlated to the coding mRNA. Adapted from Robinson et al., 2013.

Activated ESR1 receptors migrate into the nucleus, dimerise and exert their downstream effects by binding to oestrogen receptor elements (ERE) at multiple sites on the genome (Carroll et al., 2006). This binding occurs directly or via transcription factors, such as AP-1 and Sp1. Activated ESR1 proteins also interact with transcriptional coactivators that modify histone proteins covalently. In humans, two types of ER are expressed by different genes: ESR1 (aliases: ERα, ER, ESRA, and NR3A1) and ESR2 (aliases: ERβ, ESR-β). ESR1 is a 66-kDa soluble ligand-activated nuclear transcription factor, which is located on chromosome 6 (152,128,686-152,424,406 base pair, [bp]; UCSC genome browser Feb. 2009 GRCh37/hg19 assembly). ESR2 is located on chromosome 14q23.2 (64,699,747-64,761,128 bp). Both ERs have multiple splice variants, with unclear functional significances. Shortened splice variants may be non-functional themselves, but could bind to the full-length ESR1 or ESR2 protein to modulate their nuclear activities (Li et al., 2003); these may also participate in non-genomic cell membrane-based oestrogen signalling (Wang et al., 2005).

ESR1 is widely expressed in multiple tissues, including the brain, liver and colon, amongst others, but it is principally active in hormone-responsive female reproductive tissues. ESR1 shares a common high-affinity ligand, oestrogen (17β-oestrodiol), and metabolites of oestrogen that are weaker affinity ligands (oestrone and oestriol), with ESR2 (*ER*-β), but these proteins have different expression patterns and broadly independent functions that are mediated by different amino and hydroxyl termini (Klinge, 2001).

ESR1 mediates cell survival by regulating the progression of the cell into the cell cycle and apoptosis and governs organ-specific developmental processes, such as mammary gland development (Manavathi et al., 2014). ESR1 is the principal receptor responsible for the elongation of mammary ducts during puberty, whereas ESR2 governs the differentiation of lactating lobules during pregnancy. ESR1 also mediates the action of oestrogen that fuels the growth of ductal epithelium in both development and ESR1-positive (ER-pos) malignant tumours (Perillo et al., 2000, Doisneau-Sixou et al., 2003). The significance of ESR2 in breast tumours is still poorly understood (Palmieri et al., 2002), and the focus in most breast cancer research has therefore been on ESR1.

2.1.3 The epigenetics and function of ESR1: what was known at the outset of this thesis

ESR1 contains a single CGI that is associated with the TSS at the 5' end of the gene. Much of the data regarding the importance of promoter region methylation in ESR1 for transcriptional status has been conducted in cell line studies. In human breast cancers, between 60-70% of these cancers are classed histologically as ER-

pos and the other 30-40% are ER-neg (Carey, 2010). ER-pos cells have unmethylated *ESR1* promoter regions, and the promoter is also unmethylated in normal breast tissue (Lapidus et al., 1996). Several studies have shown the methylation of the promoter region is associated with a downregulation of expression (Ottaviano et al., 1994, Lapidus et al., 1996, Yoshida et al., 2000). Furthermore, it has been demonstrated that the *ESR1* promoter becomes highly methylated in ER-pos cell lines that lose ESR1 expression (Woodfield et al., 2009). Other studies have defined specific regions of the *ESR1* promoter whose methylation profile may be more important than other promoter regions for gene transcription (Sogon et al., 2007).

Epigenetic changes in *ESR1* DNA methylation patterns are observed frequently in neoplastic cells (Baylin et al., 1998), with often conflicting results depending on the method of analysis and the tissue source used (Appendix 1). In terms of the effects of methylation levels of ESR1 function, Perillo et al. showed that activated ESR1 binds to the ERE within 30 min, and that the assembly of the ESR1 transcription complex takes between 30 – 45 min (Perillo et al., 2008). This involves a fluctuation in the cycling of specific lysine methylation points on histone tails. There is increase in the dimethylation of lysine 4 (K4) in histone 3 (H3K4me2) during this time period, while the repressive histone mark, H3K9me2, is demethylated. After 45 min, the promoter and enhancer of downstream gene targets of ESR1, such as bcl-2 and pS2, are bridged, forming transcriptionally active DNA loops that can contact RNA polymerase II. After 60 min, LSD1 activation falls, and the activity of a K9 specific HMT (Suv39H1) increases. This results in increased H3K9me2 that represses further ESR1 activity. Also at that time point, ESR1-activated LSD1 acts to demethylate H3K4me2 local to the ERE.

H3K4 and H3K9 are generally dimethylated in MDA-MB-231 ESR1-neg cells, fibroblasts (ESR1-neg) and the chromatin surrounding, but not at the point of, EREs in ESR1-pos MCF7 cells (Shi et al., 2011). If ESR1 expression is induced in previously ESR1-neg MDA-MB-231 cells, H3K4me2 methylation is lost due to the activation of LSD1 by ESR1. However, there is no demethylation of H3K9me2, which would result in the transcription of ESR1 targets, suggesting an extra layer of feedback aside from the presence of activated ESR1 for this mechanism (Di Santi et al., 2014).

ESR1 functional activity is modulated in the nucleus by binding to transcriptional coregulators, which have HAT-modulating activity, leading to local histone acetylation (Leader et al., 2006b, Leader et al., 2006a). Furthermore, ESR1 associates with other coregulators, including BRCA1 and MTA1, to perform a varied range of functions (Manavathi et al., 2014).

2.1.4 Aim of this study

A previous study from our group indicated that regions of DNA methylation variability (MVRs) exist across the oestrogen receptor alpha (*ESR1*) gene in peripheral blood cells from breast cancer patients compared to healthy matched controls (Flanagan et al., 2009). However, the functional implications, if any, of this variability remain unknown. The aim of this study was to assess whether intragenic methylation levels within *ESR1* correlated with expression in breast cancer cell lines and tumour tissues, and attempt to assess whether there was a causative relationship.

The expression of ESR1 in breast cancer is significant in terms of disease biology and prognosis, as well as largely determining management strategies.

ESR1-positive tumours are generally less aggressive that ER-negative tumours, with fewer metastases and a slower disease course. There are also a broad range of targeted treatment options, such as tamoxifen, which is a direct competitive antagonist of ESR1, and the aromatase inhibitors, which inhibit the conversion of androgens into oestrogen. However, at the time that this doctorate started, very little information was available about the role of gene body methylation in ESR1 expression, and whether methylation variable loci in intragenic regions could serve as potential and diagnostic biomarkers of ESR1 expression in the absence of promoter methylation.

Therefore, based on the hypothesis that intragenic methylation may play an important role in transcription, we aimed to ascertain whether intragenic methylation patterns differed in human breast cancer cells lines that were positive (n = 3) or negative (n = 3) for ESR1 expression. Each cell line was then treated with the demethylating agent, decitabine (DAC) and the DNA methylation pattern and expression of ESR1 was assessed using pyrosequencing and by quantitative real-time PCR (RT-qPCR) compared to untreated control cells. Finally, methylation levels across the *ESR1* gene were assessed in 155 samples of human breast cancer.

2.2 METHODS

2.2.1 Cell lines

Six cell lines were obtained from stocks at the Hammersmith Hospital, courtesy of Prof. Eric Lam, or purchased from ATCC (ATCC Ltd., VA, USA). Of these, three were confirmed as ESR1-positive (MCF7, T47D and BT474) and three were ESR1-negative (MDA-MB-231, SKBR3 and BT549). Cell line authentication was also formally confirmed by Genetica DNA Laboratories Inc. (Cincinnati, OH, USA), according to an analysis of their short tandem repeat (STR) profiles against a reference set of cell lines.

For all experiments, cells were cultured in sterile conditions at 37°C in a humidified atmosphere with 5% CO_2, and maintained in either Dulbecco's modified Eagle's medium (DMEM; Sigma-Aldrich, Poole, UK) or RPMI 1600 (Sigma) supplemented with 10% fetal calf serum (FCS; Sigma-Aldrich) and 5 ml L-glutamine. Cells were passaged when the confluence was >70%, typically on every third or fourth day.

2.2.2 Decitabine (DAC) treatment

The effect of increasing concentrations of DAC on the six cell lines was assessed using the MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) dye reduction assay. Decitabine (DAC; Sigma-Aldrich) was resuspended in 2.2 ml 100% dimethyl sulphoxide (DMSO; Sigma-Aldrich), and made up to 0.5, 1, 5, 10, or 20 μM compared to growth medium (0 mM) alone as the negative control. Assays were performed in triplicate, and the MTT assay was performed using 20 μl CellTiter 96 Aqueous One Solution Cell Proliferation Assay (Promega, Madison, WI, USA)

according to the manufacturer's protocol. Decitabine was selected for use in this study over azacytidine because it does not incorporate into transcribed RNA, and therefore has fewer off target effects.

Fresh aliquots of DAC and DMSO were used for each experiment. Each cell line was cultured in 75 cm^3 flasks in 10 ml DMEM + 10% FCS with 1 µM DAC or DMSO for 3 d or 7 d in triplicate, and at three separate time points. The work done over a 3 d period was only conducted in MCF7 and MDA-MB-231 cells to assess the influence of time on the demethylating effects of DAC. After the appropriate duration of incubation, cells were trypsinised and counted. Cell pellets were collected after three washes and centrifugation at 1,500 rpm for 5 min, and divided in half for DNA and RNA extraction. DNA was extracted using the QIAamp® DNA Mini Kit (Qiagen, Crawley, UK), and concentration and quality was assessed using a Nanodrop1000 spectrophotometer (ThermoScientific, UK). DNA was stored at -20°C until bisulphite conversion.

Bisulphite conversion changes all unmethylated cytosine bases into uracil, therefore allowing the identification of unconverted cytosines as those that are methylated by pyrosequencing (Tost, 2010). DNA samples were bisulphite-converted using the EpiTect kit according to the manufacturer's protocol (Qiagen). Bisulphite-treated DNA was then desulphonated, washed and eluted prior to its use in PCR assays.

2.2.3 Methylation analyses

Primers for PCR assays were prepared using a semi-nested approach to avoid the amplification of repetitive elements, such as long-interspersed nuclear elements

(LINE) segments, which were often present in the MVRs across *ESR1* (Flanagan et al., 2009) (Figure 2.1). A biotinylated tag was placed on one of the primers, and a common biotinylated primer was used for all reactions as described in previous reports (Royo et al., 2007, Flanagan et al., 2009). The list of PCR and sequencing primer sequences is given in Appendix 2.

Reactions took place in a thermal cycler under the following conditions: incubation at 95°C for 10 min; an initial 20 s incubation at 95°C followed by 10 cycles of a 20 s incubation at 60°C (temperature decreased by 1.0°C every cycle) and incubation at 72°C for 20 s; second round PCR steps were performed using nested primers as follows: 30 cycles at 95°C for 20 s, 50°C for 20 s and 72°C for 20 s followed by a final incubation of 72°C for 5 min, with the exception of MVR 7b which only required a single-step PCR amplification. Products were assessed for quality by agarose gel electrophoresis and stored at 4°C until pyrosequencing.

Bisulphite-converted DNA samples were pyrosequenced using specific sequencing primers designed with the use of the PyroQ assay design software (Pyromark MD, Qiagen), and assay were performed on a Pyromark MD pyrosequencer using standard protocols and controls. Results were discarded and repeated if any the control assays showed positive results.

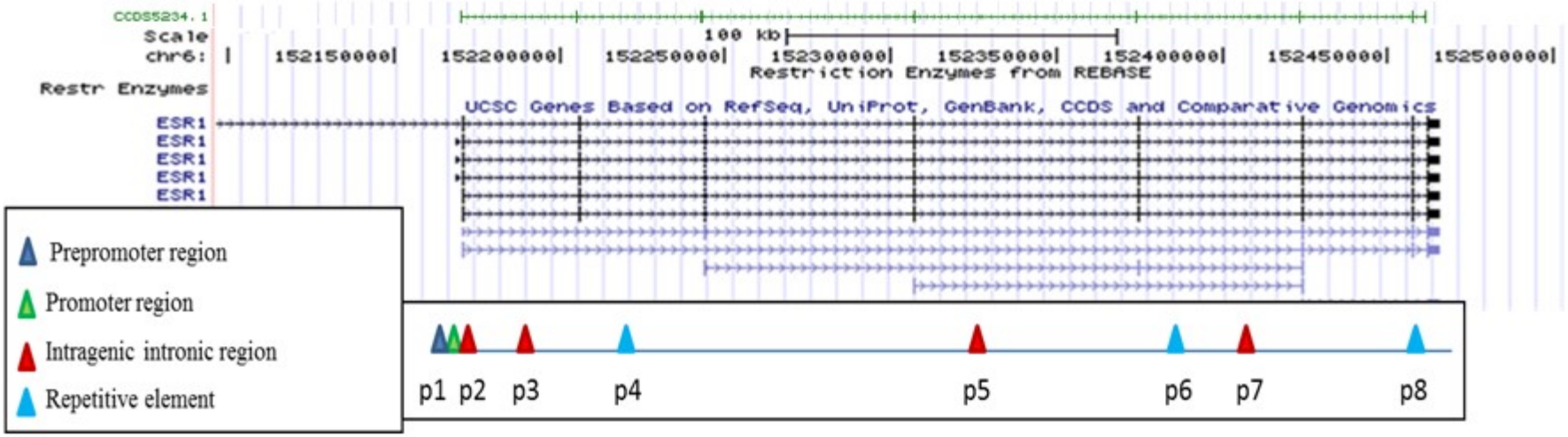


Figure 2.2: A schematic showing the position of pyrosequencing assays across the *ESR1* gene. The genomic regions and multiple transcripts are taken from a screen shot of the Human Feb. 2009 (GRCh37/hg19) genome assembly on the USCS browser (http://genome.ucsc.edu/), with the consensus sequence for the *ESR1* gene shown in green at the top. The additional promoters of *ESR1* are not shown for the purposes of simplicity, as they were not relevant to the principle aims of this study. The x-axis shows the approximate location across *ESR1* of the nine loci, including the promoter region, pyrosequenced in this study, colour-coded with regards to the nature of the region they are within. Between two and four CpG sites were assayed at each locus. Figure reproduced from BMC Cancer (Shenker et al., 2015).

2.2.4 RNA isolation, cDNA synthesis and RT-qPCR

RNA isolation was carried out from the cell pellets with the use of the Qiagen RNeasy® Mini kit (Qiagen), according to the manufacturer's instructions. The concentration of each RNA sample was assessed with the Nanodrop. cDNA was synthesised from 2 µg of each RNA sample using the SuperScript™ III First Strand Synthesis System for RT-PCR (Invitrogen, Carlsbad, CA, USA). Controls were prepared that did not contain SuperScript™ III RT for each group of samples. All samples were stored at -20°C prior to RT-PCR.

Each RT-qPCR analysis was performed in triplicate for each of the duplicate experimental sets of cDNA from the six cell lines. Each RT-qPCR assay was performed in duplicate using primers that were specific for *ESR1* mRNA and for the housekeeping gene, *GADPH* (forward, 5′-TCCCATCACCATCTTCCA-3′ and reverse, 5′-CATCACGCCACAGTTTCC-3′) (Madureira et al., 2006). The details of primers used are given in Supplemental Table 2. All primers were 100% specific for the region of interest. The plate was centrifuged briefly and placed in a C1000™ Thermal Cycler (BioRad, UK). The PCR conditions were established using the BioRad CFX Manager software as follows: 95°C for 3 min denaturing step; 42 cycles of 10 s at 95°C, 10 s at 56°C and 30 s at 72°C; 10 s at 95°C and a melt curve cycle of 5 min that ranged from 72°C to 95°C. Cycle threshold (Ct) values were recorded at a logarithmic threshold of 10^3, and the relative quantitative expression of *ESR1* mRNA in each sample was calculated by the $^{-\Delta\Delta}$Ct conversion.

Finally, RT-qPCR primer sets that were 100% unique for the MVR were designed to analyse the relative expression of antisense mRNA at the regions of interest across the ESR1 gene (Appendix 3). Antisense RT-qPCR assays were

performed in cDNA produced from RNA samples from all six cell lines as described above.

2.2.5 Breast tumour samples

A power calculation was performed to determine the number of patient samples that would be needed to show a significant difference between points of intragenic methylation across the *ESR1* gene previously assayed in cell lines. Group sample sizes of n=45 would be required to reach 98% power at alpha = 0.01 to detect >80% difference (detected at p6), and an 82% power at alpha = 0.05 to detect a significant difference of >40% methylation (observed at p7). We requested samples from the Breast Cancer Campaign Tissue Bank and received 10 formalin-fixed paraformaldehyde slides for each tumour (45 ER-negative tumours samples, 45 ER-positive grade 2 tumours and 45 ER-positive grade 3 tumours, as defined from histopathological analysis by MEB). Furthermore, we received 20 samples of fresh frozen (FF) tumours matched to formalin-fixed paraffin embedded (FFPE) samples for quality control purposes. This study was approved by the Ethics Committee of the Breast Cancer Campaign Tissue Bank (Approval no. BCC-TB00001).

Each set of unstained tissue sections came with a match haematoxylin and eosin (H+E) stained slide. Prior to cell scraping, the tumour portion of each section on these slides were marked by Dr Mona Al-Bahrawy, Consultant Histopathologist at Imperial College Healthcare NHS Trust, London, and used as a template for the tissue that was scraped off for DNA extraction. Slides (n = 4-6 depending on the size of the specimen) were then dewaxed for 10 min in Histoclear, followed by 10 min in 100% ethanol and another 10 min in fresh 100% ethanol. Slides were prepared with Levi buffer using standard techniques, and DNA was extracted using the

phenol:chloroform technique. DNA concentrations and quality were assessed by the Nanodrop. Bisulphite conversions and pyrosequencing analyses were performed as described above. For the DNA extracted from FFPE slides, different primers had to be described with amplicons of <120 bp owing to the relative fragmentation of the DNA after formalin treatment, as DNA quality was poorer in these 135 samples.

Slides were stained immunohistochemically with antibodies against Ki67 according to standard protocols to assess the rate of cell proliferation within tumour sections. Briefly, 2-µm-thick sections from formalin-fixed, paraffin embedded tissue blocks were prepared, deparaffinised and rehydrated. Immunohistochemical staining and detection was performed by Dr El-Bahrawy using an automated Leica Bond 3 machine according to the manufacturer's protocol. Antibodies raised against Ki67 (Leica, Cat No: NCL-L-Ki67-MM1, 1:100) and ER (Leica, Cat No: NCL-ER-6F11, 1:500) were used. Stain detection was performed using a bond polymer refine detection kit. Tonsil sections were used as a positive control for Ki67 staining and breast tissue was used for ER staining. Negative controls were processed in the same manner but with the substitution of PBS for the primary antibody. All sections were examined by light microscopy to assess the presence and scoring of expression. The percentage of tumour cells with nuclear expression of Ki67 was estimated. The Allred scoring system was used to assess ER staining.

Methylation data in BCC Tissue Bank tumour biopsies was validated using the TCGA breast tumour for which 450K Illumina Infinium BeadChip Array data was publically available (n = 365 ER-positive tumours, n = 109 ER-negative tumours). Data was extracted using R software and logistic regression analysis was performed to assess the relationship between ER status and methylation beta-values at each

ESR1 CpG locus, with histology as an independent variable. The Wilcox ranked test was performed with false discovery rate correction as the data was non-parametric.

2.2.6 Cell cycle analysis

This part of the study was conceived by myself and Dr J Flanagan, and performed by Ed Gore as part of his MSc studies. To analyse the influence of cell cycle status upon intragenic methylation levels, MCF7 cells were seeded into 100 mm dishes and underwent a serum starvation protocol using DMEM supplemented with L-glutamine and different concentrations of FBS (5%, 2%, 1% 0%, and 10%). Serum at 10% was used as the control. Three temporally separate experiments were performed, and each concentration was set up in triplicate. Cells were harvested by trypsinisation 24 h after the addition of the experimental medium and neutralized in growth media. Extracted DNA samples underwent PCR amplification and pyrosequencing for the p2, p3 and p4 MVRs, as these were in the most variable regions of methylation. RT-PCR expression analyses for ESR1 normalized against GAPDH were also performed RNA extracted from each sample.

For cell cycle phase analysis, the cells grown in DMEM supplemented with 10% FBS were resuspended after centrifugation with 500 µl propidium iodide (PI) solution supplemented with 50 µg/µl PI (Fluka), 10 µg/µl RNAse (AppliChem), 0.02% Triton-X (Sigma) and PBS. Samples were then incubated at 37°C for 20 min and centrifuged for 2500 rpm for 3 min. The supernatants were discarded and pellets were resuspended with 2 ml cell sorting buffer containing PBS, 1 mM EDTA (Sigma), 25 mM HEPES (Sigma; pH 7) and 1% heat-inactivated FBS. Samples were analysed by the Aria III flow cytometer (BD Biosciences), gated to the G1/G0, early S, late S and

G2/M phases. Cells were sorted in FACs tubes containing 500 µl FBS. DNA was then extracted, bisulphite-converted and pyrosequenced as described above.

2.2.7 Antisense siRNA transfection studies

This part of the study was designed by myself and performed by Emma Bell as part of her MSc studies. To assess the role of antisense RNA in transcription, MCF7 cells cultured as described above were transfected with ESR1 antisense siRNA (CAUGCUAUAAGACUUCCCA; Thermo Scientific, MA, USA) according to the DharmaFECT reagent transfection protocol (Thermo Scientific). There were three sets of controls in each transfection: ON-TARGET plus SMARTpool ESR1 siRNA to target the 3'UTR (ThermoScientific), ON-TARGETplus Non-Targeting Control #1 siRNA (ThermoScientific) and non-transfected cells. Each transfection was performed in triplicate for both DNA and RNA extraction. The siRNA transfection experiment was repeated three times.

2.2.8 Statistical analysis

All experiments were performed in triplicate unless otherwise stated. The mean ± standard deviation (SD) was calculated from each triplicate repeat of the pyrosequencing and RT-qPCR experiments. The mean ± SD were calculated after each replicate, and the standard error of the mean (SE_x) was then calculated. Parametric data, such as the methylation levels in cells incubated with DMSO and DAC or DMSO alone, were compared using paired t-tests. Non-parametric data, including average methylation levels across the gene body, were compared using unpaired Wilcoxon signed rank tests. All statistical tests were two-sided and

performed using Microsoft Excel (Microsoft, USA) and R software, v2.15 (http://www.R-project.org/).

To validate the expression changes of ESR1 after DAC treatment, two publically available expression datasets were mined for data regarding DAC treatment of two breast cancer cell lines used in this current study (gse10613 and gse13733) (Chan et al., 2008, Miranda et al., 2009). The software programme, R v2.15, was used to analyse the data from both studies.

2.3 RESULTS

2.3.1 Cell growth

The MTT assay indicated that there would be no significant cell death after incubation with 1 μM and 5 μM DAC with DMSO or DMSO alone (Figure 2.2). This finding was supported by the results of the cell counts performed before and after the end of each incubation (Table 2.1). No significant differences were found between the pre- and post-incubation counts at either 3 or 7 d in any of the treatment or control conditions ($p > 0.05$), with the exception of MDA-MB-231 cells that were incubated with 5 μM DAC for 7 d ($p = 0.39$). Both cell lines that underwent counting in this manner were more susceptible to the cytotoxic effects of DAC at 5 μM over 3 d, and that MDA-MB-231 cells did not withstand culture after exposure to DAC for 7 d incubation as well as MCF7 cells.

2.3.2 Differences in **ESR1** *IGM patterns between ER-pos and ER-neg cell lines*

In total, nine distinct regions across the pre-promoter, promoter and intragenic regions of the *ESR1* gene were assayed by pyrosequencing (Figure 2.1), and two distinct patterns were observed (Figure 2.3). Pre-promoter methylation levels in ER-positive and ER-negative cells were 79.1% vs. 22.5%, promoter methylation levels were 4.3% vs. 19.5%, and average intragenic methylation levels were 80.5% vs. 38.6%, respectively. ER-positive cells had particularly low levels of methylation at the TSS, as would be expected in a transcriptionally active gene. This region of hypomethylation extended into the first intron, with methylation levels of <5% at position 2 in ER-positive cell lines (Fig. 2.3), 1,627 nucleotides downstream from the

TSS. All six cell lines showed high methylation levels towards the 3'UTR, regardless of ER expression status.

2.3.3 In vitro ESR1 *methylation changes after DAC treatment*

Initial experiments as part of my MSc degree assess the effect of both 1 and 5 um DAC and DMSO over 3 or 7 d in MCF7 or MDA-MB-231 cells (Shenker, 2010). For the purposes of this PhD thesis, only the data from 7 d incubations with concentrations of 1 μm DAC or DMSO are shown. After 1 week of treatment with 1 μM DAC, both the ER-negative and ER-positive sets of cell lines showed a consistent decrease in methylation across the *ESR1* gene ($P < 0.05$; (Fig. 2.4a, b). The average promoter methylation level after DAC treatment was 3.72% and 5.45% in ER-positive and ER-negative cell lines, respectively, compared to the average post-DAC intragenic methylation levels which reduced from 80% to 50% (ER-positive) and 38% down to 31% (ER-negative) (Fig. 2a). ER-positive cells had persistently low levels of methylation at the CpG sites in the region of the TSS, which extended into the first intron (Fig. 2.4b).

Table 2.1: Cell counts (number of cells per ml x 10^5) after the completion of the DAC incubation experiments. The p-value represents the result of a t-test of the mean count from 3 and 7 d incubations versus the corresponding pre-incubation mean cell counts; green highlight, significant result.

		MCF7 cell line						MDA-MB-231 cell lines					
		DMSO	DAC		DMSO	DAC		DMSO	DAC		DMSO	DAC	p-
	Replicate	1 µM	1 µM	p-value	5 µM	5 µM	p-value	1 µM	1 µM	p-value	5 µM	5 µM	value
Pre-incubation	1st	10.5	10.5		5.49	5.49		12.1	12.1		10.9	10.9	
	2nd	9.5	9.5		9.12	9.12		10.2	10.2		13.5	13.5	
	3rd	1.25	1.25		4.41	4.41		16.5	16.5		14.6	14.6	
	Mean	**7.08**	**7.08**		**6.34**	**6.34**		**12.9**	**12.9**		**13**	**13**	
	SD	**5.08**	**5.08**	n/a	**2.47**	**2.47**	n/a	**3.23**	**3.23**	n/a	**1.9**	**1.9**	n/a
Three day incubation	1st	1.9	0.6		1.7	5.8		7.1	2.2		5.1	1.9	
	2nd	3.62	2.48		3.46	1.85		5.8	5.66		6.4	2.42	
	3rd	1.16	2.78		4.09	3.06		7.35	3.9		5.53	2.35	
	Mean	**2.23**	**1.95**	0.12	**3.1**	**3.56**	0.051	**6.76**	**3.92**	0.18	**5.69**	**2.23**	0.056
	SD	**1.26**	**1.18**		**1.21**	**2**		**0.84**	**1.73**		**0.64**	**0.27**	
Seven day incubation	1st	3.77	0.8		3.25	0.3		0.98	1.25		7.66	3.76	
	2nd	0.99	1.2		2.86	1.98		3.16	5.88		10.8	1.4	
	3rd	1.69	1.32		2.42	1.45		7.65	6.59		8.2	5.4	
	Mean	**2.15**	**1.11**	0.11	**2.84**	**1.24**	0.056	**3.93**	**4.57**	0.051	**8.89**	**3.52**	0.039
	SD	**1.45**	**0.27**		**0.42**	**0.86**		**3.4**	**2.9**		**1.68**	**2.01**	

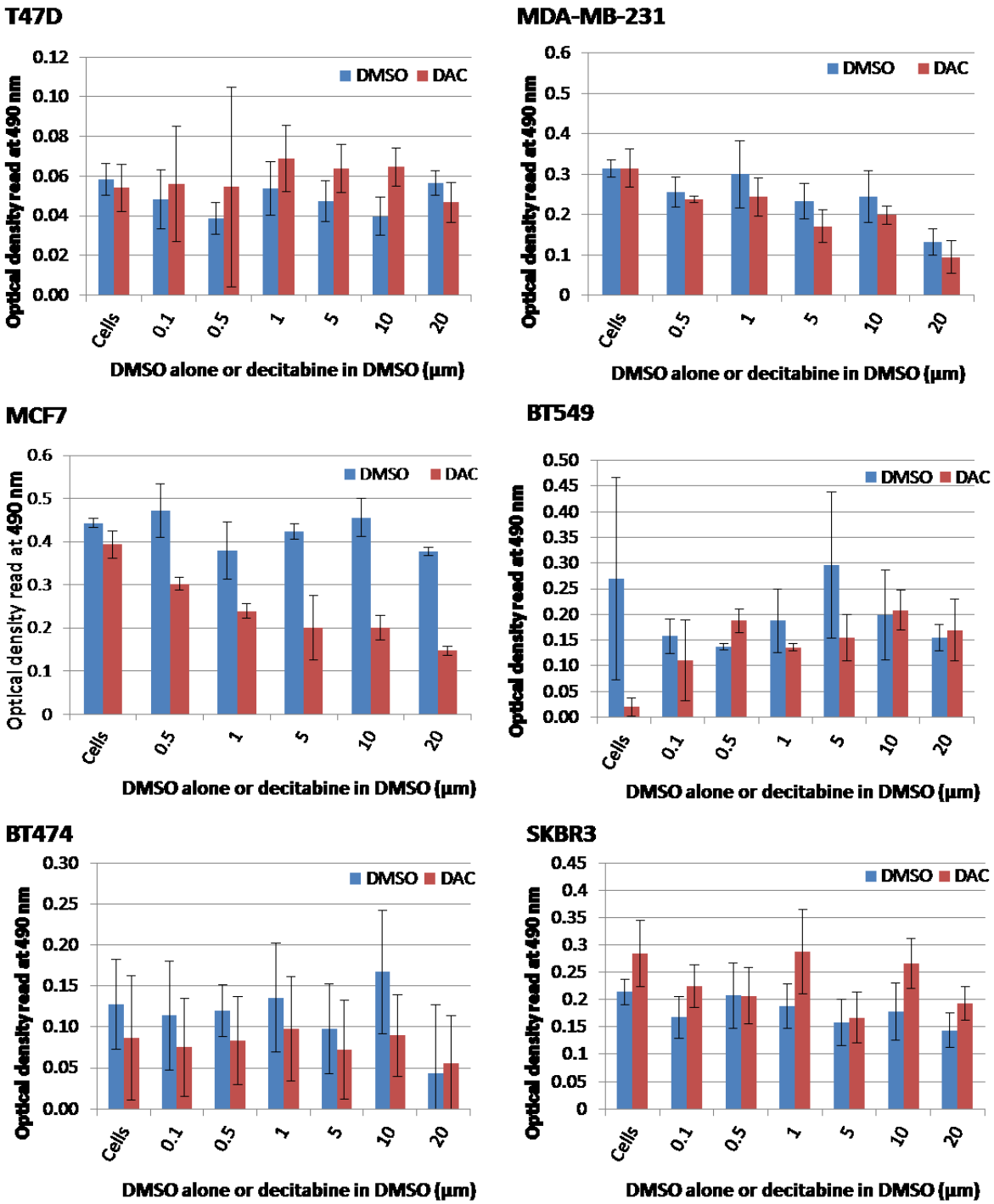


Figure 2.3: Results of the MTT assays conducted in all six breast cancer cell lines to assess the cytotoxicity of decitabine (DAC). The ER-pos cell lines are on the left, and the ER-neg are on the right side of this figure.

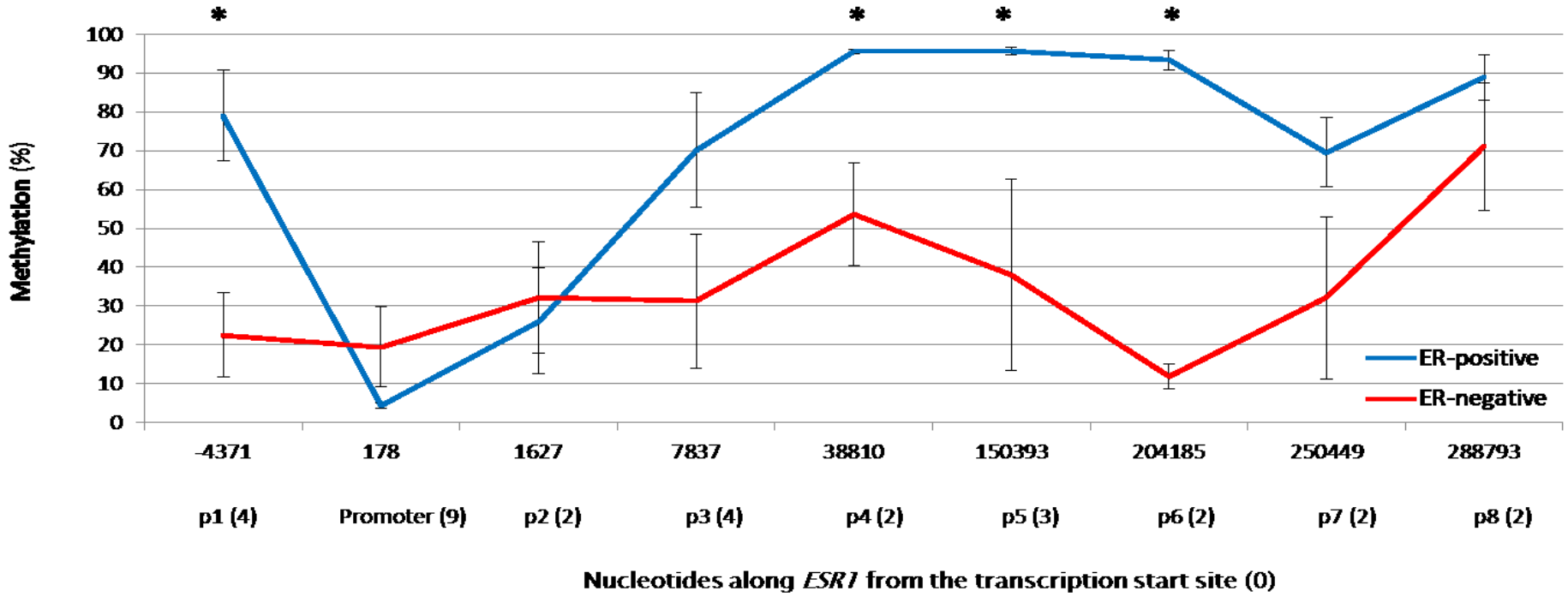


Figure 2.4: The DNA methylation levels across the *ESR1* gene in three ER-pos cell lines (blue) and three ER-neg cell lines (red). Data was collected in triplicate in each cell line at each locus (the number of CpG sites is shown in brackets at the loci names below the x-axis), and the standard error of the mean was then calculated (error bars); the loci can be mapped to specific regions of the *ESR1* gene according to the schematic shown in Figure 2.1. * $p < 0.05$, Bonferroni corrected t-test. Figure reproduced from BMC Cancer (Shenker et al., 2015).

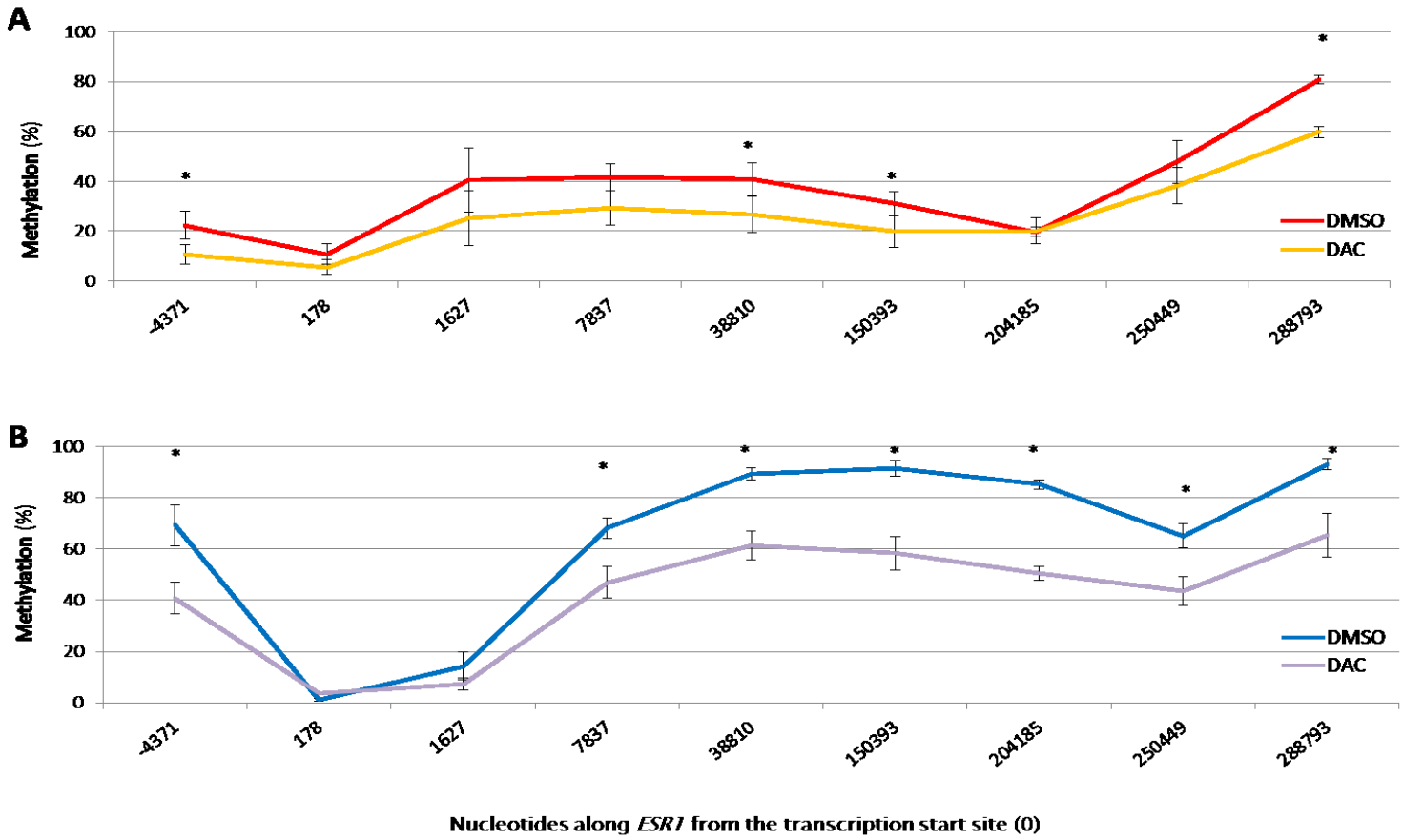


Figure 2.5: The effect of decitabine on DNA methylation in breast cancer cell lines. A) Composite graphs showing the DNA methylation levels across the ESR1 gene in three ER-neg cell lines treated with DAC (yellow) compared to DMSO-treated controls (red). Data was collected in triplicate in each cell line at each locus, and the standard error of the mean was then calculated (error bars); * $p < 0.05$, Bonferroni corrected t-test. B) DNA methylation across ESR1 gene in ER-pos cell lines treated with DAC (purple) compared to untreated (blue). The loci assessed can be mapped to specific regions of the *ESR1* gene using the schematic shown in Figure 2.1.

2.3.4 Expression levels of *ESR1* mRNA

We assayed the level of *ESR1* mRNA in all six human breast cancer cell lines in DAC-treated and control cells using RT-PCR. In ER-negative cells, 1 μM DAC increased *ESR1* mRNA expression significantly (MDA-MB-231 cells: fold change = 20.5; p = 0.003; BT549: fold change = 2.9, p = 0.037; SKBR3: fold change = 35.0, p = 0.023; Figure 2.5A). In ER-positive cells, treatment with DAC resulted in a fold change decreases that ranged from 0.31 in MCF7 cells (p = 0.023) to 0.42 in BT474 cells (p = 0.0034) and 0.37 in T47D ($p = 4.8x10^{-6}$; Fig. 2.5B).

Previously published data was used to validate this result. Gene expression profiles of MCF7 and MDA-MB-231 cells treated with DAC were downloaded from publically available databases (gse10613 and gse13733) (Chan et al., 2008, Miranda et al., 2009). Expression profiles were categorised into four quartiles, with the highest quartile representing the genes that had the top 25% expression levels prior to DAC treatment. The genes that mirrored the methylation status of *ESR1* after DAC treatment in MCF7 cells in our in vitro study could then be identified (Fig. 2.6; 4^{th} quartile genes with reduced expression after DAC treatment, n = 40). Not only were the findings regarding the methylation and expression of *ESR1* in accordance with our current findings, but these results confirmed that *ESR1* was one of the most downregulated genes in MCF7 following DAC treatment, while it was significantly upregulated in MDA-MB-231 cells. In addition, these gene expression profiles identify other genes, including several histone proteins, which are similarly downregulated upon DAC treatment (Fig. 2.6).

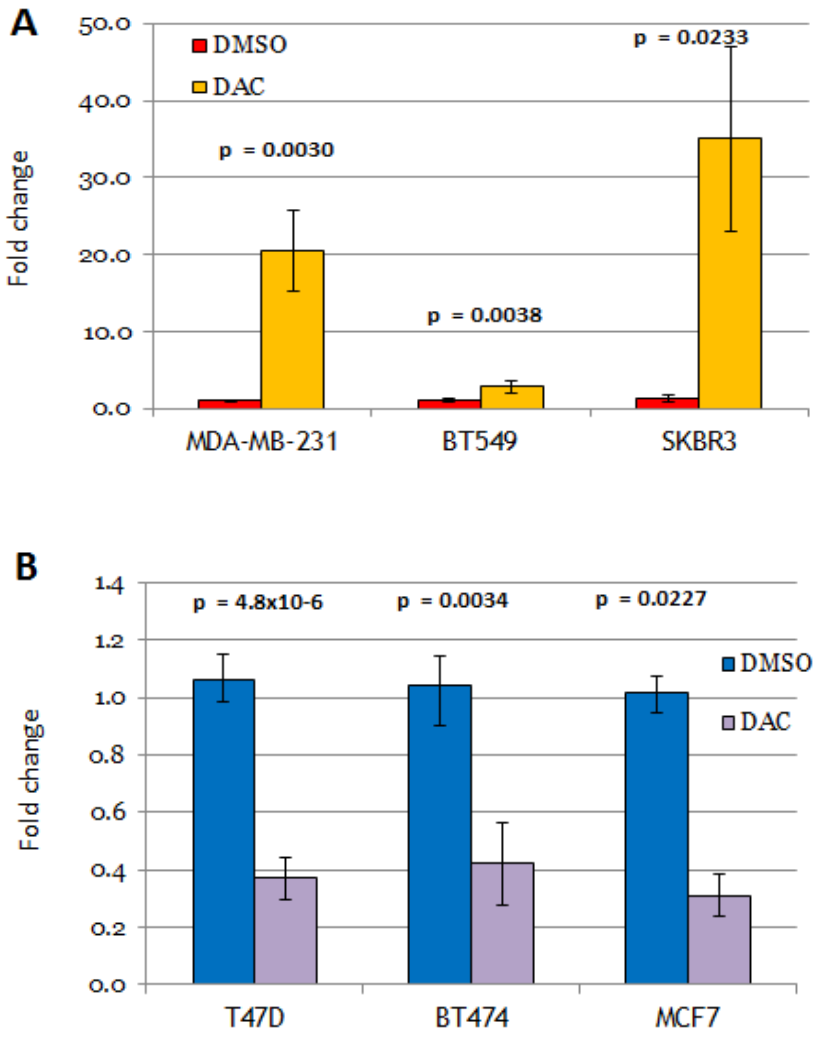


Figure 2.6: RT-qPCR gene expression in individual cell lines treated with DAC in A) three ER-neg cell lines, and B) three ER-pos cell lines. (ER-neg = yellow, ER-pos = purple) compared to DMSO-treated controls (ER-neg = red; ER-pos = blue). Expression increased significantly in ER-neg cell lines, but reduced significantly in all three ER-pos cell lines ($P < 0.05$). Figure reproduced with permission from BMC Cancer (Shenker et al., 2015).

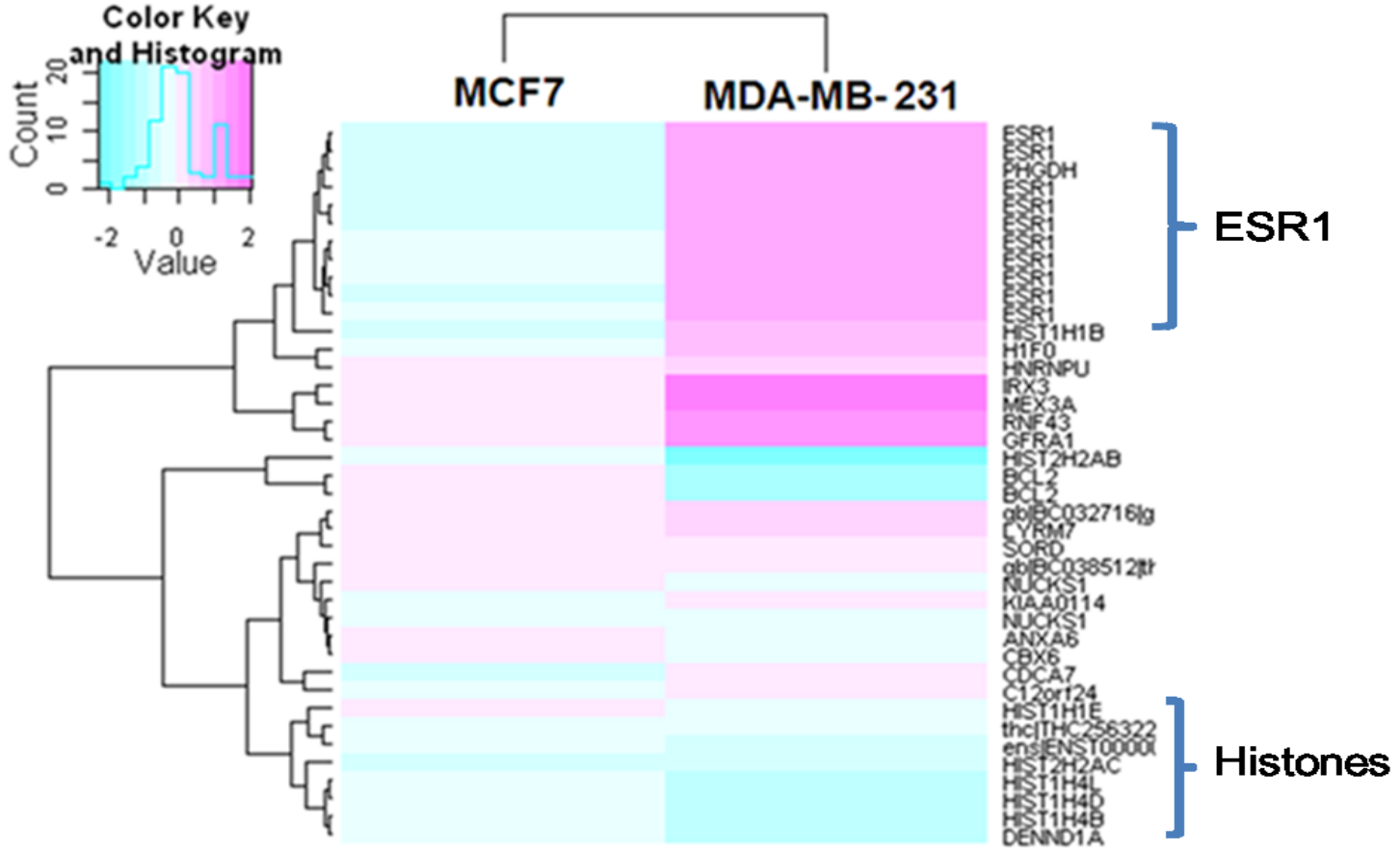


Figure 2.7: This heat map shows the relative protein expression of very highly expressed genes in MCF7 cells (n = 40 genes) that are downregulated after DAC treatment, compared to DAC-treated MDA-MB-231 cells from two publically available datasets (gse10613 and gse13733). Pink indicates upregulation, while blue indicates downregulation, with increasing intensity representing higher relative fold changes. The changes in ESR1 expression levels are in accordance with the findings from our current in vitro studies. Figure reproduced with permission from BMC Cancer (Shenker et al., 2015).

2.3.5 **ESR1** *IGM levels in DNA from breast tumour biopsies*

We received 155 breast tumour samples from the Breast Cancer Campaign Tissue Bank. These came in two separate sets: 20 fresh frozen tumour blocks with FFPE slides from the same tumours (ER-positive = 10; ER-negative = 10) and 135 tumours that were provided as FFPE slides (n = 45 ER-neg, n = 45 ER-positive grade 2, n = 45 ER-positive grade 3; 10 slides from each tumour, 3-µm sections).

New primers were designed to accommodate shorter fragments in FFPE-derived DNA (Appendix 4), and tested in the 20 fresh frozen samples vs. their matched FFPE DNA samples. There was good correlation between the results gained from the two methods of tissue preparation ($R^2 = 0.77$, data not shown), but there was significantly more variation in the FFPE compared to the FF tumours.

The promoter region of most tumours was unmethylated regardless of ER status. Furthermore, the intragenic pattern of methylation did not show as much variability across the entire gene as was observed in cell lines. The only region that was significantly different between ER-positive and ER-negative tumours were the two CpG sites 7,837 bp from the TSS in the first intron (MVR p3; $p = 0.01$). However, in contrast to the cell lines, the methylation levels at this region were higher in the ER-negative FF tumours (82.1%) compared to ER-positive tumours (57.3%; Fig. 2.7A), which was not predicted by the in vitro studies. This site was not significantly differently methylated between grade 2 and 3 ER-positive tumours, although grade 2 tumours had slightly higher levels of methylation at this region than grade 3 (62.1% vs. 51.5%; $p > 0.05$; Fig. 2.7B).

Data from the TCGA dataset showed a markedly similar pattern of methylation Fig. 2.8) across ESR1 to those of the BCC Tissue Bank tumour samples described above (Fig. 2.7). In particular, ER-negative tumours showed significantly higher beta

methylation values than ER-positive tumours at the three cg loci located within the first intron (cg04063345, cg15626350, cg00601836; $p < 1.0x10^{-7}$ for all), which were in the same region as the p2 MVR assayed by pyrosequencing.

2.3.6 Cell cycle analyses

Serum starvation resulted in an increase of cells stalled in the G1/G0 cell cycle phase (64% in controls treated with 10% FBS vs. 84% with 0% media $p < 0.05$). According to RT-PCR analyses after the serum starvation experiments, ESR1 expression was found to increase in a stepwise manner with increasing concentrations of serum in the medium ($R^2 = 0.86$), with ESR1 expression increasing by three-fold from 10% to 0% FBS. Regarding methylation, p2 DNA methylation decreased slightly from 11.8% to 9.8% in those treated with 0% serum ($R^2 = 0.92$, $p = 0.014$). Methylation did not change between samples at the p3 and p4 regions of interest.

To test whether DNA methylation and ESR1 expression are altered during different stages of the cell cycle, we used cell sorting to separate MCF7 cells grown in normal 10% serum media into different phases of the cell cycle and pyrosequenced DNA for p2, p3 and p4. There was no significant change in methylation across the cell cycle at all three assessed regions of interest. Methylation of p2 did not change significantly from 11.3% during G1, 11.35 and 11% during the S phase, and 10.6% during G2/M ($p > 0.05$). Methylation of p3 and p4 showed very similar results with methylation varying from 95.4% to 96.1% across the cell cycle ($p > 0.05$). Interestingly, one of the replicates showed a decrease of 8% at p4 during the S phase.

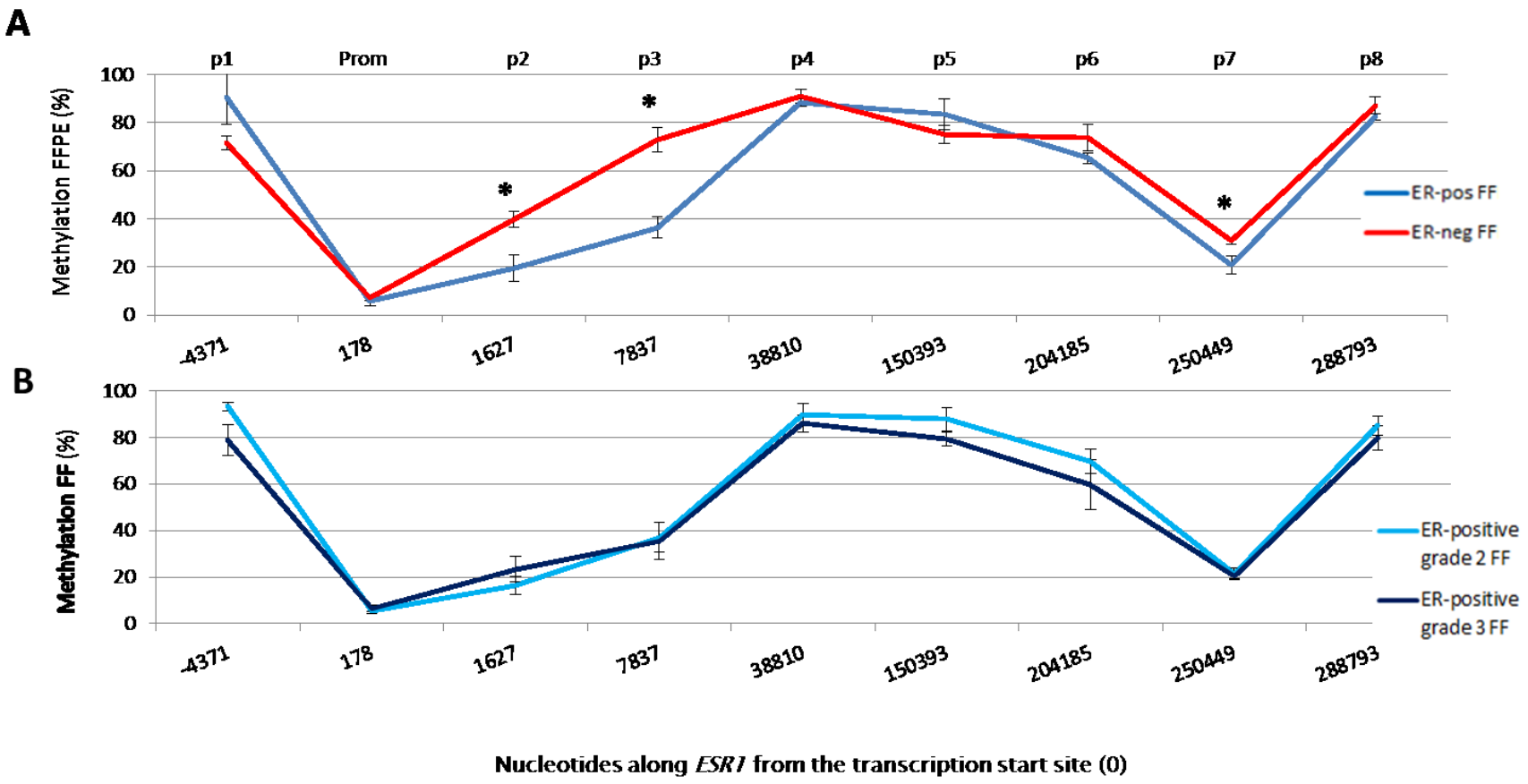


Figure 2.8: A) Pyrosequencing based methylation data across the ESR1 gene in fresh frozen (FF) human breast tumour samples (n=10 ER-negative, n=10 ER-positive [n=5 grade 2 vs. grade 3]). At regions p2 and p3 (approx. 2-8 kb downstream of the TSS), methylation levels were variable and significantly more methylated in ER-negative tumours, *p < 0.05. (B) In ER-positive FF tumours, methylation levels were very similar between grades. Figure reproduced with permission from BMC Cancer (Shenker et al., 2015). The loci assessed can be mapped to specific regions of the *ESR1* gene using the schematic shown in Figure 2.1.

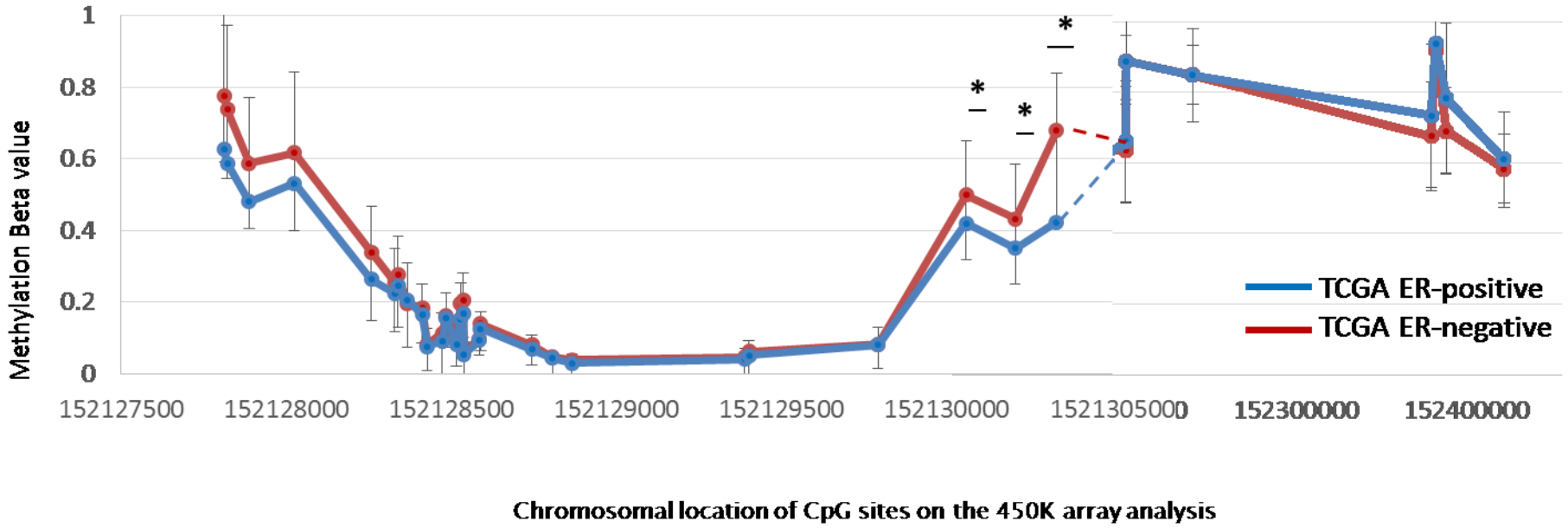


Figure 2.9: 450K Beadchip methylation analysis of ESR1 methylation from TCGA breast tumours (blue line = ER-positive tumours, n = 365; red line = ER-negative tumours, n = 109). The methylation pattern is very similar to the data shown from breast tissue biopsy material shown in Fig. 2.7A, with an unmethylated promoter region. The only differences appearing within the first intron (equivalent to regions p2 and p3 of the pyrosequenced assays in Fig. 2.7A, Wilcoxon rank sum test, * $p < 1.0x10^{-7}$. Figure reproduced with permission from BMC Cancer (Shenker et al., 2015).

2.3.7 Cell proliferation

A potential explanation for the difference between cell lines and tumour data may relate to cell proliferation rates, which also correlate with intragenic DNA methylation (Feinberg and Vogelstein, 1983). An analysis of cell proliferation was performed by staining human cancer tumour sections (n = 45 ER-pos, n=45 ER-neg) with an antibody against Ki67, which identifies the degree of cell proliferation. As expected, ER-negative tumours had significantly higher levels of Ki67 staining, and therefore higher proliferation, than ER-positive tumours (Table 2.2). The higher number of cells that are actively proliferating (i.e., in the S phase of the cell cycle; data not shown), would therefore, have higher intragenic methylation levels, according to previous work in human cells and tissues by Aran et al. (Aran et al., 2011), and may explain the higher levels of intragenic methylation observed in the ER-negative tumours. The degree of Ki67 staining was positively correlated with increased methylation levels at p2, p3 and p7, which were also the regions that were differentially methylated between ER-negative and ER-positive breast cancers (Table 2.2; Fig. 2.9).

2.3.8 Antisense mRNA levels of *ESR1*

Another potential mechanisms by which altering the methylation levels of intragenic CpG sites affects transcription is by altering the level of antisense mRNA that is transcribed. Primers were designed to amplify antisense mRNA across *ESR1*, which were then assayed in all six cells lines pre- and post-DAC treatment. The expression of the antisense strand was observed to increase markedly at a single site post-DAC in MCF7 and MDA-MB-231 cells (Fig. 2.10A, B), which was confirmed as a consistent finding in the two other ER-negative cells post-DAC treatment (Fig.

2.10C). However, two of the ER-positive cell lines had reduced levels of antisense post-DAC treatment (T47D: 0.68-fold, p = 0.029; BT474: 0.20-fold, p = 0.0086; Fig. 2.10D), but antisense at position 3 again increased by 5.07-fold in MCF7 cells (p = 0.0032). However, transfection of a specific 50 bp antisense RNA from this region into MCF-7 cells caused no change in *ESR1* methylation levels at any site or in levels of *ESR1* mRNA (Bell, 2013).

Table 2.2: Ki67 scores of stained tumour sections. Methylation levels at the intragenic variable region of interest (p3) and Ki67 scores were both higher in ESR1-negative tumours.

Ki67 score	Mean (%)	t-test	MVR p3 region	Mean (%)	t-test
ER-pos	10.00	0.0004	ER-pos	57.3	0.01
ER-neg	68.11		ER-neg	82.1	

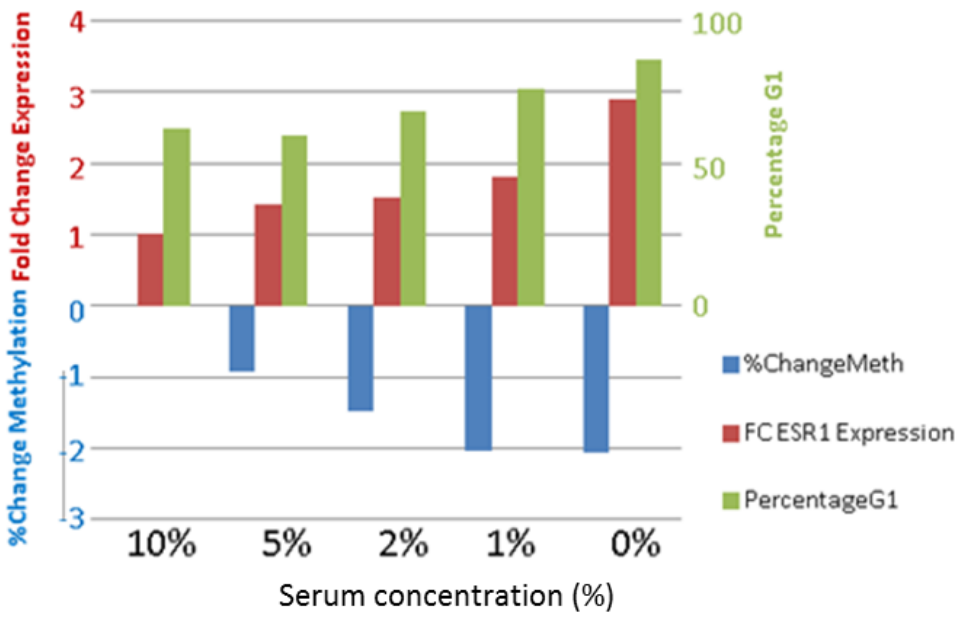


Figure 2.10: The number of MCF7 cells in the G0/G1 phase of the cell cycle increased significantly with decreasing concentrations of serum in the medium. ESR1 expression increased with decreased serum concentration ($R^2 = 0.86$; $p=8.8x10^{-4}$), and methylation levels decreased at the p2 variable region of interest.

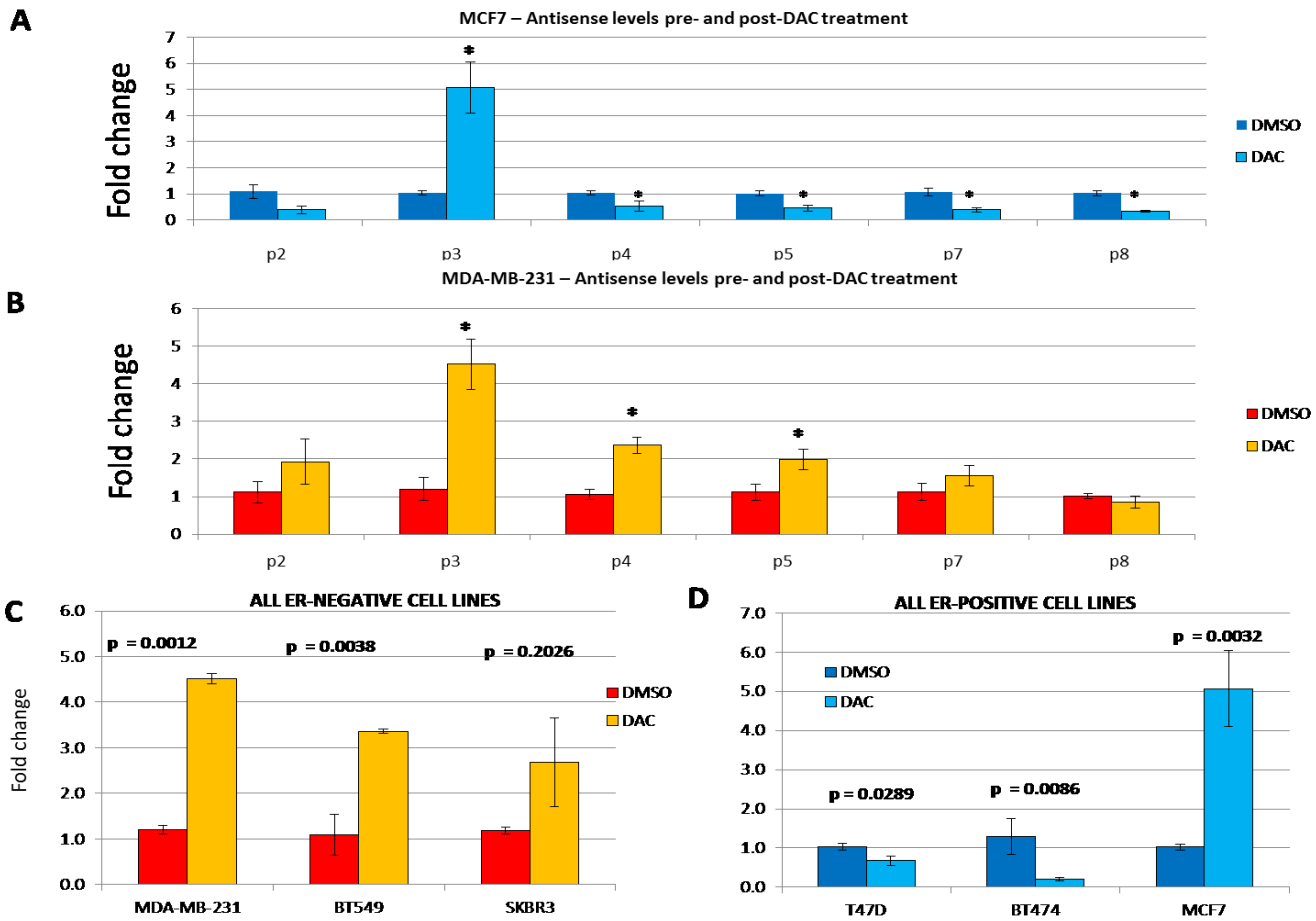


Figure 2.11: A, B) Expression changes for antisense across ESR1 in untreated (blue) and DAC-treated MCF7 and MDA-MB-231 (red). DAC treatment decreased the level of antisense expression in ER-positive MCF-7 cell lines, which was significant at almost all regions, with the exception of region p3, where antisense expression increased by 5-fold. The expression of antisense increased at every locus along ESR1 in ER-negative MDA-MB-231 cells, and this effect was most significant at region p3. * $p < 0.05$. C, D) Antisense expression changes in all six cell lines pre- and post-DAC treatment at position p3 (2nd intron of *ESR1*); C shows the three ER-negative cell lines, and (D) shows the three ER-positive cell lines.

2.4 DISCUSSION

This study identified marked and reproducible differences in the pattern of intragenic methylation across *ESR1* in vitro in ER-positive compared to ER-negative cell lines. The promoter of the ESR1 gene is contained within a 995-bp CpG island, which includes 105 CpG sites. The promoter region assayed was uniformly methylated in ER-negative cell lines, and unmethylated in ER-positive cells. As predicted, demethylation with DAC treatment increased the transcription of ESR1 in all three ER-negative cell lines, but the most surprising finding from this study was that DAC resulted in decreased levels of expression in ER-pos cells, via a mechanism independent from promoter methylation. Of note, the patterns of promoter and IGM established in tumour cell lines and a homogenous population of breast epithelial cells from EBM samples were highly similar, but differed markedly from those generated from two sources of ER-positive or ER-negative tumour biopsy samples (BCC Tissue Bank and the TCGA database). These observations are likely to reflect various caveats, including cell type heterogeneity and the tissue microenvironment, which results in a mixed epigenetic signal in tissue samples. This finding is in contrast to the artificial nature of cell lines grown on plastic in the presence of high concentrations of growth factors and has important implications on the choice of tissue for epigenetic analyses.

The principal dogma to date of the transcriptional effects of DNA methylation has focussed on promoter-associated CGI. Indeed, our results in ER-neg cell lines in this study corroborated the widely accepted view of DAC and promoter methylation function. The inhibition of DNA methyltransferases by 5’aza-2’deoxycytidine (Aza) was originally shown to cause a partial demethylation of the *ESR1* promoter CGI in

ER-neg breast cancer cell lines (Ferguson et al., 1995, Ferguson et al., 1997), enabling functional levels of ESR1 to be re-expressed. Furthermore, the selective inhibition of DNA methyltransferase 1 (DNMT1) with the use of specific antisense oligonucleotides also increased *ESR1* mRNA and protein concentrations in ER-neg cell lines, but further analysis of the ESR1 promoter showed very little demethylation (Yan et al., 2003). Other epigenetic mechanisms than DNA methylation mediated by DNMT1 may have contributed to the re-expression of ESR1 in ER-neg cell lines in this study, including the expression of non-coding RNAs, distorted DNA folding or altered upstream enhancer effects.

The more novel finding was the DAC treatment of ER-pos cell lines resulted in the expression levels of ESR1 reducing significantly in all three cell lines, which was an unexpected finding at the time. Since these experiments were performed, Yang et al. have demonstrated that DAC treatment reduces the expression levels of overexpressed genes on a genome-wide level in cancer, and may have a causal relationship (Yang et al., 2014). The *ESR1* promoter regions of these cell lines were already unmethylated, as confirmed in untreated controls, and therefore these were unaffected by DAC. However, marked intragenic demethylation occurred. One potential explanation for this finding was that DAC is a toxic drug, and could have resulted in cell stress, thereby disrupting the cellular functioning of these cells. However, MTT assays were performed in each cell line prior to the selection of the concentration of DAC (1 μm, Fig. 2.2) and cell death only occurred at concentrations of 10 μm in MCF7, MDA-MB-231 and BT474 cell lines. Another explanation is that DAC resulted in the deregulation of an upstream effector of ESR1 expression, but this hypothesis was not explored further in this set of experiments. The possibility that the decreased gene body methylation levels induced by DAC could have

contributed to reduced expression levels independent of promoter methylation levels cannot be discounted. Although the functional mechanism by which IGM exerts a transcriptional effect remains unknown, the transcriptional changes observed for ESR1 in ER-positive cells may represent an interesting insight into the basic biology of intragenic methylation levels.

The genomes of cancer cells undergo massive epigenetic changes with loss and redistribution of methylation. With the advent of array techniques that examine greater proportions of the genome, including high-density microarrays and next generation sequencing based DNA methylation analyses, the functional roles of intragenic methylation are becoming more apparent (Kulis et al., 2013) and have been linked to gene expression (Zilberman et al., 2007, Zemach et al., 2010b, Zemach et al., 2010a). Intragenic methylation levels have been shown to change markedly during carcinogenesis (Kulis et al., 2013), but in the absence of precise roles in the normal state, the effects of disrupted intragenic methylation levels in aberrant cells cannot be predicted or quantified. Several mechanisms by which intragenic methylation may be functionally important have been proposed, including the prevention of transcription from alternative start sites in the gene body, chromatin regulation, and the inhibition of transposable elements. Several studies have indicated a role of intragenic methylation with the control of alternative splicing, with increased IGM levels within exons that are included in spliced transcripts (Maunakea et al., 2013, Schor et al., 2013).

The methylation variable regions originally identified as differentially methylated in healthy control women compared with bilateral breast cancer were all in non-CGI regions (Flanagan et al., 2009). Alternative explanations require investigation to understand the functional role, if any, of the variable methylation of intragenic CpG

sites. High IGM levels may prevent the transcription of non-coding RNA in the antisense direction, although this finding was not supported by the results of this current study. Moreover, the methylation of intragenic transposable elements may affect transcription efficiency by impeding RNA polymerase II along the gene body (Lorincz et al., 2004), although a recent study proposes that intragenic DNA methylation represents a by-product of the chromatin assemblies related to transcription, and has no direct impact on transcription efficiency (Blattler et al., 2014).

The IGM levels in ER-positive cells decreased by a more substantial amount across the gene post-DAC treatment than in ER-negative cells, which raised questions regarding the cell proliferation rate of the cell lines. DAC is currently used as a clinical treatment for myelodysplastic syndrome and acute myeloid leukemia (Pleyer and Greil, 2015). DAC mediates its actions by incorporating into double-stranded DNA during cell replication, and therefore more rapidly dividing cells show greater levels of demethylation (Al-Salihi et al., 2011). The ER-positive cell lines were passaged more frequently than ER-negative ones, and all cell lines were cultured in media that contained oestrogen, fuelling a higher rate of replication. A link between DNA methylation and proliferation has previously been proposed by our group and others (Aran et al., 2011, Shenker and Flanagan, 2012). Aran et al. observed that proliferating cells and tissues tended to have higher levels of intragenic methylation (Aran et al., 2011). This observation was corroborated by our Ki67 and methylation data (Table 2.2, Fig. 2.9). We can therefore hypothesise that methylation can be influenced by proliferation rate, but is dependent on the specific genomic locus.

In terms of the DNA from tumour biopsies, ER-negative tumours had a much faster rate of cell proliferation, as indicated by the significantly higher levels of staining with Ki67 compared to ER-positive cells (Table 2.2). It is possible that the higher levels of intragenic methylation observed in ER-negative tumours was influenced by the higher levels of cell proliferation in these tumours. Velicescu et al. noted that serum starvation stalls cells at the G0/G1 phase of the cell cycle, preventing cell division and that DNA methyltransferases were predominantly expressed during the S phase (Velicescu et al., 2002). Active demethylation of promoter regions is known to be initiated by the same enzymes that induce methylation (DNMT3A and 3B) (Kangaspeska et al., 2008, Metivier et al., 2008), but is a long and energy expensive process (Ooi and Bestor, 2008). To establish if cell proliferation does actively affect intragenic methylation levels, and the subsequent effects on transcription, larger studies that investigate different cell lines and clinical samples with a genome-wide microarray approach may be required. Of note, the DMSO control methylation levels were comparable to those of the control DNA (untreated cells) in almost every instance. However, it is true that there were small differences at specific points, mainly affecting regions 4 and 5, which may indicate a certain amount of methylation cycling at these points, possibly at different points in the cell cycle, or in response to cell injury caused by the low concentration of DMSO used.

In terms of the antisense findings, although no definitive results were gained from the transfection studies, it is of significant interest that levels of antisense were raised significantly at region p2 across *ESR1*, the region at which there was most variable methylation between 2000-8000 kb from the TSS. The transcription of antisense mRNA was determined over a decade ago (Yelin et al., 2003, Katayama

et al., 2005), and it is largely thought to act in a repressive manner, in much the same fashion as siRNA transfections dampen the expression of sense strand mRNA by interfering with translation, resulting in the breakdown of the mRNA (Enerly et al., 2005). Initial speculation regarding the potential role of antisense transcripts in carcinogenesis (Tufarelli et al., 2003) via mRNA blockade have been largely supported by results from in vitro single gene assays (Yu et al., 2008b, Yap et al., 2010). Moreover, antisense transcripts could play a role at the promoter of certain genes (Yan et al., 2003), leading to a tantalising potential insight about the true function of promoter methylation. However, the rapid degradation of antisense transcripts in vivo and a poor understanding of their expression patterns mean that determining them accurately is currently difficult.

2.4.1 Limitations

This study had several limitations. Firstly, we have not investigated the effect of passive demethylation via DAC on distant enhancers or the many alternate promoters of *ESR1*. Only a single promoter was assessed, which generates the most abundant mRNA transcript in breast epithelial cells (Amaral et al., 2009), and the functions of the other transcripts generated from other promoters may hold clinical relevance. However, the gene body, which was the major focus of this research, was consistent. It would be interesting to determine whether gene body methylation levels vary dependent on which TSS initiates transcription of *ESR1* mRNA.

Secondly, the interesting findings determined in cell lines were not validated by the breast tumour biopsy results. A potential reason for this could have been confounding of the pyrosequencing because of the presence of 5-

hydroxymethylation, which is present in primary tissue but not in in vitro cultured cell lines. Pyrosequencing will not distinguish between these two marks. In Year 2 of my PhD, I attempted to develop the use of potassium permanganate in our laboratory, prior to the release of commercial products (Cambridge Epigenetix, Cambridge, UK) that had been previously published in Nature Protocols (Booth et al., 2013). However, this assay could not be optimised to determine the presence of 5-hydroxymethylation in these samples. Future studies will be required that use novel techniques for the detection of this epigenetic mark. This limitation is caveated by the fact that 5-hydroxymethylation levels are generally very low in cancer tissues (Delgado-Cruzata, personal communication), and could not explain the differences of 5-30% seen between cell lines and tumour biopsies in this study.

Thirdly, the quality of DNA extracted from FFPE tumour sections was relatively poor compared to that of fresh frozen samples, leading to higher technical variation. It will be important to optimise methylation assays further in this DNA to enhance the quality of the pyrosequencing results and enable a more direct comparison between in vitro findings.

Fourthly, the pyrosequencing analysis of the promoter region was very limited. This is a region that is densely packed with CpG site, which makes the design of suitable pyrosequencing primers highly challenging. From the 20 different assay pairs originally tested, just three worked consistently and reliably, without primer dimers forming or false positives from alternative binding regions. However, this limited the number of CpG assays in the *ESR1* promoter to just 9 out of the 105 CpG sites, and all of these were in the proximal region. It would be interesting to determine the methylation status of the entire promoter region in relation to the expression status of individual cell types, perhaps by using NGS based methods.

Additionally, this project did not attempt to understand the relative methylation levels at different ESR1 promoter regions, of which there are eight. The additional promoters of *ESR1* were not studied for the purposes of simplicity and cost, as they were not relevant to the principle aims of this study.

Finally, the choice of reference gene, *GAPDH*, for the RT-qPCR analyses was subsequently determined to be flawed. Although the results of the RT-qPCR assays were validated by publically available transcriptome data, used of the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines published in 2009 (Taylor and Mrkusich, 2014, Bustin, 2010, Bustin et al., 2009) suggest that GAPDH is not the optimal choice for breast cancer cell line work, due to variability in the expression of this reference gene (Gur-Dedeoglu et al., 2009). Conversely, GAPDH levels have been shown to remain stable in breast cancer cell lines after DAC treatment (Kastl et al., 2010).

2.4.2 Conclusions

Numerous studies that have investigated demethylating agents have shown that genes can be reactivated by the demethylation of promoter-associated CGIs. However, the repression of gene transcription by demethylation is often shown but widely ignored. The pattern of intragenic methylation was also interesting, with high levels being positively correlated with gene expression in *ESR1* in cell lines. However, this finding was not validated in human breast tumour biopsies, probably as a result of the significant heterogeneity represented by biopsy specimens, even after pathology review and macrodissection. This finding reflects the importance of

tissue selection for epigenetic analyses, and why findings should be interpreted with caution in both cell line and tissues experiments.

The reduction of ESR1 transcription after DAC treatment in ER-positive cells was investigated further in the search for functional insights into intragenic methylation, but few definitive conclusions can be made as yet about the role of gene body methylation, if any. Methylation may indeed represent a by-product of the transcriptional machinery processing along the DNA strand and altering chromatin conformation, rather than an active mechanism that affects transcription.

Regardless, this study offers insights into the importance of the methylation status of individual intragenic CpG sites. Future work will focus on the role of the cell cycle in methylation cycling, and its impact on gene expression. We will also focus on variable sites of intragenic methylation as predictive and prognostic biomarkers for breast cancer, as have already been proposed in ovarian cancer (Huang and Murphy, 2013).

CHAPTER 3

INVESTIGATING THE DNA METHYLATION SIGNATURE OF SMOKING AND ITS ASSOCIATION WITH BREAST CANCER RISK

3.1 INTRODUCTION

Epigenetic profiles, including patterns of DNA methylation, can be used to discern environmental exposures in cells, and may yield clues for resulting aberrant gene expression (Michels, 2010, Ogino et al., 2013). Such profiles, or signatures, are being defined that can unravel the pathogenesis of numerous complex diseases, in particular for diseases that have an environmental component to their aetiology, including a broad range of cancer types. Since the first studies by Richard Doll in the 1950s, including the seminal British Doctors Study in 1954 (Doll and Hill, 1954), smoking has been identified as a major risk factor for numerous cancers, most importantly lung cancer.

The smoking of tobacco-based products has been carried out for at least 7,000 years (Gately, 2001). As paintings of Aztec tradition in the Florentine Codex record (Reeves, 2006), ritualistic shamanistic practices throughout the Americas used smoke from the burning of tobacco plants. It was after European settlement of the continent started in the 1600s that this custom was adopted in the form of pipes, and exported firstly to Europe and from there around the world. Modern cigarettes are the most common form of tobacco delivery system, and enable the smoker to receive smoke into the lungs efficiently and relatively cheaply. Cigarette smoke contains over 5000 chemicals, many of them toxic and carcinogenic, including carbon monoxide, polynuclear aromatic hydrocarbons, dioxins, tar, cyanide, radioactive compounds and arsenic (Talhout et al., 2011). The addictive nature of cigarette smoke is largely the result of the action of nicotine, a mild stimulant with central nervous system effects. Nicotine acts directly as an agonist for nicotinic cholinergic receptors in both the dopamine system (Herman et al., 2014), the central nervous system reward system, and the endogenous cannabinoid system (Gamaleddin et al., 2015).

Nicotine can also enable smokers to maintain a low weight as a result of the replacement of reward effects from food (Zoli and Picciotto, 2012). Furthermore, nicotine withdrawal usually leads to anxiety and irritability, which can be severe (Jackson et al., 2014). Taken together, these observations explain why some smokers find cigarettes almost impossible to stop smoking, despite the well-known warnings about their deleterious health effects and a vast range of public health interventions that aim to reduce smoking rates.

Smoking is a risk factor for a wide range of cancers throughout a range of organ systems and cell types (Table 3.1) (Sasco et al., 2004, Carbone, 1992, Pirie et al., 2013). Smokers are also at a higher risk than non-smokers of developing subsequent cancers after being successfully treated for a range of primary tumours (Shiels et al., 2014). Tissues with a highly increased relative risk (RR) for the development of cancers in smokers tend to be those in direct contact with the tobacco smoke, including the mouth and throat, lungs, liver, and kidneys, which filter many of the toxic compounds excreted after their absorption in the lungs. Colorectal cancer is moderately association with smoking (Hannan et al., 2009), but tumour-type differences play a significant role in terms of the RR. Limsui et al. determined that ever-smoking conferred a risk for all colorectal tumours of just 1.19 (RR), while current smoking was associated with a RR of up to 1.99 in tumours with microsatellite instability caused by errors in the DNA mismatch repair systems (Limsui et al., 2010).

For breast cancer, the data regarding risk is complex, and depends strongly on the timing of smoking initiation. Two large studies published during the early 2000s, one meta-analysis of over 50,000 women and one report from the USA Surgeon General's office, indicated that smoking was not a risk factor for cancer (Glantz and

Johnson, 2014, Office, 2004, Hamajima et al., 2002). However, a large body of evidence has now accumulated to suggest that smoking in pre-menopausal women is a risk factor, either by the individual smoking (Glantz and Johnson, 2014, Reynolds, 2013) or by second-hand smoke exposure, or passive smoking (Johnson and Glantz, 2008, Johnson et al., 2011, Dossus et al., 2014).

Furthermore, a dose-response relationship exists whereby women who start smoking as teenagers are at an even higher risk than women who start smoking in their 20s or 30s (Millikan et al., 1998). In post-menopausal women, some evidence suggests that smoking may even be protective against breast cancer as toxins induce the cytochrome p450 system in the liver (Kendall et al., 2007). This enzymatic system metabolises tobacco smoke toxins, as well as oestrogen. Lower circulating levels of oestrogen are therefore protective. However, the theory is confounded by the increase in carcinogenic byproducts from tobacco smoke constituents, and the complexity of this relationship remains opaque.

The complexity is deepened by the influence of the individual's genotype on risk related to smoking. Variants in *ESR1* have been implicated in increased risk of breast cancer in passive smokers (Tang et al., 2013). Most of the related literature however focuses on polymorphisms within N-acetyltransferase (NAT)-1 and -2 and breast cancer risk (Ochs-Balcom et al., 2007, Weber and Nathanson, 2000). NAT1 and NAT2 catalyse the metabolism of a number of aromatic and heterocyclic amine compounds within cigarette smoke, producing carcinogenic metabolites (Hein, 2000). Certain genotypes in postmenopausal women may lead to increased risk in smokers as shown by a meta-analysis almost a decade ago (Terry and Goodman, 2006), but these authors noted that all such epidemiological studies are limited by

the inter-study methodological heterogeneity, debatable questionnaire accuracy, underpowered studies and the lack of a reliable biomarker of smoke exposure.

Table 3.1: Organs at risk of carcinogenesis as a result of carcinogens in tobacco smoke, as assessed by the International Agency for Research on Cancer Working Groups, adapted from data in (Sasco et al., 2004, IARC, 2004, Hannan et al., 2009, Wenbin et al., 2013, Bennett et al., 2015). RR: Relative risk.

Organ	RR (average)
Lung	15-30
Urinary tract	3
Oral cavity	4-5
Pharynx	4-5
Larynx	10
Oesophagus	2-5
Pancreas	2-4
Nasal cavity, sinuses	1.5-2
Stomach	1.5-25
Liver	1.5-2.5
Kidney	1.5-2
Acute myeloid leukaemia	1.5-2
Breast	1.14
Gallbladder	1.5
Colorectal	1-2
Small intestine	1.25

Improvements in array-based technology have enabled the molecular effects of altered environments, such as cigarette smoke inhalation, to be assessed in larger populations. The first study that investigated methylation levels in smokers' blood was conducted by a group in Germany in 2011 (Breitling et al., 2011). They used the 27K Illumina HumanMethylation Beadchip array to assess the DNA methylation profiles of plasma samples from individuals that had never smoked compared to those who were current smokers or who had smoked within the last 10 years. Their results showed a significant link between smoking and the epigenome at a single locus in the *F2RL3* gene, out of 27,578 loci tested, that was less methylated in current smokers (n = 65) compared to former (n = 56) and non-smokers (n = 56). The observation of lower methylation levels at this locus was validated using a

different technique and in a validation cohort, and remained consistent. They also showed the same trend between former smokers and non-smokers, but this was not significant in this study. *F2RL3* is associated with platelet activation and coagulation. Changes in *F2RL3* methylation in smokers may represent altered expression in PBMCs, which could contribute to smoking-related pathological processes such as coagulopathies, stroke and cardiovascular disease.

The introduction of the Illumina 450K methylation bead array has enabled the analysis of the DNA methylation pattern across the genome with additional coverage of promoters, 5'UTRs, first exons, gene bodies and 3'UTRs. The molecular changes that occur as a consequence of the neoplastic process itself, in addition to changes caused by chemo- and radiotherapy, imply that any investigation into epigenetic alterations induced by cancer risk factors must be performed prospectively in pre-diagnostic samples. This present study aimed to assess the impact of current and former smoking on DNA methylation in an epigenome-wide approach using prospectively collected blood samples from age- and sex-matched pairs of healthy individuals. One individual in each pair subsequently developed breast or colon cancer. A methylation signature was determined in both current smokers and former smokers, and a model was created to predict whether an individual was or had been a smoker, and validated against the current gold standard for smoke exposure, cotinine (Yuchuan et al., 2011, West et al., 2011, Florescu et al., 2009, Benowitz et al., 2009).

Our findings of a reduction in methylation in former and current smokers enabled the further generation of a methylation index that could be used to identify former and current smokers with a much higher level of specificity and sensitivity than cotinine quantification, the current gold standard assay.

3.2 METHODS

3.2.1 Study participants

Study participants were drawn from the Italian component of the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort, a large general population cohort consisting of approximately 520,000 individuals with standardized lifestyle and personal history questionnaires, anthropometric data and blood samples collected for DNA extraction (Riboli and Kaaks, 1997, Riboli, 2001). For the microarray component of this study, participants were sampled from 10,604 individuals collected in Turin, and included 95 incident breast cancer cases and 95 matched controls. We excluded three male breast cancer cases and their matched controls leaving 92 female cases and 92 female controls for the analysis. These included 41 ER+ and 10 ER- breast tumour patients and 41 tumours with an unknown ER status. The second nested case-control study included 95 incident colon cancer cases and 95 matched controls including 68 male and 27 female pairs. Controls were individually matched on age (±5 years), sex, seasonality of blood collection and length of follow-up. Blood samples from cancer cases were taken 55 months before diagnosis on average (range, 24-108 months) for breast cancer cases and 84 months before diagnosis (range, 0.2-173) for colon cancer cases (Table 1).

Ethics approval was gained by the Turin EPIC group from the IARC Ethics Committee, and all necessary biobank access and material transfer documentation were completed and approved.

3.2.2 Microarray hybridisation (conducted by Dr Silvia Polidoro at the HuGeF Foundation, Turin, Italy)

DNA samples were extracted from buffy coats using the QIAsymphony DNA Midi Kit (Qiagen, Crawley, UK). Bisulphite conversion of 500 ng of each sample was performed using the EZ-96 DNA Methylation-Gold™ Kit according to the manufacturer's protocol (Zymo Research, Orange, CA). Then, 4 µl of bisulphite-converted DNA was used for hybridization on the Infinium HumanMethylation 450 BeadChip, following the Illumina Infinium HD Methylation protocol. Briefly, a whole genome amplification step was followed by enzymatic end-point fragmentation and hybridization to HumanMethylation 450 BeadChips at 48°C for 17 h, followed by single nucleotide extension. The incorporated nucleotides were labelled with biotin (ddCTP and ddGTP) and 2,4-dinitrophenol (DNP; ddATP and ddTTP). After the extension step and staining, the BeadChip was washed and scanned using the Illumina HiScan SQ scanner. The intensities of the images were extracted using the GenomeStudio (v.2011.1) Methylation module (1.9.0) software, which normalizes within-sample data using different internal controls that are present on the HumanMethylation 450 BeadChip and internal background probes.

3.2.3 Validation of microarray data findings

For the validation component of this study, 180 healthy control individuals were randomly sampled from the EPIC-Turin cohort and DNA samples diluted in distilled RNase-free water were delivered frozen on ice to our laboratory from Turin. This group comprised 33 current smokers, 45 former smokers and 102 individuals who had never smoked. Extensive information on smoking and tobacco use was collected from a baseline questionnaire at enrolment, including smoking duration and

intensity, and this data is available in Table 3.2 in the Results section. The DNA concentration and quality were checked by the Nanodrop1000 spectrophotometer (ThermoScientific, UK) and stored at -20°C prior to their analysis.

Genomic DNA (250 ng) from each subject in the validation cohort was bisulphite-converted using EZ-96 DNA Methylation-Gold kit according to the manufacturer's protocol (Zymo Research, Irvine, CA, USA). For pyrosequencing, specific primers were designed for each assay using PyroMark software (Qiagen, Hilden Germany), and PCR conditions were as described in Chapter 2 of this thesis, and in Flanagan et al. (Flanagan et al., 2010). A semi-nested approach was used for each primer set to increase specificity and yield. A list of genes and primers can be found in Appendix 5. The Pyro Q-CpG Software conducts an inbuilt overall quality assessment for each sample, which flags any sequence that deviates from the expected pattern. Any sample that failed quality control was removed from the analysis. Additionally, a commercially available fully methylated genomic DNA sample was used as a positive control (Zymo Research) and in-house whole genome amplified genomic DNA (Genomiphi, GE Healthcare) was used as an unmethylated negative control, in addition to water controls. Methylation values were calculated as an average of all high quality CpG sites, which were determined as "passed" by the quality control thresholds within the Pyro Q-CpG software (Qiagen).

3.2.4 Human lung analyses

Pyrosequencing assays for the loci of interest as identified from the microarray and validated by pyrosequencing were conducted on bisulphite-converted genomic DNA extracted from 27 human lung samples, kindly provided from a collaboration with the

Respiratory Pharmacology Group led by Professor Maria Belvisi at the National Heart and Lung Institute, Imperial College. Human lung samples (n = 27; 14 smokers, 13 non-smokers) were obtained either from lung transplants performed at The Royal Brompton or Harefield Hospital or purchased from IIAM (International Institute for the Advancement of Medicine, Edison, NJ, USA). In all cases, consent for the use of tissue for scientific research was provided by all study participants, and ethics approval was obtained from the Royal Brompton and Harefield Trust.

Accordingly, DNA was extracted and pyrosequencing was performed using the techniques and assays described above. RNA was also extracted from lung tissue (n = 5 smokers, n = 5 non-smokers) using the Qiagen mini-RNA extraction kit (Qiagen, Hilden, Germany), and quantitative RT-PCR (RT-qPCR) was performed on a Bio-Rad PCR machine with SyBR green (Sigma), for AHRR normalized against GAPDH, using standard protocols. DNA and RNA quality and concentrations were assessed using the Nanodrop 1000 Spectrophotometer (BioRad).

3.2.5 Murine model of smoking

Lung tissue samples were collected from 10 male C57BL/6 mice (weight: 16-20 g; obtained from Harlan UK Ltd.) exposed to air (n = 5) or cigarette smoke for 3 days (n = 5) or 28 d (n = 5). These mice were from an experiment conducted by Dr. Mark Birrell, a postdoctoral fellow in the Respiratory Pharmacology Group. The smoking model was achieved by a previously published technique for smoke exposure (Eltom et al., 2011). Ethics approval for this part of the study was obtained from an institutional ethics committee (project no: MTS 01-10) as part of a project licence (PPL 70/6681), and was conducted in accordance with UK Home Office Guidelines for animal welfare, according to the Animals in Scientific Procedures Act of 1986.

Briefly, mice were exposed to cigarette smoke (3R4F non-filter cigarettes, Tobacco Health Research Institute, University of Kentucky, Lexington, KY) after being placed in exposure chambers (Teague Enterprises, CA, USA) for 50 minutes twice a day, followed by a 10-min period for venting of the chamber. Control mice were placed into adjacent chambers for the same time periods, but were exposed to room air instead of cigarette smoke. The smoke was dispersed evenly throughout the chamber by a fan, and the total suspended particulate levels were assayed every 15 minutes to ensure that the concentration of smoke was consistent between chambers. The smoke exposure experiments continued for 3 or 28 days. Mice were humanely sacrificed at 24 hours after the last exposure by an intraperitoneal injection with phenobarbitone and samples of lung tissue were dissected, flash-frozen with liquid nitrogen and stored at -80°C until use.

RNA was extracted using above-mentioned techniques and RT-qPCR assays were performed for *Ahrr*, *Ahr*, *Cyp1a1* and *F2rl3*, also as described above (primers shown in Appendix 5). These genes were selected as homologous forms of the human genes with CpG loci of interest (*Ahrr* and *F2rl3*), as well as important genes in the dioxin metabolism pathway in which *AHRR* is a critical component (Evans et al., 2008). A ribosomal gene, *Rpl7*, was used as the internal reference standard.

3.2.6 Development of the methylation index to predict smoking status

All study participants were drawn from the Turin component of the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) cohort, a general population cohort that consists of approximately 10,000 individuals with standardized lifestyle and personal history questionnaires, anthropometric data and blood samples collected for DNA extraction (Riboli, 2001, Riboli and Kaaks, 1997). Smoking status

was ascertained from questionnaire data. For the test sample set, 81 healthy individuals were sampled from 1,805 that had been previously measured for serum cotinine (Timofeeva et al., 2011), including 33 non-smokers, 30 former smokers and 18 current smokers. These included 66 males and 15 females with no significant gender bias in smoking status (chi squared test, p = 0.12). Smoking duration was calculated as age at recruitment minus age start smoking (for current smokers) and age at quitting minus age start smoking (for former smokers). For the validation component of this study, 180 healthy individuals were randomly sampled from females enrolled in the EPIC-Turin cohort (n = 102 non-smokers, n = 45 former smokers, n = 33 current smokers).

The loci included one in the *AHRR* gene (Chr5: 373,299), two intergenic loci at 2q37 (Chr 2: 233,284,112 and Chr 2:233,284,661) and one intergenic locus at 6p21.33 (Chr 6: 30,720,080). Individuals (test, n = 11 and validation, n = 42) that were heterozygous for a previously identified G>A SNP at the first locus in *AHRR* (Chr 5: 373,299) were removed from the analysis to avoid confounding of the *AHRR* methylation results. The pyrosequencing assay for this locus was used to determine both the methylation assay and the genotype at this CpG site. Alternative CpG sites that could be used in place of this AHRR site included cg21161138, for which a pyrosequencing assay was developed, or cg05575921, for which a pyrosequencing assay could not be developed.

The Illumina 450K DNA methylation microarray processing and analysis methods were as described above, and the pyrosequencing data for the four loci was taken from that of the validation cohort (n = 180) also described above. As these data represent actual methylation values, no normalization was performed on the pyrosequencing data.

3.2.7 Data analysis (conducted by myself and Dr. James Flanagan)

For the microarray data, the methylation score for each CpG was represented as a β-value according to the fluorescent intensity ratio representing any value between 0 (unmethylated) and 1 (completely methylated). The raw microarray data and processed normalized data were submitted to the Gene Expression Omnibus (GEO) (accession no: GSE51057) in order that our data would be publically available. For the statistical analysis, raw data was exported from GenomeStudio (Illumina) as background subtracted β-values with corresponding detection p-values. Any probe not detected ($p > 0.05$) was converted to "NA", and any sample with >3% of probes not detected, and any probe not detected in >20% of samples was removed from the data set. Quality control measures identified a general failure of chip "5" (Barcode 5809079028) which was removed from the final breast cancer EWAS analysis. The resulting data sets included 86 breast cancer cases with 87 matched controls (86 matched pairs) with 95 colon cancer cases and 95 matched controls. Overall, 484,804 probes were analyzed in the breast cancer data set and 485,152 in the colon cancer data set. SNP variants were removed before data analysis. Data was normalized using quantile normalization using the R package "preprocessCore". Multivariate linear regression was used to identify associations between methylation β-values as the outcome and coded smoking status as the exposure (0, 1 and 2 for 'Never', 'Former' and 'Current' smokers, respectively), adjusting for age and batch. A secondary analysis was performed that also adjusted for case-control status, which did not significantly alter the results (Supplementary Table 2). P-values $<1x10^{-7}$ were considered to be significant at the level of epigenome-wide significance. False discovery rate (FDR) analysis was used to adjust for multiple testing, using FDR of 5% at the level of epigenome-wide significance. For association with smoking

intensity, coded intensity categories were used as the exposure. Time to quitting (Age at recruitment – Age stopped smoking) and duration of smoking (Age stopped – Age started smoking) in former smokers were analysed as continuous variables. Manhattan plots and quantile-quantile plots were generated using mhtplot function from the “gap” R package and the qqplot function from the “stats” R package, respectively. Leucocyte subpopulations could not be formally assessed in this analysis.

For the pyrosequencing data, we used multivariate linear regression to compare groups adjusting for patient age and batch (plate for pyrosequencing), and p-values less than 0.05 were considered to be significant. False discovery rate (FDR) analyses were used for each locus to adjust for multiple testing (n = 5).

To construct the methylation index, a stepwise iterative generalized linear regression model was generated for the data from never smokers versus former smokers starting with the pyrosequencing data from six loci. 14 We excluded one CpG site each in AHRR (cg21161138) and F2RL3 (cg03636183) as these were not found to be independent of the other loci ($p > 0.05$ in the linear regression model) and were indeed highly correlated with the other four markers ($R^2 > 0.6$). Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated using receiver operative curve (ROC) analyses using R software. Area under the curve (AUC) values were calculated for each genomic locus and the overall methylation index (MI) and compared to those for cotinine in differentiating between never versus former smokers. Binomial tests were used to ascertain the significance of PPVs. All statistical analyses were performed in R, v2.13.1 and significance was set at the level of $p < 0.05$.

3.3 RESULTS

3.3.1 Epigenome-wide association studies

In this study, we performed two epigenome-wide association studies (EWAS) on genomic DNA from peripheral white blood cells (WBC) that were prospectively collected from two nested case-control studies within a large cohort from the general population (Table 3.2). All individuals were healthy at the time of blood collection, but cases were selected from individuals who subsequently developed either breast or colon cancer (average duration to diagnosis = 7.07 and 7.37 years, respectively).

Table 3.2: Characteristics of the three cohorts investigated in this study.

	EPIC-Turin (Breast CaCo)*		EPIC-Turin (Colon CaCo)*		EPIC – Turin (Validation)	All subjects
	Controls	Cases	Controls	Cases	Controls	
Individuals, n	92	92	95	95	180	554
Age at blood, yr, median (range)	50 (34-63)	51 (34-65)	55 (35-64)	55 (35-64)	49 (35-65)	211
Sex – male, n	0	0	68	68	0	136
Sex – female, n	92	92	27	27	180	238
Time before diagnosis, months, median (range)	n/a	55 (24-108)	n/a	84 (0.2-173)	n/a	
Smoking status **						
Never, n	46	48	34	32	102	160
Former, n	15	18	30	37	45	100
Current, n	29	18	24	20	33	91
Excluded from analysis after quality control, n	6	5	0	0	n/a	n/a
Smoking intensity, median (range) ***	4 (2-7)	4 (2-7)	4 (2-9)	4 (2-7)	4 (2-9)	

* CaCo, case-control study. Cases and controls were individually matched on age, recruitment centre, date and time of blood collection.
** Differences in numbers within categories of population characteristics and total numbers are due to missing/unknown data for these variables.
*** Smoking intensity: 2 (1-3 cig/day), 3 (4-8), 4 (9-13), 5 (14-18), 6 (19-23), 7 (24-28), 8 (29-33) and 9 (>34). n/a, not applicable.

For the present analysis, we used multivariate linear regression to investigate the association between DNA methylation levels and smoking status, adjusting for age and batch. Using a cutoff of $p < 1x10^{-5}$, we identified 17 and 19 loci in the Breast Cancer and Colon Cancer EWAS data, respectively, which were differentially methylated between smokers, former smokers and those who had never smoked; eight of these loci were shared by both studies ($p < 1x10^{-7}$) (Fig. 3.1). The top hits for both studies ($p < 1x10^{-5}$) are shown in Appendix 6. In all instances, the degree of methylation was lower in smokers than non-smokers. Of note, the *F2RL3* locus previously identified was in this list.

There was no strong association between methylation levels at any of these loci and disease status ($p < 1x10^{-5}$), despite smoking being a weak risk factor for colon cancer (Sasco et al., 2004, Hannan et al., 2009). We observed no evidence for an association between smoking and breast cancer risk in this study ($p = 0.857$), which was likely to be a consequence of the small numbers of participants in the study. One of the validated loci (cg01940273) at 2q37.1 showed a weak association with developing breast cancer, after adjustment for smoking ($p = 0.003$) and estrogen receptor status ($p = 0.035$) (Table 3.3). This region warrants further investigation in larger studies as a cancer risk marker.

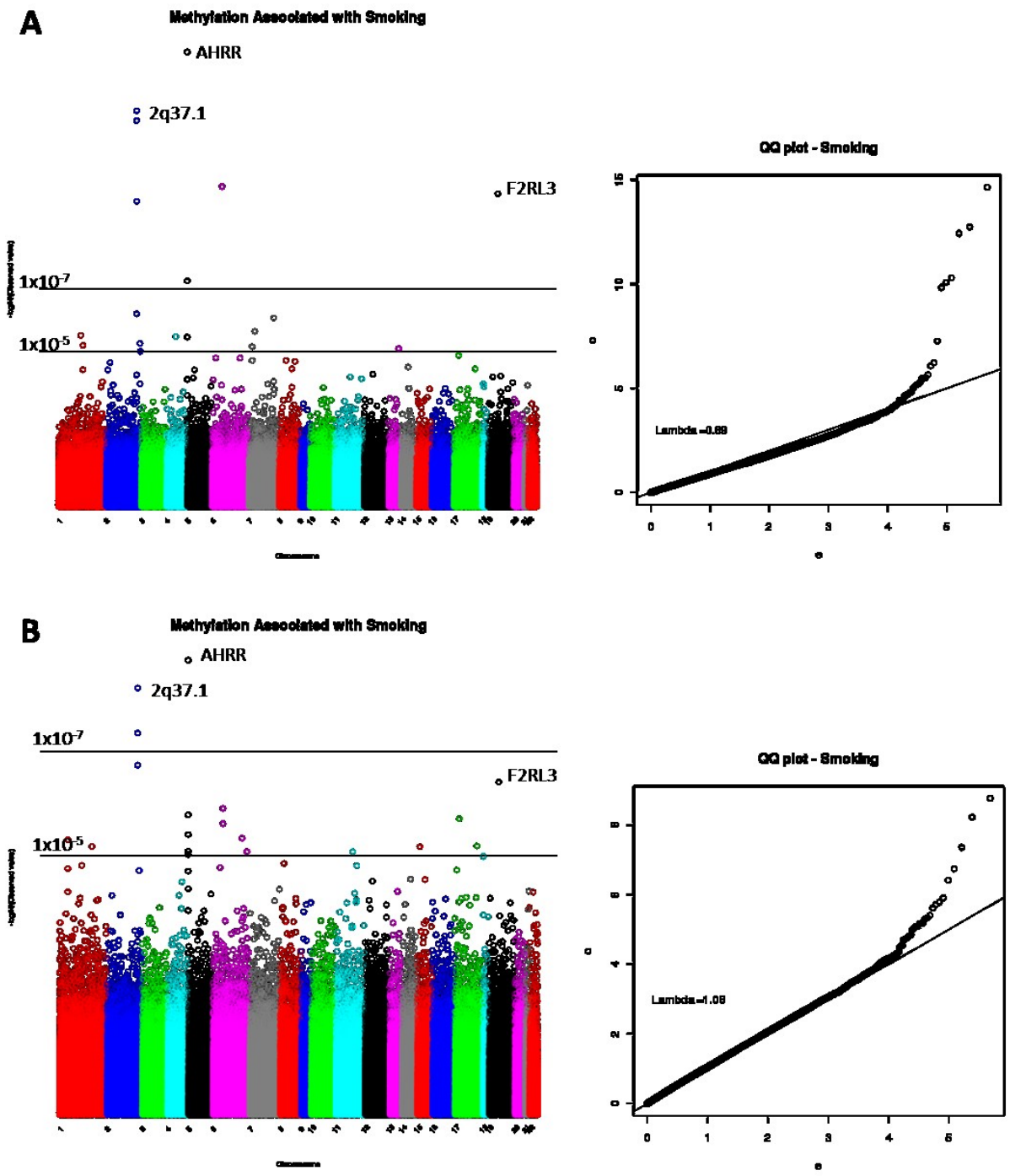


Figure 3.1: Manhattan plot and quantile-quantile plot for EWAS results for smoking status in two case-control studies. A) Breast cancer case-control study. B) Colon cancer case-control study. In the Manhattan plot, the vertical axis indicates ($-\log_{10}$ transformed) observed p-values, and the horizontal thresholds indicate the significance levels of $p = 1 \times 10^{-5}$ and $p = 1 \times 10^{-7}$. In the quantile-quantile (QQ) plot, the horizontal axis shows ($-\log_{10}$ transformed) expected p-values, and the vertical axis indicates ($-\log_{10}$ transformed) observed p-values. The lambda inflation factor (median[obs]/median[exp]) is shown. Reproduced with permission from Hum Mol Genet (Shenker et al., 2013a).

Table 3.3: Leading differentially methylated genomic loci between smokers and former or non-smokers in a breast and colonic case-control cohort. This list contains the top eight overlapping CG sites between the breast cancer (BCa) and colon cancer (CCa) EWAS studies, in addition to an additional site within the 2q37.1 locus that was also validated by pyrosequencing. Rows in bold indicate the data that was independently validated by bisulphite pyrosequencing of that locus in a separate EPIC cohort of healthy individuals (n = 180) with known smoking status. Reproduced with permission from Hum Mol Genet (Shenker et al., 2013a). N:

Target ID	*Chr*	*MAPINFO (bp)*	*Symbol*	*Beta values in BCa case-control cohort*					*Beta values in CCa case-control cohort*					*Pyrosequencing of the validation cohort*				
				N	F	C	P-value *	Effect size (%) **	N	F	C	P-value *	Effect size (%) **	N	F	C	P-value (FDR)	Effect size (%) **
cg06644428	2	233284112	2q37.1	0.07	0.05	0.05	6.17E-07	2	0.11	0.09	0.09	3.38E-04	3	12.11	11.02	8.11	1.48E-05	4
cg05951221	2	233284402	2q37.1	0.39	0.33	0.28	1.80E-13	11	0.41	0.37	0.34	1.83E-07	7	nd	nd	nd	nd	n/a
cg21566642	2	233284661	2q37.1	0.44	0.37	0.32	3.73E-13	12	0.51	0.47	0.39	4.41E-08	12	58.86	53.04	40.11	2.22E-08	18.75
cg01940273	2	233284934	2q37.1	0.58	0.56	0.49	1.47E-10	9	0.62	0.59	0.54	5.96E-09	8	nd	nd	nd	nd	n/a
cg23576855	5	373299	AHRR	0.66	0.64	0.5	3.46E-06	16	0.73	0.68	0.53	9.66E-06	20	82.68	80.17	66.91	1.51E-12	15.77
cg05575921	5	373378	AHRR	0.84	0.79	0.68	2.31E-15	17	0.84	0.81	0.7	1.73E-09	14	nd	nd	nd	nd	n/a
cg21161138	5	399360	AHRR	0.66	0.65	0.6	5.44E-08	5	0.72	0.71	0.68	8.29E-06	4	75.47	74.67	68.3	3.22E-05	7.17
cg03636183	19	17000585	F2RL3	0.64	0.61	0.56	8.38E-11	8	0.68	0.65	0.61	3.84E-07	7	62.19	60.92	54.32	0.00172	7.87
cg06126421	6	30720080	6p21.33	0.65	0.61	0.57	4.96E-11	8	0.71	0.7	0.64	2.46E-06	7	nd	nd	nd	nd	n/a

Never smoker; F: former; C: current.

Chr, chromosome; FDR, false discovery rate; nd, not done; n/a, not applicable.

* Logistic regression adjusting for age and batch effect; ** effect size represents percentage methylation difference between current smokers and individuals who have never smoked.

Table 3.4: Relationship to case-control status adjusted for smoking status, and for oestrogen receptor status in breast and colon cancer.

Target ID	Chr	MAPINFO (bp)	Symbol	BrCa (Ca-Co)	BrCa (adj. for smoking)	ER+ BrCa (adj. for smoking)	CoCa (Ca-Co)	CoCa (adj. for smoking)
cg06644428	2	233284112	2q37.1	0.589	0.315	0.197	0.718	0.512
cg05951221	2	233284402	2q37.1	0.266	0.261	0.079	0.591	0.447
cg21566642	2	233284661	2q37.1	**0.038**	**0.018**	**0.034**	0.983	0.890
cg01940273	2	233284934	2q37.1	**0.008**	**0.003**	**0.035**	0.512	0.300
cg23576855	5	373299	AHRR	0.930	0.752	0.885	0.121	0.197
cg05575921	5	373378	AHRR	0.254	0.168	0.868	0.255	0.517
cg21161138	5	399360	AHRR	0.240	0.272	0.846	0.419	0.501
cg03636183	19	17000585	F2RL3	0.017	0.012	0.131	0.299	0.565
cg06126421	6	30720080	6p21.33	0.535	0.662	0.660	0.511	0.965

3.3.2 Validation of microarray loci of interest by pyrosequencing

Bisulphite pyrosequencing validated the results of five CpG sites identified by the 450K array (2 CpG sites in *AHRR,* 2 CpG sites in 2q37 and 1 CpG site in *F2RL3*). Primers were designed to amplify four further loci, but these failed at either the PCR stage or the pyrosequencing stage as a result of poor quality amplification, the formation of primer dimers or non-specific binding at another site. Furthermore, one of the five sites amplified in *AHRR* (AHRR_p1), was an A>G single nucleotide polymorphism (SNP). Therefore, the results at this locus of individuals who carried the A SNP (either homozygotes or heterozygotes) were excluded as a CpA dinucleotide cannot be methylated. Of the 180 DNA samples, the number of passed individuals for each loci were 149 (AHRR_p1, cg23576855; 31 excluded), 176

(AHRR_p3, cg21161138), 153 (2q37_p1, cg06644428), 117 (2q37_p3, cg21566642) and 174 (F2RL3, cg03636183).

The direction of methylation change and effect sizes were of the same magnitude in the validation groups, and again, a significant association with smoking was observed (Table 3.3). There was also remarkable evidence for an association between methylation levels and smoking intensity for the *AHRR* and 2q37 loci, but not *F2RL3*, with individuals smoking ≥4 cigarettes per day having significantly lower methylation levels at these genomic loci (Figure 3.4). In former smokers, methylation levels at these genomic loci returned to the levels of non-smokers with increasing time from cessation and those who had smoked more intensively had methylation levels that were closer to those of current smokers (Table 3.5, Figure 3.4).

The CpG site in one of the *AHRR* regions of interest (AHRR_p1, cg23576855) was also the site of a CG→CA SNP (rs6869832), with an A allele frequency of approximately 10% as confirmed by pyrosequencing. Minor allele (A) carriers (n = 31) were excluded from the pyrosequencing analysis, as the cytosine could not be methylated in the CpA dinucleotide. Interestingly, three of the current smokers in the validation set were heterozygous carriers; their methylation levels (mean, 37.7%) were approximately half of the value found in smokers who were homozygous for the G allele (mean, 66.9%; Table 3.3). This indicated that the CpG site on the G allele was methylated to a similar degree as homozygous G alleles. We did not have genotyping on the individuals in the EWAS sample sets, but predict that the individuals with ~40% and 5% β-values are likely to be heterozygous and homozygous for this SNP, respectively (Figure 3.2).

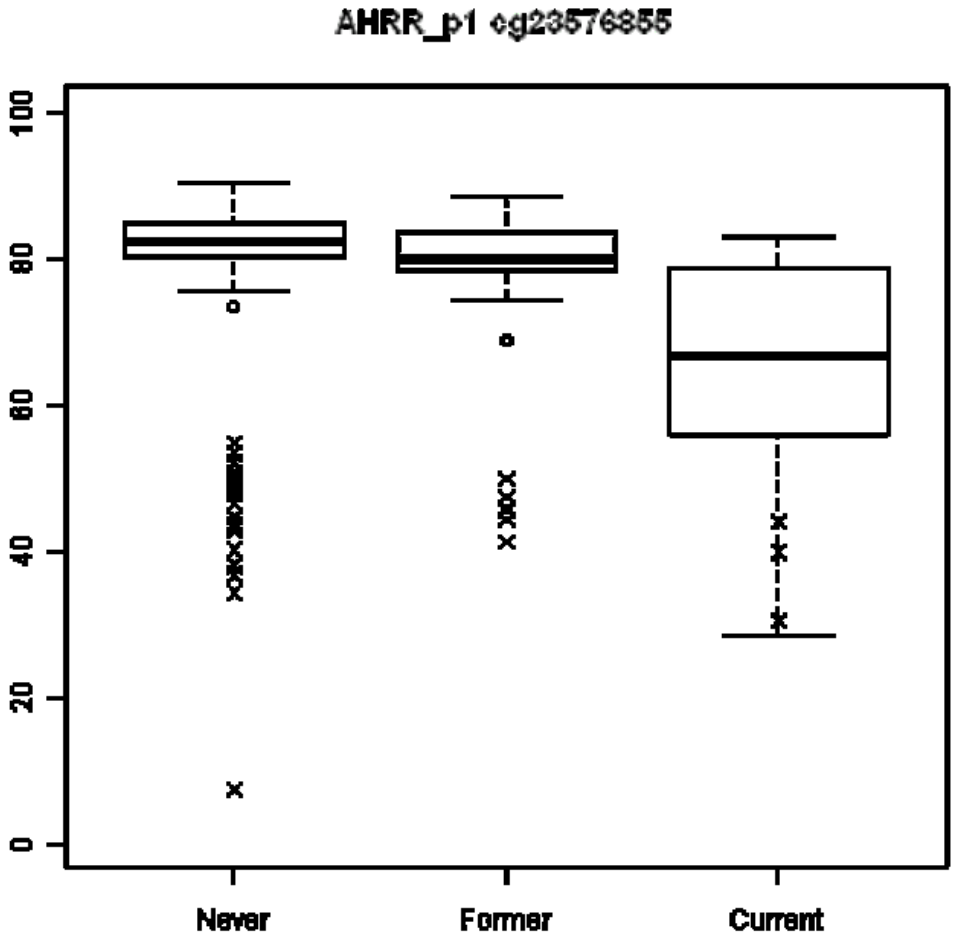


Figure 3.2: Box-plot representation of pyrosequencing-derived methylation percentages for the first locus identified from AHRR (cg23576855) in the validation control population (n = 180).

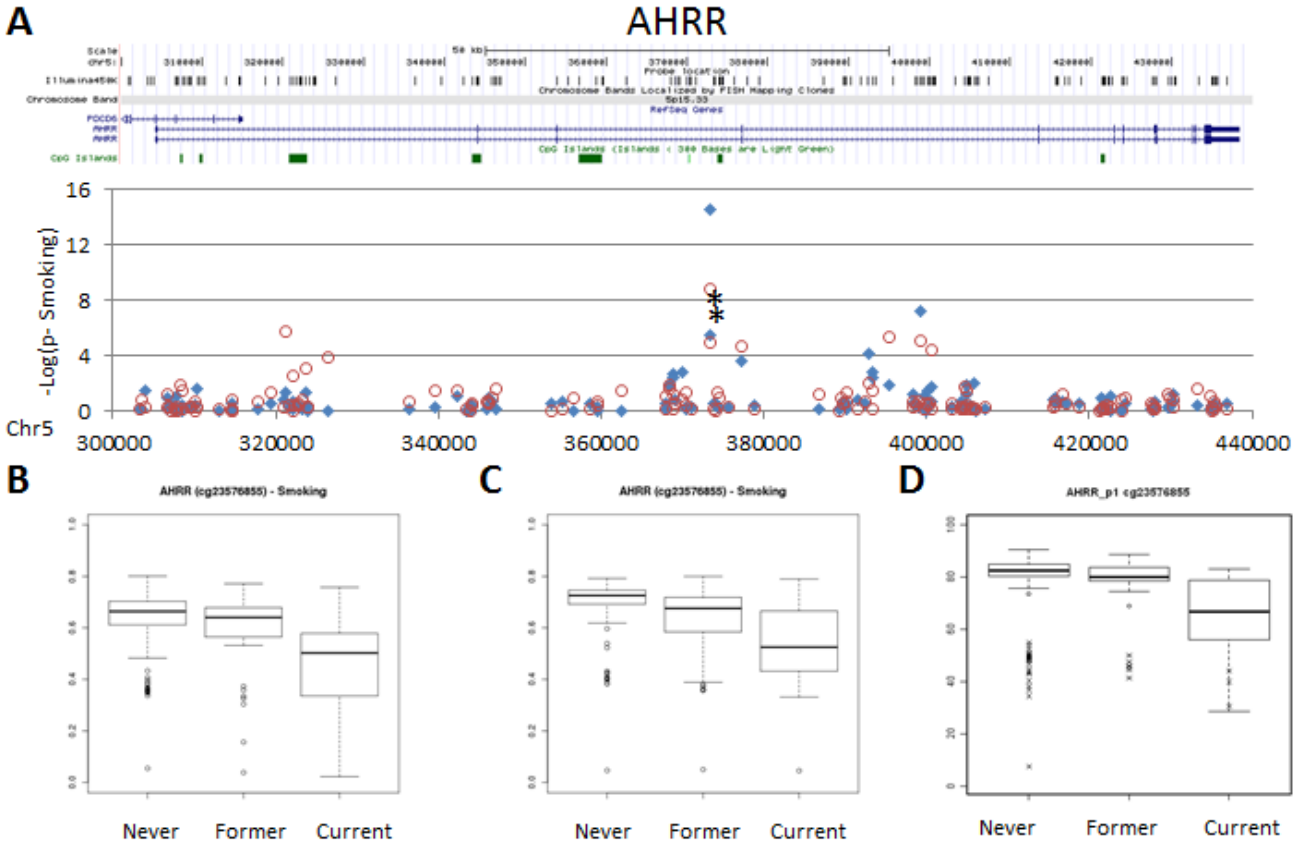


Figure 3.3: Association between smoking status and AHRR. (A) A regional association plot is shown with the gene map (from the UCSC genome browser, hg19) with a graph of –log10 p-values on the y-axis and the nucleotide position on the x-axis for the Breast Cancer EWAS (blue diamonds), Colon-Cancer EWAS (red circles) and pyrosequencing validation (black asterisks). (B-D) Boxplots represent β-values for the top hit in the Never, Former and Current smoker cohorts in the breast cancer EWAS (B), colon-cancer EWAS (C), and pyrosequencing-based methylation levels in the validation control population (D). Reproduced with permission from Hum Mol Genet (Shenker et al., 2013a).

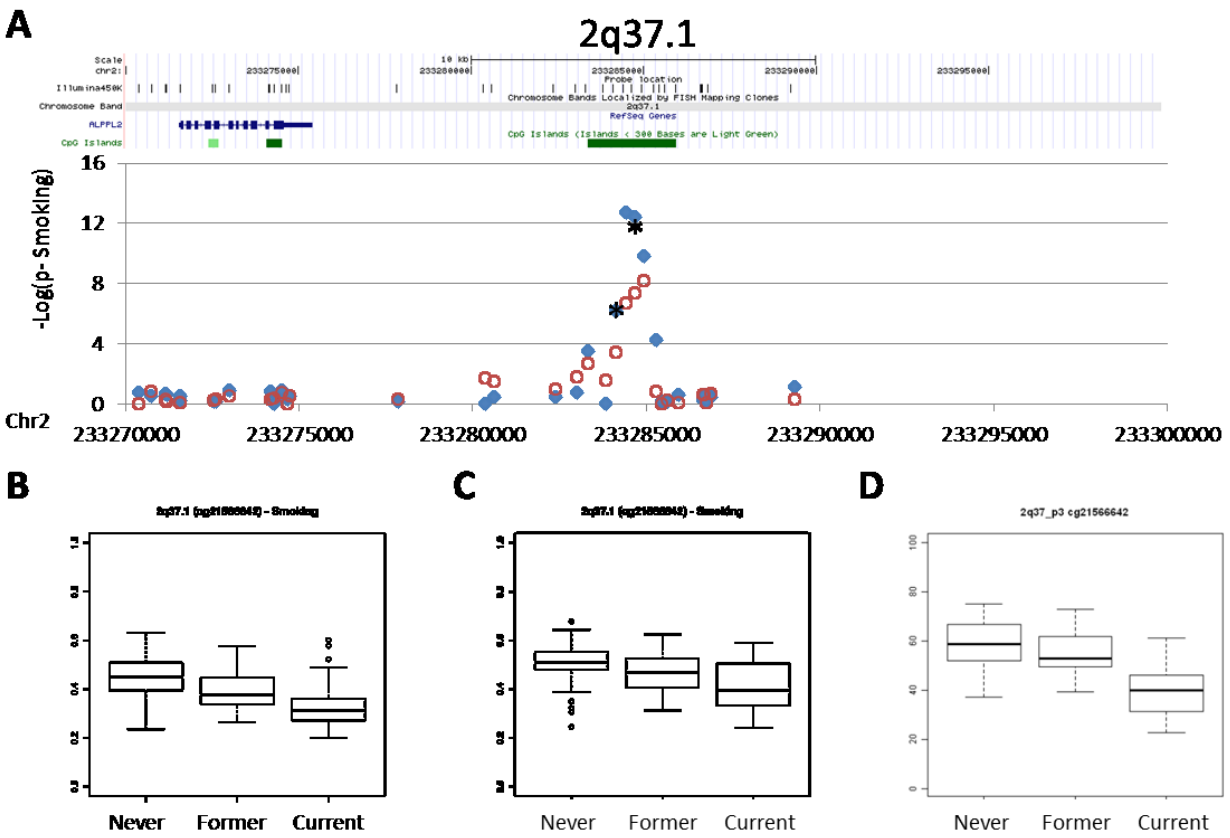


Figure 3.4: Association between smoking status and 2q37. (A) As shown in Fig. 3.2, the regional association plot is shown with the gene map (from the UCSC genome browser, hg19) with a graph of –log10 p-values on the y-axis and the nucleotide position on the x-axis for the Breast Cancer EWAS (blue diamonds), Colon-Cancer EWAS (red circles) and pyrosequencing validation (black asterisks). (B-D) β-values for the top hit in the Never, Former and Current smoker cohorts in the breast cancer EWAS (B), colon-cancer EWAS (C), and pyrosequencing-based methylation levels in the validation control population (D). Reproduced with permission from Hum Mol Genet (Shenker et al., 2013a).

Table 3.5: Relationship between methylation and smoking intensity. General linear regression adjusting for age and batch, in smokers only, was used to investigate the interaction between methylation (outcome) and smoking intensity (exposure). General linear regression adjusting for age and batch, in former smokers only, was used to investigate the interaction between methylation (outcome) and time since quitting smoking or length of time smoking (exposure). The p-value for each probe is shown within the table for the breast cancer EWAS, colon cancer EWAS and the pyrosequencing validation cohort. In former smokers, methylation levels returned to the levels of non-smokers with increasing duration after cessation, and those who smoked more intensively had methylation levels closer to those of current smokers. Increasing intensities in current smokers was coded as 2 (1-3 cigarettes per day), 3 (4-8), 4 (9-13), 5 (14-18), 6 (19-23), 7 (24-28), 8 (29-33) and 9 (>34). nd, not done. Reproduced with permission from Hum Mol Genet (Shenker et al., 2013a).

				Breast cancer EWAS			**Colon cancer EWAS**			**Pyrosequenced controls**		
Target ID	Chr	MAPINFO (bp)	Symbol	Smoking intensity (smokers only)	Time since cessation (former smokers)	Duration of smoking (former smokers)	Smoking intensity (smokers only)	Time since cessation (former smokers)	Duration of smoking (former smokers)	Smoking intensity (smokers only)	Time since cessation (former smokers)	Duration of smoking (former smokers)
cg06644428	2	233284112	2q37.1	0.948	0.34	0.591	0.436	0.009	0.015	0.643	0.505	0.2495
cg05951221	2	233284402	2q37.1	0.046	0.048	0.083	0.791	0.015	0.007	nd	nd	nd
cg21566642	2	233284661	2q37.1	0.04	0.049	0.155	0.265	0.018	0.009	0.027	0.144	0.008
cg01940273	2	233284934	2q37.1	0.097	0.034	0.097	0.023	0.022	0.007	nd	nd	nd
cg23576855	5	373299	AHRR	0.007	0.275	0.476	0.006	0.488	0.379	0.005	0.808	0.725
cg05575921	5	373378	AHRR	0.003	0.204	0.328	0.045	0.002	3.76E-05	nd	nd	nd
cg21161138	5	399360	AHRR	0.002	0.151	0.326	0.078	0.002	0.015	0.095	0.007	0.089
cg03636183	19	17000585	F2RL3	0.047	0.179	0.317	0.149	5.66E-04	1.97E-05	0.404	0.397	0.333
cg06126421	6	30720080	6p21.33	0.04	0.085	0.188	0.479	8.49E-04	2.40E-05	nd	nd	nd

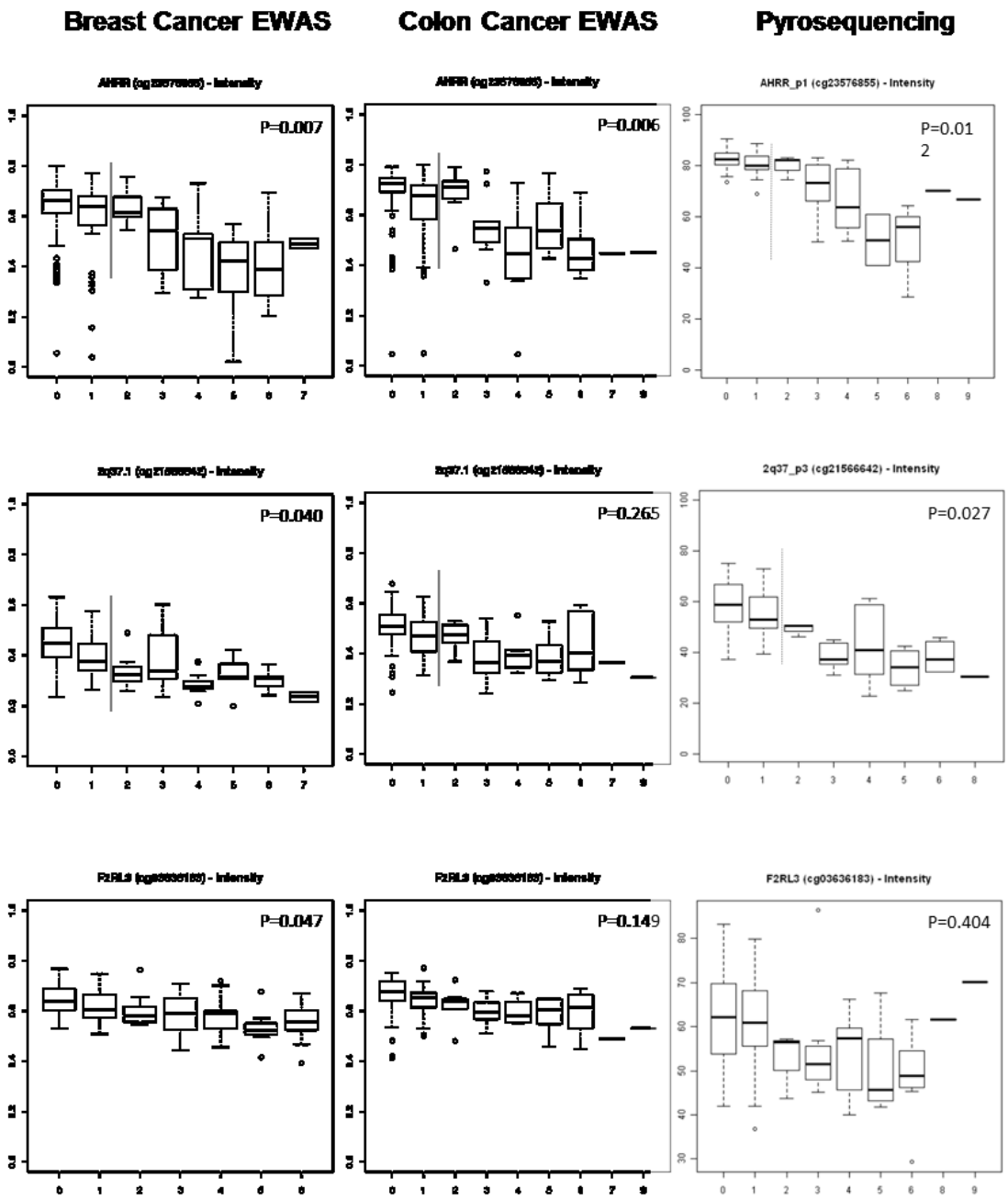


Figure 3.5: Smoking intensity is directly correlated with hypomethylation at *AHRR* and *2q37.1* in smokers. Methylation β-values are presented for never smokers (0), former smokers (1) and increasing intensities in current smokers from 2 (1-3 cigarettes per day), 3 (4-8), 4 (9-13), 5(14-18), 6(19-23), 7(24-28), 8(29-33), 9(>34). Y-axis represents methylation beta values (for EWAS) or methylation percentage (pyrosequencing) presented as box and whisker plots showing the median line, 25th-75th percentiles and whiskers showing the 95th percentile range for each probe. General linear regression adjusting for age and batch, in smokers only, was used to investigate the interaction between methylation and smoking intensity (p-value for each probe is shown within the plot). Never smokers and former smokers are plotted only for comparison. Reproduced with permission from Hum Mol Genet (Shenker et al., 2013a).

3.3.3 Human lung tissue methylation and expression analyses

Pyrosequencing assays were conducted on bisulphite-converted genomic DNA extracted from 27 human lung samples. We have investigated lung tissue in this case to assess the most relevant tissue to the initial exposure. A marked association was present between the smoking status and methylation levels in human lung tissue. As in peripheral circulating mononuclear cell DNA, the methylation levels at cg23576855 and cg21161138 in *AHRR* were significantly decreased in current smokers (P = 0.00327 and 0.00143, respectively) (Fig. 3.7A). The methylation values in lung tissue were also identical to those shown in peripheral blood mononuclear cells, which suggests that blood sampling could offer a useful surrogate for future biomarker studies of lung tissue methylation for this gene.

Expression analyses using RT-qPCR assays for *AHRR* in human lung samples from smokers versus non-smokers (n = 5 for each group) showed that expression was inversely correlated with methylation levels (R^2 = 0.157). *AHRR* mRNA levels increased by 5.7-fold (P = 0.0047) in lung tissue from current smokers compared with non-smokers (Fig. 3.7B).

3.3.4 Expression analyses in a mouse model of smoking

In a mouse model of smoke exposure developed by collaborators at Imperial College (Eltom et al., 2011), smoking significantly increased the levels of *Ahr* and *Cyp1a1* in a time-dependent manner compared with controls (Fig. 3.7C). *Ahrr* mRNA transcript levels were initially reduced by 2.6-fold after initial smoke exposure (3 d; P = 4.56×10^{26}), but increased by 1.7-fold after 28 days (P = 0.003). Methylation of the *Ahrr*

locus was not performed as this intragenic CpG island region is not conserved in the mouse genome.

3.3.5 Development of the methylation model for smoking

The association between methylation levels in four of the genomic loci and smoking status that was observed in the above-described studies was confirmed in the test set of 81 individuals (Table 3.6, Figure 3.6). Cotinine levels above the cut-off of 15 ng/ml (Florescu et al., 2009, Soo-Quee Koh and Choon-Huat Koh, 2007) were significantly associated with current smoking status ($p < 0.00025$, Table 3.6) with a high predictive value for current smokers compared to never smokers using ROC analysis (AUC = 0.97). Table 3.6 also indicates the AUC values based on the methylation values for each genomic locus in predicting former smoker status for the test set, in addition to that for cotinine: the AUC for AHRR_p1 (AUC = 0.71), 6p21 (AUC = 0.63), 2q37_p1 (AUC = 0.68) and 2q37_p3 (AUC = 0.66) individually had a greater ability to distinguish former from non-smokers than cotinine levels (AUC = 0.47).

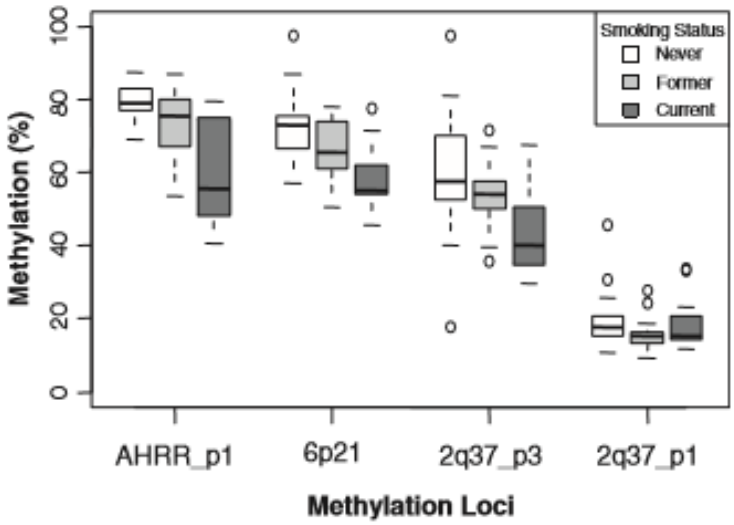


Figure 3.6: Box plots of distribution for methylation at four genomic loci for never, former, and current smokers (n = 81).

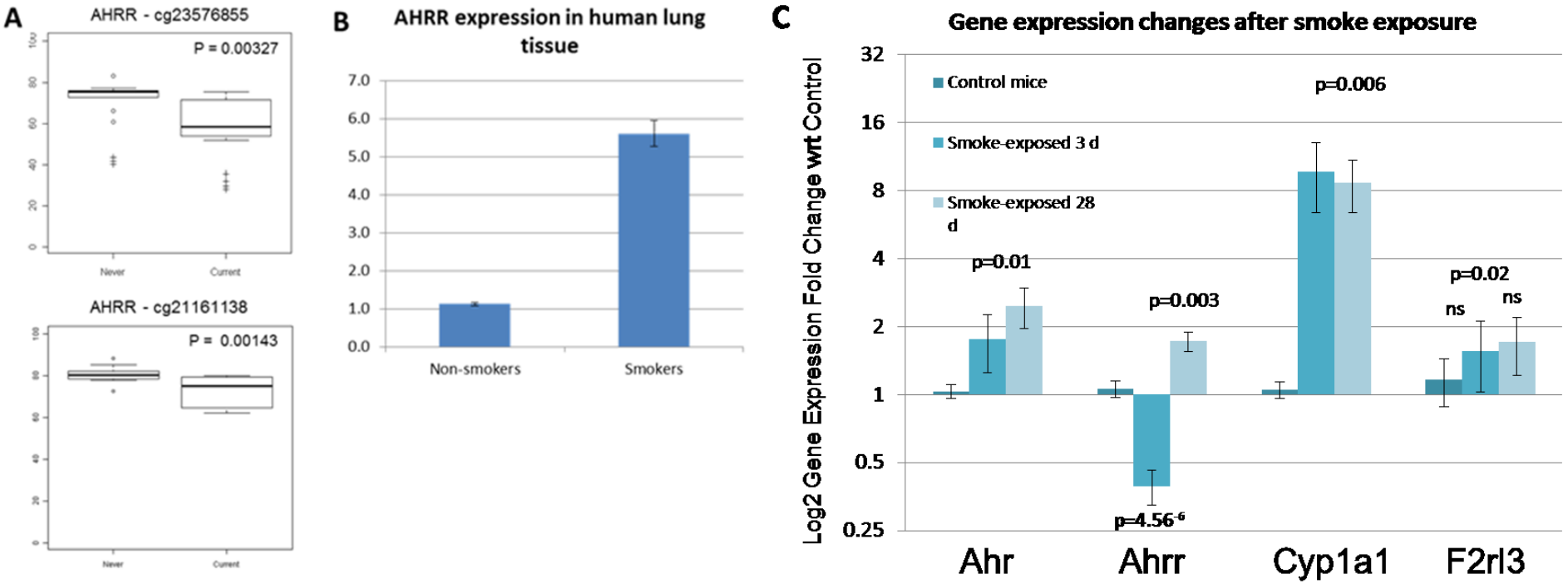


Figure 3.7: Smoking induces methylation and expression changes in human and mouse lung tissues. (A) Pyrosequencing data from two regions of interest, presented as boxplots, in human lung tissue samples from smokers (n = 14) versus never smokers (n = 13). For cg23576855 individuals heterozygous for the CG>CA polymorphism are marked as crosses. (B) AHRR mRNA expression data from the human lung tissue of smokers (n = 5) versus non-smokers (n = 5), showing a 5.7-fold increase in expression in smokers compared with non-smokers (upper panel) and correlation between methylation and expression in the same five non-smokers (grey circles) and five smokers (black triangles) (lower panel). (C) Mouse lung expression changes in *Ahrr* and *F2rl3,* in addition to two other genes from the aryl hydrocarbon receptor pathway, *Ahr* and *Cyp1a1*. As predicted, *Ahr* and *Cyp1a1* increase in response to cigarette smoke, and the expression of *Ahrr* initially decreases after 3 days, before increasing after 28 days. There was no significant change in the expression of F2RL3 after exposure to cigarette smoke. Reproduced with permission from Hum Mol Genet (Shenker et al., 2013a).

Table 3.6: Mean methylation percentages for the four cytosine-guanine nucleotides identified from the microarray study as differentially methylated in current smokers vs. former and non-smokers, in addition to the mean cotinine values (ng/ml) in each group of individuals. AUC values are shown for the individual markers to distinguish former smokers from never smokers in the test set.

		Smoking status						
Marker	Genomic locus (Illumina ID)	Never (n=33)	Former (n=30)	Current (n=18)	Sensitivity (%)	Specificity (%)	AUC never vs. former (95% CI)	PPV
Test set (n = 33 never smokers, 30 former smokers, 18 current smokers)								
AHRR_p1	Chr5:373,299 (cg23576855)	79.64	73.33	59.63	0.65	0.67	0.71 (0.56 – 0.86)	0.57
2q37_p1	Chr2.233,284,112 (cg06644428)	18.96	15.88	18.00	0.67	0.77	0.68 (0.56 – 0.85)	0.72
2q37_p3	Chr2.233,284,661 (cg21566642)	59.56	53.94	43.45	0.61	0.64	0.66 (0.46 – 0.81)	0.58
6p21.33	Chr6:30,720,080 (cg06126421)	73.03	66.09	58.45	0.63	0.65	0.63 (0.45 – 0.85)	0.52
Cotinine (ng/ml); mean		4.96	3.77	980.74	0.04	0.90	0.47 (0.32 – 0.63)	0.25
Methylation Index					0.69	0.90	0.82 (0.64-0.99)	0.92
Validation set (n=102 never smokers, 45 former smokers, 33 current smokers)								
Methylation Index					0.71	0.80	0.83 (0.70-0.96)	0.85

PPV: positive predictive value using optimum thresholds; 95% CI: 95% confidence intervals

Table 3.7: Cotinine levels (ng/ml) in the test sample set (n = 81).

	n	*Median*	*Minimum*	*Maximum*	*IQR*
Never Smoker	32	2.03	0.83	45.15	1.70-3.15
Former Smoker	25	2.04	0.57	24.47	1.59-3.41
Current Smoker	18	904.30	3.27	3181.89	583-1156

The methylation values of all four loci were combined into a single methylation index using the MI model = ($\beta_1 M_1$ * $\beta_2 M_2$ * $\beta_3 M_3$* $\beta_4 M_4$), where β represents the beta coefficient for the methylation locus in association with smoking status and M represents the methylation level of each locus as a percentage (equivalent to raw beta values from 450K methylation array data). For the purpose of this analysis we have not transformed methylation values to M-values as this did not alter the performance of the model. The AUC value for the combined methylation index of these four loci in differentiating never- from former smokers was 0.82 (95%CI = 0.64-0.99) in the test set and 0.83 (95%CI = 0.70-0.96) in the validation set (Figure 3.8) with a sensitivity of 69% and 71%, respectively.

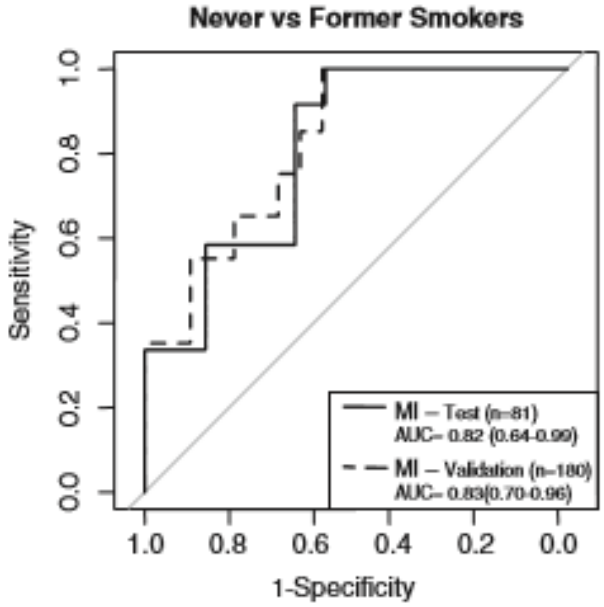


Figure 3.8: ROC for prediction of former smoking status based on MI.

Using the 450K methylation array data from the EPIC-Turin sample described above, we showed that the methylation index correlated strongly with the duration of smoking in former smokers (Pearson's correlation $R^2 = 0.47$, $p < 0.0001$, Figure 3.9A). Furthermore, using the current pyrosequencing data in the validation set, we showed that the methylation index also correlated strongly with the time since quitting in former smokers (Pearson's correlation $R = -0.51$, $p = 0.01$; Figure 3.9B), as well as duration of smoking ($R^2 = 0.63$, data not shown).

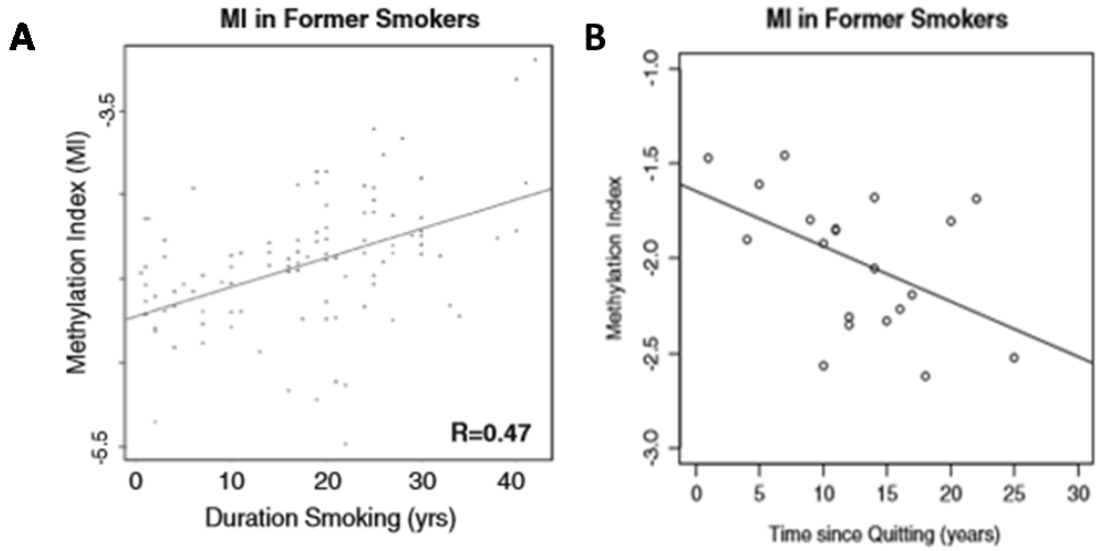


Figure 3.9: A) MI compared with duration of smoking in former smokers. B) MI versus time since quitting in former smokers.

3.4 DISCUSSION

3.4.1 General discussion

The findings from these three groups of individuals give strong evidence for the role of smoking in inducing changes in DNA methylation levels. The affected loci identified in intragenic regions are of particular interest with regards to the presence of gene expression changes in genes with altered methylation levels.

Previous evidence has suggested that intragenic methylation levels are correlated with expression levels, with highly expressed genes having high levels of intragenic methylation, and vice versa (Zilberman et al., 2007, Aran et al., 2011, Shenker and Flanagan, 2012). According to this rationale, we predicted that decreased levels of methylation in the *AHRR* gene in smokers would indicate lower levels of expression. In the smoke-exposed mice, this was indeed the case with short-term exposure (3 d), but the consequent increase in *Ahrr* expression after 28 d of smoke exposure suggest that other compensatory mechanisms of gene induction override the short-term decrease in expression marked by lower intragenic DNA methylation levels. In humans, the expression of *AHRR* in lung tissue mirrored the long term exposure in mice, however, the observed hypomethylation of this locus in lung tissue may be an indicator of past expression changes in this differentiated tissue type.

Toxic components of cigarette smoke enter the bloodstream via the alveolar capillary system, after which they could directly affect the epigenetic profile of circulating WBC. We hypothesize that the DNA methylation levels of *AHRR,* in particular, could be due to transcriptional repression via a negative feedback loop due to the induction of *AHR* pathway by dioxin-like compounds. However, we have to consider the possibility that other confounding factors may influence this

association. We did not adjust for alcohol and body mass index, which may be smoking-associated or methylation-associated confounding factors, principally because of the small numbers of subgroups in each population. Furthermore, we have assessed methylation levels in the genomic DNA of all circulating WBC. Smoking increases the circulating WBC count (Wannamethee et al., 2005). Although the effect of cigarette smoke on specific WBC populations have not been assessed in large-scale populations, it may be that subsets with different methylation levels are clonally expanded, which might affect the overall methylation result. This could be tested by separating WBC fractions and analysing methylation of these candidate loci in the different cell types or by adjusting computationally for cell types using the recently reported Houseman method (Houseman et al., 2012).

We, and subsequently others, have determined that a particular region of DNA methylation in the *AHRR* gene is strongly associated with smoking and can be a marker of past smoking exposure (Zudaire et al., 2008, Tsaprouni et al., 2014, Shenker et al., 2013a, Richmond et al., 2015, Philibert et al., 2013, Philibert et al., 2012, Novakovic et al., 2014, Nielsen et al., 2014, Lee et al., 2015, Kupers et al., 2015, Joubert et al., 2012). Reduced methylation at the CpG sites identified in AHRR appear to be consistent between ethnic populations (Elliott et al., 2014, Dogan et al., 2014), unlike methylation at 6p21.33, which showed marked differences between European current and never smokers in comparison to no differences in South Asian populations (Elliott et al., 2014). Further studies that incorporate a wide range of ethnicities with larger sample sizes will be required to tease out the ethnic-specific methylation states at these smoking-affected loci.

In a study conducted contemporaneously to ours, Monick et al. used 450K array to show that a single probe in the *AHRR* gene was associated with smoking status

(FDR P < 0.05) in cultured lymphoblast cell lines (Monick et al., 2012). Philibert *et al*. also using the same 450K platform, identified the same site as significant in both males and females in peripheral blood lymphocyte DNA (FDR correction, p < 0.05) (Philibert et al., 2013). Additionally, in a much larger study (n = 1062) on the 450K platform, Joubert *et al*. extended this by identifying 26 CpG sites associated with smoking in umbilical cord blood DNA (Bonferroni correction, p < 0.05) (Joubert et al., 2012). These effects appear to result from in utero exposure to maternal smoking (Joubert et al., 2014), which could be one mechanism by which the recognised effects of smoking (either maternal or environmental exposure, i.e., passive) are exerted on the neonate of lower birth weights (DiFranza and Lew, 1995, Kleinman and Madans, 1985, Kline et al., 1987, Hegaard et al., 2006), smaller head circumferences (Wang et al., 1997, Herrmann et al., 2008) and a higher risk of congenital abnormalities such as cleft palate (Khoury et al., 1989, Sabbagh et al., 2015).

AHRR (aryl hydrocarbon receptor repressor) is part of the aryl hydrocarbon pathway that metabolises cigarette smoke components, including the carcinogenic dioxins and dioxin-like compounds (Figure 3.10) (Evans et al., 2008, Chiba et al., 2011, Kawajiri and Fujii-Kuriyama, 2007). The aryl hydrocarbon receptor (AHR) is a crucial receptor in the pathway that metabolizes a range of biological compounds and synthetic environmental pollutants. Benzopyrene and dioxin-like compounds are highly toxic organic molecules that are released from various components of cigarettes during smoking and enter the circulation via the pulmonary vasculature. These compounds are metabolized by AHR (Shimizu et al., 2000), releasing further carcinogenic metabolites that have been implicated in lung cancer development (Chiba et al., 2011). AHR ligands (star in Figure 3.10), such as dioxins, dioxin-like

compounds, aromatic hydrocarbons and other synthetic chemicals, diffuse into the cytoplasm and bind to the AHR protein, which in turn changes conformation and binds to AHRNT (AHR nuclear transporter). This complex moves into the nucleus, where it binds to response elements in the promoters of a range of genes, including Cyp1A1, a critical component of the cytochrome p450 system (Badal and Delgoda, 2014, Go et al., 2015). The cytochrome p450 system is based in the liver, and is responsible for the metabolism of a range of exogenous compounds. These include ligands of AHR, which can produce carcinogenic metabolites (Rignall et al., 2013, Kuo et al., 2013), which exert effects that are independent of the tumorigenic effects of primary dioxin-AHR binding (Moennikes et al., 2004, Hosoya et al., 2008). However, the cytochrome p450 system also metabolises certain forms of cholesterol, bilirubin and the steroid hormone family, of which oestrogen is a member.

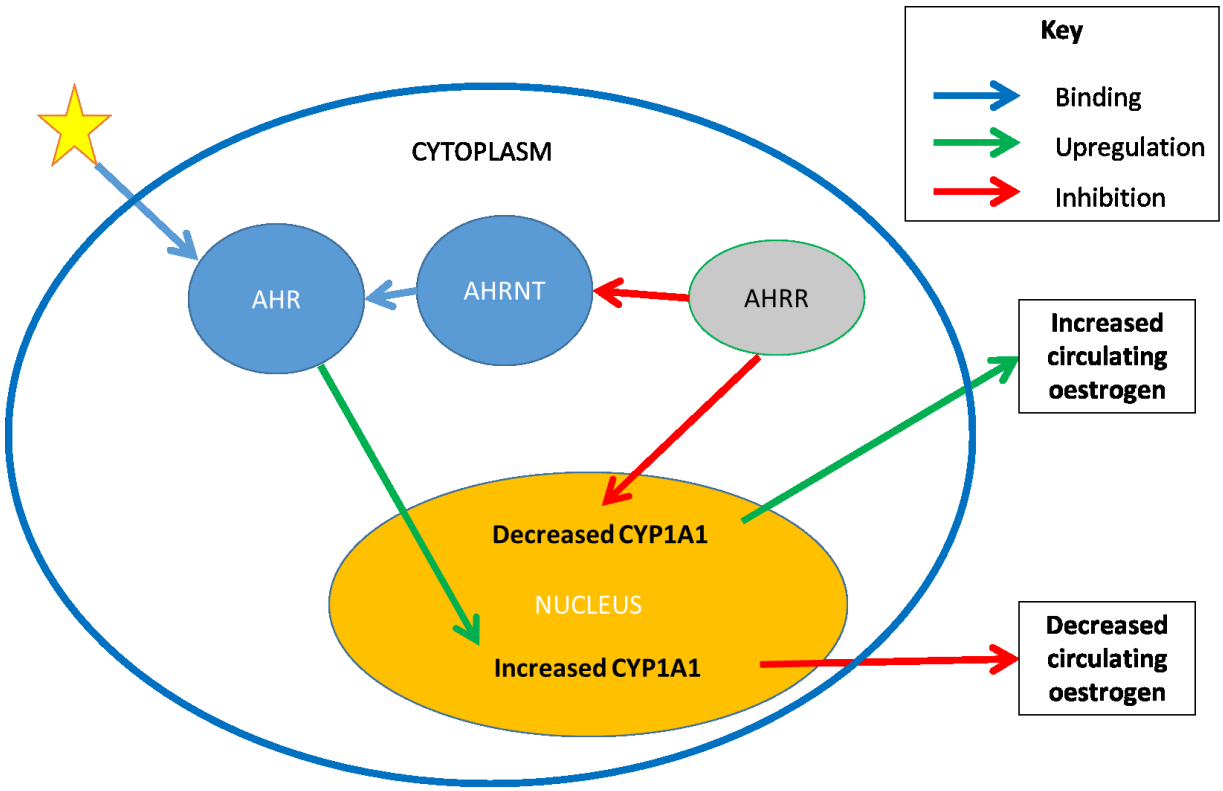


Figure 3.10: The AHR pathway for the metabolism of dioxins and dioxin-like compounds, found in cigarette smoke.

Although the link between *AHRR* (aryl hydrocarbon receptor repressor; chromosome position 5p15.33) demethylation and gene expression has not been proven by this study, the results from the mouse model of smoking indicate that gene expression of AHRR is increased in mice exposed to cigarette smoke (Fig. 3.7C). *AHRR* encodes for a class E basic helix-loop-helix protein (Baba et al., 2001), which is mainly found in the cytoplasm. It inhibits the translocation of the AHR-ligand complex into the nucleus by disrupting the binding of AHR to AHRNT (Evans et al., 2008), therefore reducing the induction of CYP1A1 (Hosoya et al., 2008), and reducing the activity of the cytochrome p450 system. The AHR-AHRNT heterodimer also reciprocally inhibits the transcription of *AHRR*. AHRR-AHRNT heterodimers compete with the AHR-AHRNT heterodimers to bind to promoter-associated xenobiotic response elements at numerous locations. The decreased expression of AHRR could, therefore, mediate increased levels of circulating oestrogen, which stimulates the growth of oestrogen-receptor positive breast cancers that constitute 70% of cases. The knockdown of *AHRR* has been shown to produce increased tumour cell invasiveness in a range of tissue types, including breast, colon, lung, and ovarian (Zudaire et al., 2008, Kanno et al., 2006), and the loss of *AHRR* expression has been shown in multiple solid tumours (Zudaire et al., 2008).

With regards to breast cancer, this model of modified AHRR function in the context of smoking-related demethylation could have both risk-promoting and risk reduction effects. With reduced cytochrome p450 activity, increased levels of circulating oestrogen would be expected, with associated increased risk for breast cancer. However, the metabolism of carcinogenic compounds in cigarette smoke would be reduced, yielding fewer toxic metabolites. Whether this effect is more marked in hepatocytes, which would be subjected to lower levels of toxins in

smokers accordingly, than in breast epithelial cells needs to be investigated in future in vitro studies.

The conflicting predictions in terms of risk modification or induction may, however, help to explain in part why smoking is not one of the strongest risk factors for breast cancer compared to its role in other cancers. To add to the complexity of this model, AHR is a major upregulator of multiple downstream genes, and therefore any disruption in the balance between AHR and AHRR expression could mean profound changes in cellular genome expression patterns.

Numerous other loci were associated with methylation changes in smokers across the 450K studies, but these are largely intergenic. The functions of these regions are currently unknown. Of particular interest was that differential methylation was observed between smokers and former or non-smokers across four consecutive probes located within 824 bp of an intergenic CpG island (chr2: 233,284,112 – 233,284,935; http://genome.ucsc.edu/, hg19; Fig. 3.11). This region maps to a DNase hypersensitivity site, indicating a possible regulatory region (Crawford et al., 2006). This locus is adjacent to two RNA polymerase II binding sites, and is upstream of several minus strand expressed sequence tags (ESTs) that have been identified in multiple human embryonic cell lines. BlastN analysis of the sequence indicates that this region is a proposed pseudogene of *ECEL1* (endothelin-converting enzyme-like 1; chr2: 233,344,537 to 233,351,464 bp, 60,425 bp downstream of 2q37), as it possesses a high level of sequence homology (>95%). *ECEL1,* a member of the M13 family of zinc-containing endopeptidases, is crucial for the development of neural components during embryonic lung development; knockouts of *ECEL1* result in embryonic lethality (Valdenaire and Schweizer, 2000). We hypothesize that the cluster of ESTs located at 2q37 may represent either a transcribed coding sequence

or a non-coding RNA that is normally repressed by a hypermethylated promoter. Further studies will be required to assess the functional nature of this genomic locus.

3.4.2 Methylation index of former smoking

Any biomarker of tobacco exposure should be a parameter that reflects the degree of exposure, including the intensity and duration of smoking (Florescu et al., 2009). The findings from our study give strong evidence for long-standing methylation changes as a result of smoking. Methylation is a relatively stable DNA modification (Tost, 2010). However, methylation changes may be reversible after the cessation of an exposure, although in the case of tobacco the timing of reversion is unknown.

In the majority of loci, smoking induced hypomethylation (loss of methylation) which has been observed in all similar studies to date. This finding was validated using an alternative method of bisulphite pyrosequencing for six loci. Regarding one of the top hits in the *AHRR* gene (e.g., cg05575921), the findings of Monick *et al*., Joubert *et al*., Philibert *et al.* and our own study have all shown a decrease in methylation, which suggests that the direction of association is consistent. A notable exception is CYP1A1 for which increased methylation in relation to maternal smoking in pregnancy was identified and replicated in an independent population (Joubert et al., 2012). The contrasting effects of maternal smoking during pregnancy on methylation at CpGs in AHRR and CYP1A1 are of interest because of the opposing function of these genes in the AhR pathway (Kawajiri and Fujii-Kuriyama, 2007). Numerous other loci have been associated with methylation changes in smokers including two that we previously identified, the intergenic loci 2q37 and 6p21; however, the functions of these are currently unknown.



Figure 3.11: UCSC genome browser image of genomic locus 2q37, which is demethylated in current and former smokers. This region maps to a cluster of DNaseI hypersensitivity regions (shown in grey shading), as well as a CpG island (http://genome.ucsc.edu/).

3.4.3 Longevity of methylation changes

We have shown that the epigenetic changes associated with smoking are detected in blood DNA in former smokers many years (median, 13 yr; IQR, 9-18 yr) after they have quit smoking (Figure 3.9B). More recent analysis suggests that methylation changes associated with smoking can be detected up to 40 years later (Guida et al., 2015). This methylation index has significant advantages over the current gold standard for the detection of smoking, the cotinine assay, which is used worldwide. Cotinine, a metabolite of nicotine, only has a half-life of 16 hours (Benowitz et al., 2009, Benowitz, 2008). As a result, cotinine does not distinguish between former smokers and those who have never smoked, and can validate only whether former smokers have actually quit smoking (Florescu et al., 2009). The identification of a persistent biomarker of tobacco exposure would not only be useful for molecular epidemiology, but would suggest a paradigm for quantification of other exposures that are also difficult to measure.

Given that the majority of white blood cell types have lifespans of ~30 days, this suggests that the exposure must also be affecting the haematopoietic stem and progenitor cells which perpetuate the epigenetic alterations in the daughter differentiated cells. Further evidence of long term perpetuation of methylation associated with exposures was observed in the Dutch famine study with IGF2 hypomethylation in exposed individuals detected 60 years later (Heijmans et al., 2008). If exposures such as smoking can increase an individual's risk of cancer, even in former smokers, then we hypothesise that those exposures throughout life must also affect the tissue specific stem and progenitor cells and potentially the cell of origin for the initial carcinogenic events. While we have measured this biomarker in blood DNA, which proves a suitable DNA source, other sources of DNA, such as

from buccal swabs, may be an equally accessible source of DNA for measuring this long-term biomarker of cigarette smoke exposure. Further investigation into DNA methylation of other cell types affected by tobacco smoking is clearly warranted.

3.4.4 Limitations

One of the limitations of this study is that it has been conducted in a single population of individuals enrolled in the EPIC cohort in Turin. However, multiple studies have been conducted by other groups that have further validated this smoking methylation signature in other populations. Another limitation in this analysis is that there is a gender imbalance in the test and validation sample sets. While not significantly different in smoking status distribution or discriminatory power, an analysis of larger cohorts may yet identify further gender-specific differences (Joubert et al., 2014). While we had too few individuals who were self-described as former smokers with high cotinine levels (>15 ng/mL, n=1) or current smokers with low cotinine levels (<15 ng/mL, n = 2), it would be interesting to assess the performance of the methylation index versus self-reported smoking status in a larger study.

From a technical point of view, DNA and RNA quality was solely assessed using the Nanodrop, which may have improved downstream analysis if more advanced methods were used, such as Picogreen assays and Agilent Bioanalyzer technology, respectively. Furthermore, the methylation model was limited by the fact that it was impossible to design pyrosequencing assays that worked with sufficient reliability for a number of loci identified from the microarray data, for a number of reasons. These included the presence of repetitive elements, a high density of CpG sites, and some

primers that generated excellent PCR bands but could not be pyrosequenced, despite all efforts. Ultimately, a time limit for their optimisation had to be imposed to enable the progression of the study, but an alternative quantitative technique to pyrosequencing should be used to validate the 450K array data for these CpG sites. If this is successful, these CpG sites can also be incorporated into a model with the aim of increasing its sensitivity and specificity.

Limited numbers of mouse and human lung tissue samples were available in this study, as a consequence of the ethical requirement to use the minimum number of animals possible (n = 5 in a single arm of the study), and the paucity and precious nature of the clinical tissue samples, respectively. It should also be noted that the lung tissues samples were obtained from patients who had smoking-related or non-smoking related lung disease, and therefore the gene expression and methylation profiles may not have been strictly normal.

3.4.5 Conclusions

Taken together, these studies report a strong link between tobacco use and a direct biological result on the epigenome, which is detectable in blood, and may impact on the cancer risk associated with smoking. The level of methylation was directly associated with smoking intensity and duration, however, the biological consequences of these epigenetic alterations will require further investigation. Indeed, it is unclear at present as to which component of cigarette smoke is responsible for the epigenetic changes, how these changes are effected, and whether these changes are limited to circulating blood cells and lung tissue. Our results regarding a key gene within the AHR pathway, *AHRR*, will have relevance

beyond tobacco smoke exposure given the key role played by this pathway in the metabolism of many environmental carcinogens.

Furthermore, the set of differentially methylated genomic loci dependent on tobacco exposure have been combined into a single DNA methylation index that can predict former smoking status with high levels of positive predictive value and sensitivity. This provides a direct molecular measure of prior exposure to tobacco that can be performed using pyrosequencing, and could be developed into a marketable assay as a result of its significant advantages over the currently available gold standard cotinine assay.

As yet, the specific factor and underlying mechanism responsible for the demethylation at specific loci identified in this and subsequent studies remain unidentified. With the advent of vaping, or e-cigarette use (Oh and Kacker, 2014), a natural experiment is developing that will enable us to assess whether the demethylation results from nicotine alone, or is effected by another component of cigarette smoke. Further work will be required to develop a cohort of blood samples from current vapers to assess whether they exhibit similar PBMC methylation changes, to understand whether this is a result of either nicotine or any of the other chemicals found in both vaping liquids and cigarettes.

These current data also show that epigenetic patterns detected in blood may provide molecular biomarkers of other exposures that are also difficult to quantify in epidemiological studies, but need to be validated in specific cell types of interest. Of particular interest to our group is the role of smoking in the aetiology of breast cancer. As described in the introduction of this thesis, smoking is a risk factor in breast cancer under specific circumstances (early age of smoking commencement

and smoking by pre-menopausal women). However, without an assessment of the direct effects of smoke on the methylation profiles and biology of breast epithelial cells, it is impossible to conclude conclusively whether smoking is a true risk factor for breast cancer.

CHAPTER 4

ASSESSING THE FEASIBILITY OF THE USE OF EPITHELIAL CELLS EXTRACTED FROM BREAST MILK FOR EPIGENOMICS

4.1 INTRODUCTION

Epigenetic signatures vary between cell types, most of which are inherited between mitotic cell cycles and generations (Reik, 2007, Hawkins et al., 2010, Cedar and Bergman, 2008). A significant volume of work is ongoing to characterise the inherent cell type variability in humans in terms of both chromatin modifications (Ernst et al., 2011, Barski et al., 2007, Heintzman et al., 2009, Kasowski et al., 2013, Flanagan et al., 2006) and methylation (Lister et al., 2009, Ziller et al., 2013), both in healthy and diseased tissue (Maurano et al., 2012). Data is currently being generated and amalgamated into accessible online databases, such as the Human Epigenome Atlas; data generation and analyses are overseen by the Roadmap Epigenomics Consortium (Kundaje et al., 2015), which ensure that data is produced from different cell types in a comparable manner. The aim is that the interplay between epigenetic modifications and transcriptional regulation can then be inferred (Yu et al., 2008a), particularly in the context of how histone modifications and methylation variability interact with regulatory elements (Consortium, 2012) in different tissues, and subsequently, how environmental exposures may interfere with normal cell functions (Zhu et al., 2013).

The consequent effects of epigenetic variability on gene expression largely determine different cell phenotypes, and explain the marked phenotypic differences between organisms despite having much of the same genetic sequence in common (Zeng et al., 2012, Weber et al., 2007, Molaro et al., 2011). Environmental exposures, such as alcohol or smoking, have been shown to perturb DNA methylation in consistent patterns between individuals across the genome (Christensen et al., 2010, Travis et al., 2010, Herceg and Paliwal, 2011, Flanagan et al., 2015, Monick et al., 2012, Philibert et al., 2013, Shenker et al., 2013a, Dogan et

al., 2014, Elliott et al., 2014, Nielsen et al., 2014, Tsaprouni et al., 2014, Lee et al., 2015), which may explain in part how these exposures act as risk factors for cancers.

However, the vast majority of work into these exposures and cancer risk markers has been conducted in blood cells, specifically mononuclear cells (PBMCs) (Flanagan et al., 2015, Flanagan et al., 2009, Brennan et al., 2012, Teschendorff et al., 2009, Zhang et al., 2011, Van De Voorde et al., 2012). The epigenetic signatures in normal blood cells and those in different tissues, in addition to those exposed to carcinogenic compounds over a lifetime, may bear little or no similarity to those in the cell types that develop cancer (Heijmans and Mill, 2012). In addition, the epigenetic signature from blood cells may be distorted by the different cell types within the PBMC population, including neutrophils, eosinophils, basophils and monocytes (Hodges et al., 2011).

In terms of breast cancer research, the most relevant cell type is the breast epithelial cell (BEC), as these are the principal cell of origin of most breast carcinomas (Ince et al., 2007, Santagata and Ince, 2014). By studying the normal BEC in the context of lifetime exposures and family history (Swerdlow et al., 2011), a convincing ‘methylation signature of risk’ could be developed (Michels et al., 2013). Unfortunately, to date, the acquired knowledge about breast epithelial cells is insufficient to achieve this aim. Most published methylation data regarding breast cancer risk has been derived from PBMCs, which are not appropriate surrogates for BECs. These cells have different exposures, functions, phenotypes, differentiation states and half-lives. However, given the importance of the healthy breast epithelial cell, surprisingly little is known about its normal biology, and even less about the lactating state of the terminally differentiated BEC. Previous efforts to collect normal

BECs have focussed on ductal lavage (Konstandiadou et al., 2012, Bentz, 2009, Sauter et al., 2010). However, this method is not suitable for large-scale epidemiology studies (Hornberger et al., 2015), as a consequence of the intrusive nature of the test and the low numbers of cells acquired in the aspirate.

Although normal breast epithelial cell lines are available, cell lines are notoriously difficult to study normal biology and these cells grow poorly on plastic, with low passage rates. There is also concern in the epigenetic community regarding the rapid loss of 5-hydroxymethylation and altered methylation levels in culture (Nestor et al., 2015). This controversy was encountered in the work described in Chapter 2 of this thesis. A large period of time during the first year of my own PhD was spent at the Institute of Cancer Research in Kensington, in a collaboration with the Royal Marsden Hospital breast surgeons. I tried to continue to develop a protocol to culture BECs from mammoplasty procedures – breast reductions of large 'healthy' breasts – as well as from BRCA carriers. Unfortunately, the cells did not grow well and the project was abandoned, leaving me with personal experience of the frustration of growing these cells.

4.1.1 Breast epithelial cells from milk for breast cancer research

A small number of previous studies have proposed using mammary epithelial cells from EBM, from as far back as 1998 (Thompson et al., 1998, Arcaro and Anderton, 2008). Breast milk contains high numbers of BECs, which are sloughed into the fluid during the process of expressing milk. Extracted BECs have been used in genotyping studies and in assessments of the effects of environmental carcinogens and genotoxins on breast tissue (Ambrosone et al., 2007, Martin et al., 2001, Martin et al., 1999, Martin et al., 2000, Yilmaz et al., 2012). However, although there is a

small body of literature that illustrates that BECs can be extracted and used to generate limited genetic and epigenetic data, no systematic studies have been conducted to date that have characterised these cells with regards to lifetime exposures and breast cancer risk, and there is only a scant amount of available epigenetic data (Sturgeon et al., 2014, Wong et al., 2010). Therefore, there is a considerable need for a large cohort study to collect and investigate the normal biology of BECs and their potential for breast cancer risk analyses.

4.1.2 Aims

In the most recent Gap Analysis published by Breast Cancer Campaign (Eccles et al., 2013), 10 recommendations were made for areas of research focus in order to overcome critical shortcomings in breast cancer biology, treatment and risk prediction. Of these, this study aimed to contribute to gaps (1) and (10), reproduced below:

"Gap 1: Understanding the functions and contextual interactions of genetic and epigenetic changes in normal breast development and during malignant transformation"

"Gap 10: A continuing need for clinical material for translational research derived from normal breast, blood, primary, relapsed, metastatic and drug-resistant cancers with expert bioinformatics support to maximise its utility"

We propose that although lactating BECs are a highly specific cell type, they may be the most appropriate resource to advance the current understanding the oestrogen receptor (ER)-responsive cell biology, a function shared with the large majority of breast cancers (luminal type). We therefore aimed to develop a

prospective cohort of collected exfoliated breast epithelial cells from expressed breast milk.

The specific aims of this study were as follows: i) to develop standard operating protocols for the recruitment, quality control and cell extraction from both fresh and frozen milk samples; ii) to investigate whether these cells are viable for use in gene expression, DNA methylation and chromatin assays; iii) to generate pilot data from these cells with regards to candidate epigenetic risk biomarkers in order to support the expansion to a large cohort study to increase the recruitment and collection of EBM samples nationwide; and iv) to explore the intragenic methylation state of ESR1 in breast epithelial cells from healthy young women and compare these results to cell line and tumour biopsy material in Chapter 2.

4.2 METHODS

4.2.1 Breast milk collection

During the pilot stage of this project, we collected 20-30 mL aliquots of breast milk, with up to 20 samples from each donor, which had been stored at the Queen Charlotte and Chelsea (QCCH) Milk Bank from 150 individuals. The Milk Bank collects donated milk in 250 mL containers, but aliquoted two 25-mL samples from each donation. One of these is sent for microbiological assessment at Charing Cross Hospital, and the other is kept at -20°C in case the first sample is lost. It was these samples that were analysed in this current study, as the milk would otherwise have been discarded once the microbiology results were returned. Ethics approval for this part of the study was obtained from the Hammersmith Hospital Human Imperial NHS Tissue Bank access committee (project reference: R13020).

Breast milk samples were processed in the Milk Bank using standard operating procedures based on NICE guidelines (nice.org.uk/cg93) (NICE, 2010). Samples were stored in standard fluid collection white-capped tubes, labelled on the top and side with a consecutive 5-digit Milk Bank sample identifier. Each donor was given a unique donor ID, against which the sample numbers were recorded. Demographic and sample data was collected on a Microsoft Excel spreadsheet, which was anonymised before leaving the Milk Bank according to ethical guidelines for the use of patient data. Demographic data recorded by the Milk Bank includes age, parity, smoking and alcohol history, nutrition (vegetarian, vegan, allergies) details of the individual's milk donation history, details of the birth (date, mode and place), medical history, medication use (prescription and non-prescription) and details of their home freezer and breast pump provision. Sample IDs were updated

on receipt by the Milk Bank staff to standardise and streamline record keeping. All participants in the study provided specific consent for the use of their milk for research purposes prior to the removal of the samples from the Milk Bank. Samples were transferred within an igloo storage box with ice packs to reduce any chance of defrosting between the Milk Bank and our laboratory freezer. Overall sample/ donor recruitment numbers were collated each month to assess the pattern of donor recruitment.

4.2.2 Cell pellet generation

Samples were defrosted on the bench at room temperature or in a 20°C water bath. An initial comparison was conducted to check that there was no difference between DNA yields and quality from the two methods of defrosting, and this was indeed the case (n = 5 samples, data not shown). The water bath enables the time to defrost to be reduced from 1.5-2 hours to 20 minutes, greatly speeding up the processing time.

Samples were poured into labelled 50-mL Falcon tubes (Corning, NY, USA) and diluted 1:1 in phosphate buffered saline (PBS) to reduce the viscosity of the milk, in accordance with previously described techniques (Martin et al., 2000, Wise et al., 2009, Browne et al., 2011). Tubes were then spun at 2200 rpm for 20 – 25 min at room temperature. Supernatants were decanted into clean labelled Falcon tubes and stored at -20°C for future use, or discarded into appropriate biowaste containers according to the risk assessment prepared in advance of this study. The lipid layer formed at the top of the sample; this was relatively thick (≈4-8 mm), and firmly adherent to the side of the tube. Falcons were therefore kept horizontal to prevent unnecessary contamination of the cell pellet by lipid falling back down to the bottom of the tube. Cell pellets were then resuspended in 1 mL PBS, pipetted into labelled

1.5-mL Eppendorf tubes, and spun again for 10 min at 6000 rpm at room temperature. The supernatant was discarded and the cell pellet was either stored at -80°C until use (recorded appropriately on the sample database), or processed immediately. If the tube was grossly stained with lipid residue, a second spin cycle was performed.

4.2.3 Cell characterisation

On resuspension in 1 mL PBS, 20 µl of the cell suspension was removed and cells were counted using a Countess Automated Cell Counter (Invitrogen, Carlsbad, CA, USA). The first 20 samples were also counted under light microscopy using a Neubauer Chamber cell counter, and results were compared. As the results of the automated cell counter were accurate, this was used thereafter. Cells were also examined under light microscopy (x20, x200), and found to be spherical and intact in the main. Staining with Trypan blue was conducted to assess cell viability in fresh milk samples, and this was correlated with the viable percentages recorded by the automated cell counter.

Sixty-six samples underwent flow cytometry for the assessment of an epithelial cell marker and oestrogen receptors. Samples were analysed using flow cytometry with a FITC-labelled antibody against epithelial membrane antigen (EMA; CD227, Sigma Aldrich; n = 68) and a PE-labelled antibody against intracellular ESR1 (Sigma Aldrich; n = 6), according to standard techniques in accordance with the guidelines of the manufacturer. Briefly, 250 mL cell suspension was removed from the initial cell pellet suspension described above before the first PBS wash. For EMA staining, a surface membrane receptor, 1 mL FACS buffer (PBS + 0.5% bovine serum antigen, BSA) was added and cells were spun at 1000 rpm for 5 min.

Supernatant was removed, and the primary antibody was prepared in FACS buffer, and 80 µl antibody solution was added to each sample. Samples were incubated with the primary antibody (10 µl) for 30 min at room temperature in the dark. The primary antibody for EMA was already conjugated to FITC (green fluorescence), and therefore no secondary antibody step was required. Samples were then washed three times in 1 mL FACS buffer and spun at 1000 rpm for 5 min. At this stage, the samples for the assessment of EMA staining were resuspended in 300 mL FACS buffer. Five samples were prepared with CD10 antibody (Sigma Aldrich, FITC conjugated), which is specific for stem cells, but this was abandoned as a result of technical difficulties.

For samples that were assessed for both EMA and ESR1 expression (n = 6), an initial lysis step was required as ESR1 is predominantly an intracellular protein. The primary antibody against ESR1 (5 µl) was prepared in 0.01% Triton X to permeabilise the cell membrane and samples were incubated for 30 min at room temperature. Appropriate controls were prepared using 50 mL of the cell suspension (1×10^5 cells), using both primary antibodies only + Triton X, no antibodies + Triton X and double-staining + Triton X, all compared against untreated cells (Figure 4.1). For the secondary antibody (PE, phycoerythrin: red fluorescence), 80 µl antibody solution was added to each sample and samples incubated for 45 min at room temperature before being washed as described after the primary stain. All cell samples were resuspended in 300 µl FACS buffer.

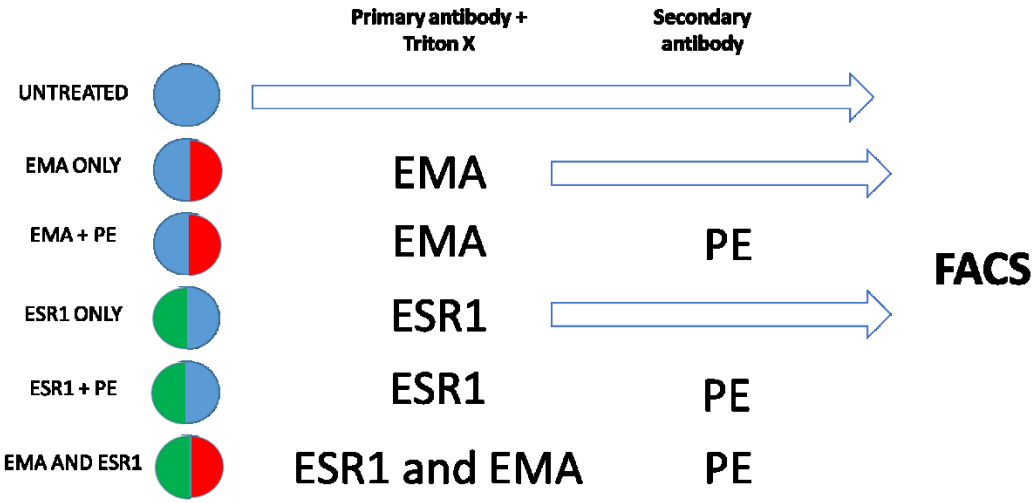


Figure 4.1: Flow diagram of the flow cytometry protocol, showing the controls and antibody steps used.

Samples and controls were then run on a FACScalibur flow cytometer (BD Biosciences, San Jose, CA, USA) and data were analysed using the following gates: FSC (forward side scatter): E-1, 5.24, linear; SSC (side scatter): 230, 6.06, linear; FL1: 412, logarithmic; FL2: 396, logarithmic. Cells were counted until 10,000 events were recorded.

4.2.4 DNA extraction and bisulphite conversion

DNA was extracted using a phenol:chloroform technique, with duplicate phenol and chloroform steps to optimise yields and reduce phenol contamination, respectively. Briefly, each cell pellet was resuspended in 360 µl Levi buffer (stock: 100 µl 10 mM Tris-HCl (pH 8.3), 500 µl 50 mM KCl, 25 µl 2.5 mM $MgCl_2$, made up to 10 mL with 9.4 mL H_2O) .With the sample in a fresh Eppendorf, 20 µl fresh Tween (10% Tween 20, diluted in distilled H_2O) and 20 µl proteinase K were added. Each sample was briefly vortexed and centrifuged for 10 s on maximum, and wrapped with Parafilm™ to reduce losses from evaporation. Samples were incubated for 2-3 hours at 55C.

Longer incubations were initially tested, but did not increase DNA yields. Proteinase K activity was terminated by incubating the sample at 95C for 15 min, and samples were adjusted to 1 mM EDTA by the addition of 16.7 µl 25 mM EDTA to each 400 µl. Samples were then placed into labelled phase-lock tubes that had been prepared by centrifugation on maximum speed for 30 s. In the fume hood, 400 µl phenol:chloroform was added, and samples were mixed by inversion for 1 min. Samples were spun at 13,000 rpm for 5 min, and the phenol incubation was repeated in the same phase-lock tube, mixed and spun again. The sample above the gel was then pipetted into a new phase-lock tube, and 400 µl chloroform was added. Samples were inverted for 1 min, spun at 13,000 rpm for 5 min and then added to a third phase-lock tube for the repeat chloroform step. The repeated phenol and chloroform steps increased DNA yield by 4-fold in a trial of three DNA extractions vs. a single phenol and chloroform steps (data not shown) by increasing the yield and reducing phenol contamination, respectively. Finally, the clear upper phase was added to a labelled screw-top tube, to which 1,200 µl ice-cold ethanol, 200 µl 3M sodium acetate and 4 µl 20 µg/µl glycogen were added. Tubes were frozen at -20°C overnight, and then spun at 4°C for 30 min. The supernatant was discarded, and DNA pellets resuspended in 1 mL ice-cold ethanol. Samples were spun again at maximum speed for 20 min at 4°C, the supernatants were discarded, and pellets were air-dried before being resuspended in 20-50 µl RNAse-free water. DNA quality and quantity was assessed by the Nanodrop Spectrophotometer (BioRad, UK) and the Picogreen assay.

For the Picogreen assay, each DNA sample was examined in triplicate against a standard curve of standard solutions ranging from 0–20 ng/µl. Briefly, 2 mL of standards and each DNA sample was pipetted into a black flat bottomed plate

(Costar 3915; Corning, USA). TE (1x, 98 µl) was added to each well, followed by 100 µl Picogreen reagent (diluted 1:200 in TE) made up immediately before the assay and stored in the dark. The plate was incubated for 2-5 min at room temperature in the dark, and fluorescence was then read using a plate reader. For analysis, triplicate values were averaged, and the average background value from the blank control was subtracted. A standard curve was then plotted against the fluorescence, and the line of best fit was centred through 0.0. From the equation of the standard curve line, the concentration of each DNA sample could be calculated (Figure 4.6).

DNA was bisulphite-converted (500 ng) using the EZ-96 Methylation-Gold™ kit (Zymo) according to the protocol given by the manufacturer, and samples from 36 donors were used for hybridisation onto the Infinium HumanMethylation 450 BeadChip array, using the Illumina Infinium HD methylation protocol (conducted by UCL Genomics). Pyrosequencing was subsequently performed for the ESR1 regions described above on 250 ng bisulphite-converted DNA samples (n = 53) to validate the ESR1 regional methylation.

4.2.5 Genotype analyses

The Illumina OncoArray is a custom-made SNP-based array that includes 533,631 single nucleotide polymorphisms that are associated with risk for a range of cancers, which has been used for research since October 2013. The OncoArray Network brings together a range of scientific consortia investigating breast, ovarian, prostate, colorectal and lung cancers. The SNPs on the array include 260,000 common variants across the genome, in addition to all of the risk markers identified in gnome-wide association studies (GWAS) and fine mapping studies, amongst others, for these five cancers.

This current project was approved by the OncoArray Steering Committee as part of an application entitled, “eQTL and meQTL analysis of breast cancer risk alleles in breast epithelial cells”, written by Dr James Flanagan. Aliquots of genomic DNA (2 µg) from 80 donors were also plated onto a 96-well plate (#Cat MSP9601, Microseal PCR Plate, BioRad), diluted in RNase-free water, with the requisite water controls for analysis by the OncoArray platform. The plate was sealed with a MicroAmp Clear Adhesive Film (Cat# 4306311, Applied Biosytems), and couriered on dry ice to Genome Quebec in Canada. All samples were >100 ng/µl according to the Nanodrop, but Picogreen assays showed substantially lower levels of double-stranded genomic DNA, with most samples well below the required cut-off of 50 ng/µl. On receipt in Quebec, the plate underwent further QC checks based on a Picogreen assay and over 75% of the plate failed. The decision was made to go ahead, based upon the fact that the DNA from BECS cells would be degraded, and 98% of the samples passed the array QC.

Five SNPs on the OncoArray were validated with the use of the PyroMark genotyping program on the pyrosequencer in 27 of the same donor samples. The SNPs analysed and associated genes were rs10759243 (*RAD23B*), rs2236007 (*PAX9*), rs2981579 (*FGFR2*), rs614367 (*CCND1*) and rs3757318 (*ESR1*) (Michailidou et al., 2015 individuals identifies 15 new susceptibility loci for breast cancer). Although the latter SNP was actually within another gene, it was located in the same linkage block as *ESR1*, and therefore associated with that gene (personal communication with Michailidou).

This genotyping project was largely undertaken by a Masters student, Isobel Steer, who I co-supervised throughout. This role included selecting the SNPS,

designing the primers, and teaching her how run PCR assays, gels and genotyping assays. I also provided support to her with data analysis and thesis writing.

Table 4.1: Primer sequences used for the five SNP PCR assays and genotyping reactions.

Gene	SNP	Primer	Sequence
ESR1	rs3757318	F	AAGGGGTAGGCAGGGAACTG
		R	CCTCTTTCTCCCAAGCTTACAGG
		S	CACATATGGGTCAGAGTC
CCND1	rs614367	F	CCTCTTGGCTTCTCTGCAACT
		R	TGTTTGGGGCCTAAAGAGATG
		S	GGGCCTAAAGAGATGTAAT
RAD23B	rs10759243	F	GGGTAAACACAGCAAAATGTAAGC
		R	TGCCCAGCCCCTTTTTATT
		S	CCCTTTTTATTTTCTTCG
PAX9	rs2236007	F	TCAGAGGCTGGGCGTGTG
		R	GGGACCTCCGCGAGAGTG
		S	TGTGGATGTGCAGCG
FGFR2	rs2981579	F	CCGCAGATGCCACTCAAGT
		R	TGGTGTGACTCCCTTCATCG
		S	TTTTGGATACGACCTCT

F: Forward; R: reverse; S: sequencing.

4.2.6 RNA extraction and analysis

RNA was found to be highly fragmented from frozen milk samples according to assessment by the Bioanalyser 2100 (Agilent Technologies, Santa Clara, CA, USA) using standard techniques, and was unsuitable for further use. Furthermore, the kit-based systems for RNA extraction were unsuitable for use with breast milk samples. Therefore, as a pilot study, a small number of fresh breast milk samples were collected from donors. Donors were asked to express 25 mL of milk 30 min prior to the arrival of the SERV driver. These samples were then placed in a separate plastic

bag on top of the frozen milk samples and arrived at the Milk Bank chilled but not frozen.

RNA was extracted from fresh samples using a standard TRIzol® technique. Briefly, the newly extracted pellet from each fresh breast milk sample was resuspended in 250 µl PBS. Samples were not washed again to avoid disrupting the RNA even more. Cells were lysed in 750 µl chilled TRIzol® in the fume hood (1 mL TRIzol® per 5-10 x 10^6 cells), incubated for 5 min at room temperature and spun at 4°C for 10 min at 12,000 rpm to remove lipid contamination. The liquid phase was removed into a new Eppendorf and 150 µl chloroform was added (0.2 mL per 1 mL TRIzol®). The tube was shaken vigorously for 15 s and incubated at room temperature for 3 min before being spun at 4°C at 12,000 rpm. The aqueous phase was pipetted into a fresh 2-ml labelled screw-top tube, and 500 µl isopropanolol (0.5 ml per 1 ml TRIzol®) was added. Samples were incubated at room temperature for 10 min before a further centrifugation step at 12,000 rpm for 10 min at 4°C. The supernatant was discarded and the RNA pellet washed twice with 1 mL 75% ice-cold ethanol, with spins for 5 min at 7500 rpm at 4°C after each wash. The RNA pellet was air-dried on the bench for 5-10 min and then dissolved in 30 µl RNAse-free water, before being incubated at 55°C for 10 min.

The RNA concentration was assessed on the Nanodrop and quality was assessed by the Agilent Bioanalyser RNA 6000 Nano Assay (Agilent Technologies, Santa Clara, CA, USA), according to the following protocol. With the RNA Nano reagents at room temperature, the RNA Nano Chip was selected and the electrodes decontaminated using standard techniques. All RNA samples were <500 ng/µl according to the Nanodrop, and therefore did not need to be diluted. The gel-dye mix was prepared by vortexing the RNA dye concentrate for 10 s with a brief spin; 0.5 µl

RNA dye concentrate was then added to 32 µl of the filtered gel aliquot, and vortexed. After the gel and dye were confirmed visually to be thoroughly mixed, the tube was spun at 13,000 g for10 min. RNA samples were then denatured and placed on ice with the RNA 6000 Ladder. With the chip prepared with 9 µl gel-dye mix, 5 µl RNA 6000 Marker and 1 µl RNA 6000 Ladder were pipetted into the empty wells (2 wells had gel-dye mix only as controls). Then, 1 µl each sample (n = 12; 11 fresh milk samples, 1 duplicate extracted from a frozen sample from the same donor) was pipetted into their assigned wells. The chip was vortexed for 1 min, and after bubbles were removed, the chip was placed in the Bioanalyser. The chip was then analysed using the 'Eukaryote Total RNA Nano Series II' programme.

cDNA was prepared from each fresh sample (n = 11) according to the techniques described in Chapter 2. An RT-qPCR assay for *ESR1* mRNA was performed on cDNA made from the 11 RNA samples from fresh EBM, normalised against GAPDH and using MCF7 and MDA-MB-231 cell lines as positive and negative controls for ESR1 expression, respectively, along with a negative reverse transcriptase sample. Each sample was analysed in triplicate.

4.2.7 Statistical analysis

All experiments were performed in triplicate unless otherwise stated. The mean ± standard deviation (SD) was calculated from each triplicate repeat of the pyrosequencing and RT-qPCR experiments. The mean ± SD were calculated after each replicate, and the standard error of the mean (SEx) was then calculated. Parametric data were compared using paired t-tests. Nonparametric data, including average methylation levels across the gene body, were compared using unpaired

Wilcoxon signed rank tests. All statistical tests were two-sided and performed using Microsoft Excel (Microsoft, USA).

The methylation scores from samples on all three chips that underwent the Illumina 450K arrays were processed using standard quality control measures, and normalised (colour correction) and batch adjusted using COMBAT, resulting in beta methylation values according to the fluorescent intensity ratio from 0 (unmethylated) to 1 (completely methylated). R was also used to analyse the probes related to ESR1. The R code for this analysis is shown in Appendix 7.

Observed versus expected genotype values were compared using the Chi-squared test, using data from the NCBI Short Genetic Variation database (http://www.ncbi.nlm.nih.gov/projects/SNP) as the evidence for expected genotypes within a European population. P-values were taken as significant if <0.05.

4.3 RESULTS

4.3.1 Establishment of the QCCH Milk Bank collaboration

Over the course of the pilot study period from August 2014 to July 2015, 158 donors consented to provide part of their donation samples or research purposes, and each signed the appropriate Queen Charlotte and Chelsea donor consent form. The samples used were the duplicates for microbiological screening, kept by the Milk Bank in case the first sample was lost; these would otherwise have been discarded. Of these 158, eight donors were excluded from any analysis: one subsequently removed their consent by telephone; seven had incomplete demographic data records. The recruitment of donors and sample numbers is shown in Figure 4.2.



Figure 4.2: A) Number of donors recruited monthly throughout the pilot study (August 2014 – August 2015) from a single Milk Bank; B) number of samples (20-25 ml) collected throughout the sample period. The study was paused for a month in May while data analysis was carried out. As this showed viable results, collection recommenced in June 2015.

Demographic data was collected via an anonymised questionnaire, and showed that their median age was 33.9 years (interquartile range, IQR: 31.5-37.0 years). Ethnicity data was provided by 125 donors, of which there were 67 white British, 33 white other, 24 Asian and 1 African. The median duration of breastfeeding at the time of sample donation was 153 days (IQR: 94 – 223 days). Women could only donate if they passed stringent inclusion criteria; accordingly, current smokers were not permitted to become milk donors because of the risk of introducing toxic compounds to the premature infant recipient. However, 28 donors in the cohort were former smokers, having quit between 1998 and 2013 (median time from quitting, 4 years). These donors were therefore a useful cohort for analysis of the smoking methylation index developed in Chapter 3, as the direct effect of cigarette smoking on breast epithelial cells has not been shown previously (see section 4.3.8).

4.3.2 Cell pellet generation and cell characterisation

After an initial intensive 2-month period developing and optimising the most successful centrifugation and wash protocol, samples could be processed *en masse.* With experience, batches of 20 (the maximum number of Falcon tubes that could fit in the centrifuge at any one time) could be processed, with three to four batches processed at any one time. Given that with optimised DNA extraction techniques, only one 25-ml sample tube was required from any one donor, the numbers and speed at which donor samples could be processed increased rapidly.

Sixty samples underwent flow cytometry analysis for EMA, and showed high rates of positivity with a median of 97.8% (range, 85.3%-99.4%). The single outlier with 85.3 epithelial cells came from milk from a donor within the first postnatal week, when the ratio of neutrophils and monocytes to epithelial cells is much higher than

established milk. Flow cytometric analyses in a small number of EBM samples (n = 6) was conducted to assess the presence of intracellular ESR1, and the median percentage of positive cells was 90.5% (range, 84.4%-93.3%; Fig. 4.3).

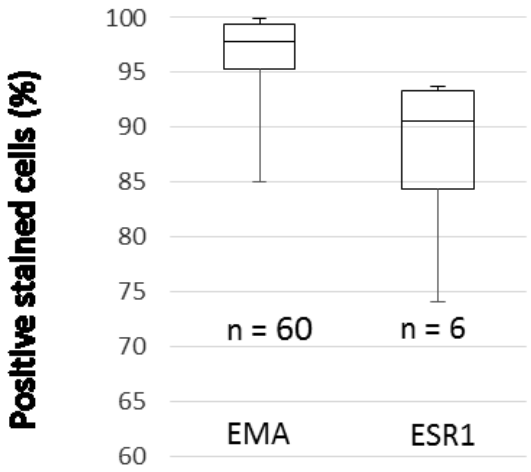


Figure 4.3: Flow cytometry results for epithelial membrane antigen (EMA; median positive staining, 97.8%; 95.3%-99.4%; n = 68) and intracellular ESR1 (median, 90.5%; IQR, 84.4%-93.3%; n = 6).

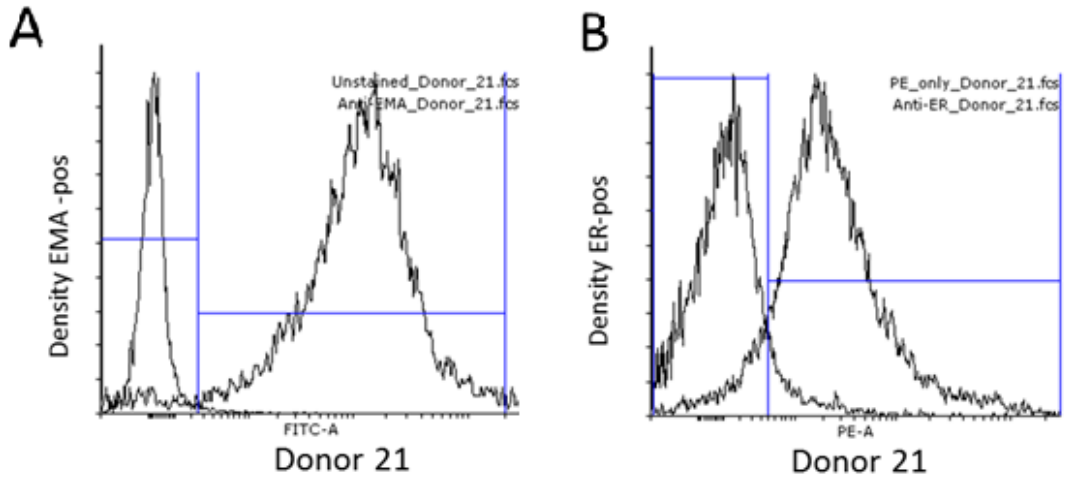


Figure 4.4: A) Flow cytometry for epithelial membrane antigen (EMA) from a single donor (#21) compared to negative control unstained cells; B) cells from the same donor stained with anti-ESR1 primary antibody + PE secondary antibody vs. PE-stained control cells. A marked shift is shown in both, showing the strong levels of expression of the epithelial cell marker and ESR1 in this cell type. Images representative of n = 68 and n = 6 for each antigen, respectively.

4.3.3 DNA extraction and quality control

Cell counts were performed for the first 68 EBM donor samples to gain an understanding of the range of cell numbers in each sample. For each 25 ml milk sample, cell counts ranged from 4.0 x 10^3 to 8.0 x 10^6. Interestingly, there was no correlation between the DNA yields from the phenol:chloroform technique and cell numbers.

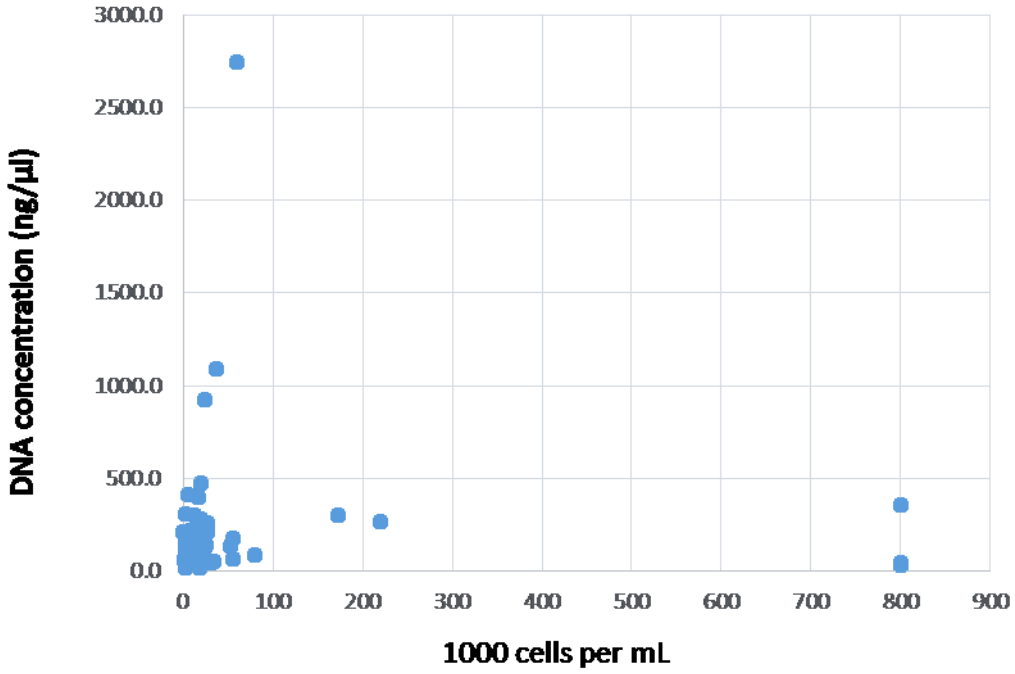


Figure 4.5: Scattergram to show that there was no correlation between cell counts and DNA concentration).

For pyrosequencing and 450K analyses, a 260:280 ratio as recorded by the Nanodrop should be at least >1.8. Some of the early samples fell below this ratio, reflecting technical difficulties that were largely overcome by the introduction of the second phenol and chloroform steps, and inherent sample issues. The results of the Picogreen assays appeared to be significantly lower, suggesting that the reasonably high DNA concentrations recorded by the Nanodrop were not replicated.

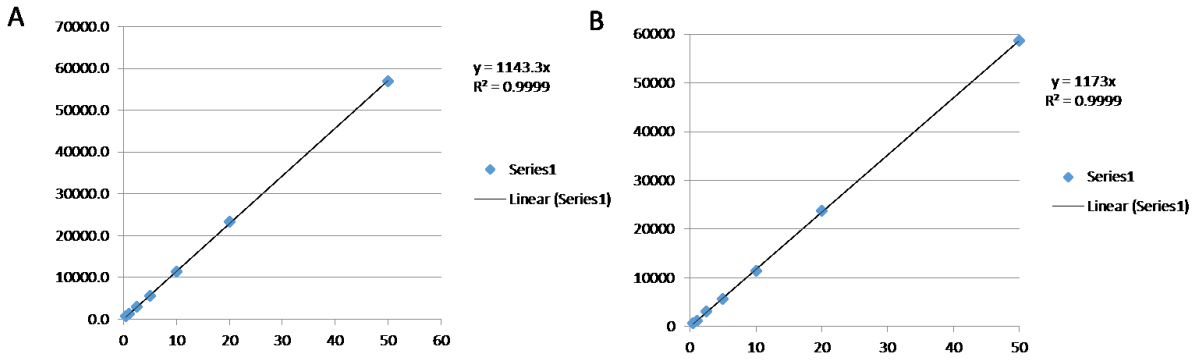


Figure 4.6: A, B) Standard curve generated from two Picogreen assays of DNA standards (conc. 0 – 20 ng/µl), with averages generated from triplicate repeats. A and B relate to the two different batches (A = donor IDs 1-4, B = donor IDs 25-49).

Table 4.2: DNA concentrations for 49 samples from Picogreen assay compared with Nanodrop concentrations. The final Picogreen DNA concentration was calculated by dividing the average OD value by the coefficient generated from the standard curves in Fig. 4.6. The two batches are shown in different grey shading (A) or no shading (B).

Donor/ sample ID	Nanodrop conc	Average Picogreen value	1:10 dilution	ng/ul
1	*212.9*	11994.0	10.5	104.9
2	*24.1*	15419.7	13.5	134.9
3	*29.3*	11880.7	10.4	103.9
4	*23.4*	5454.3	4.8	47.7
5	*133.2*	5760.7	5.0	50.4
6	*125.0*	10500.0	9.2	91.8
8	*93.6*	2219.3	1.9	19.4
10	*66.9*	3246.3	2.8	28.4
11	*41.5*	1762.0	1.5	15.4
12	*59.9*	1611.3	1.4	14.1
13	*114.8*	9439.3	8.3	82.6
14	*42.2*	8902.7	7.8	77.9
15	*113.9*	1449.3	1.3	12.7
16	*132.0*	1507.0	1.3	13.2
17	*30.1*	1753.7	1.5	15.3
18	*129.6*	627.3	0.5	5.5
19	*94.2*	7325.7	6.4	64.1
20	*234.7*	7244.0	6.3	63.4
21	*51.0*	4045.0	3.5	35.4
22	*44.1*	1768.3	1.5	15.5
23	*64.0*	19215.7	16.8	168.1
24	*50.7*	5928.0	5.2	51.8

MCF7	*212.0*	35939.3	31.4	314.3
H20	*0.0*	4.0	0.0	0.0
25	*94.8*	522.0	0.4	4.5
27	*561.2*	651.0	0.6	5.5
28	*57.5*	2923.7	2.5	24.9
29	*172.9*	317.7	0.3	2.7
31	*36.7*	1800.3	1.5	15.3
32	*53.9*	5305.7	4.5	45.2
33	*91.6*	72.3	0.1	0.6
34	*26.4*	339.3	0.3	2.9
35	*184.6*	326.0	0.3	2.8
36	*199.4*	17069.7	14.6	145.5
37	*134.5*	2461.7	2.1	21.0
38	*87.4*	2498.0	2.1	21.3
39	*141.2*	2821.0	2.4	24.0
40	*471.6*	4821.0	4.1	41.1
42	*308.9*	2544.3	2.2	21.7
43	*29.5*	3414.7	2.9	29.1
44	*73.3*	1790.7	1.5	15.3
45	*58.8*	1732.3	1.5	14.8
46	*30.9*	1259.7	1.1	10.7
47	*43.9*	1298.3	1.1	11.1
48	*142.7*	1683.0	1.4	14.3
49	*60.3*	4029.7	3.4	34.4
MCF7	*212.0*	24051.7	20.5	205.0
H20	*0.0*	-2.0	0.0	0.0

Furthermore, almost 70% (55/80 samples) failed the initial OncoArray quality control also using a Picogreen assay. After discussion, we decided to continue with the OncoArray, and over 98% of samples passed the data analysis quality control measures to produce good data (Appendix 8). This illustrates that conventional QC checks, based on the presence of long sections of double-stranded DNA, are unsuitable for this resource because of the fragmentation of DNA prior to extraction. However, hybridisation of DNA onto a SNP array should not be affected because the SNP will either be present or not, and DNA fragmentation appeared to not affect these results. Similarly, provided that the PCR primers are designed to produce short

amplicons (<100 bp), results can be generated. The issues of DNA quality with this resource highlights more than ever the need for validation of results using a sample source specific technique.

4.3.4 450K analysis and pyrosequencing

We profiled DNA methylation in 36 individuals using the Illumina Infinium HumanMethylation 450 BeadChip and all chips passed the quality control procedures for samples and probes. Quality control measures for samples included assessment of the DNA concentration, bisulphite conversion and quantity, while those for probes included checking that the probe detection cut-offs and minimum bead counts at each probe (>3) were reached, in addition to cross-reactive probes and those containing SNPs were removed from the final analysis (Morris and Beck, 2015). The ComBat analysis, which is used to adjust for the chip-specific batch effect and conducted in R, showed acceptable low levels of variability between the three chips combined into a single data file after batch correction (Fig. 4.7).

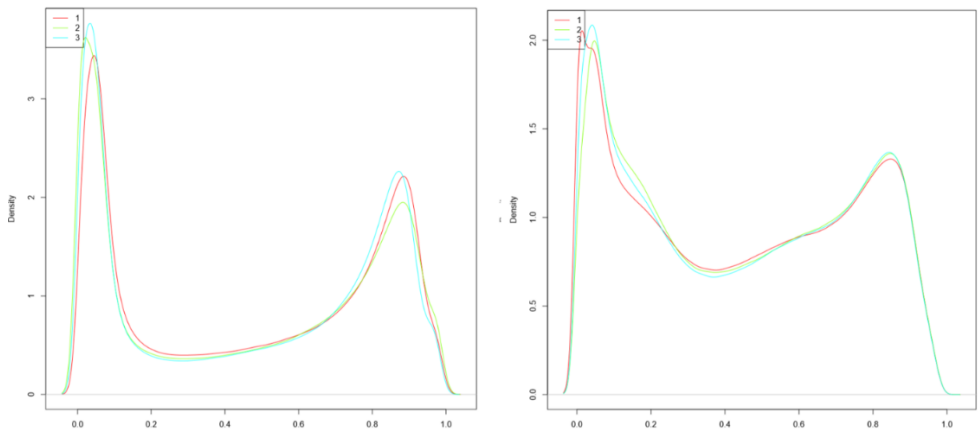


Figure 4.7: COMBAT analysis for the three chips of EBM samples (n = 36 in total, 12 in each chip) showing a high level of similarity in the bimodal peaks of the three colours for each.

Data from the ESR1-associated 450K probes was extracted and showed low levels of promoter methylation, with high levels of intragenic methylation (Fig. 4.8A). Pyrosequencing of a separate set of bisulphite-converted DNA samples (n = 53) validated the 450K ESR1 methylation profiles (Fig. 4.8B). Methylation levels increased rapidly within the first intron, but was highly variable between samples as evidenced by the wide error bars in both the 450K data and at the region analogous to p2 in the pyrosequencing assays. IGM levels remained high at most regions assayed along the ESR1 gene, in a pattern that was very similar to that shown for ER-positive tumour cell lines (Fig. 4.8B, vs. Fig. 2.3). The raw data is also shown in Table 4.3.

Table 4.3: Results of triplicate repeat pyrosequencing assays across *ESR1* in bisulphite-converted DNA samples from donated EBM (n = variable, stated in table). Samples that also underwent 450K analysis were excluded from this data.

Locus	p1	Promoter	p2	p3	p4	p5	p6	p7	p8
Genomic locus	-4371	178	1627	7837	38810	150393	204185	250449	288793
Methylation (%)	97.94	2.07	38.69	94.52	94.75	55.14	93.38	76.60	87.13
SD	2.76	2.02	20.74	4.66	4.86	21.60	3.88	10.08	7.67
n	53	50	38	31	52	29	52	46	53

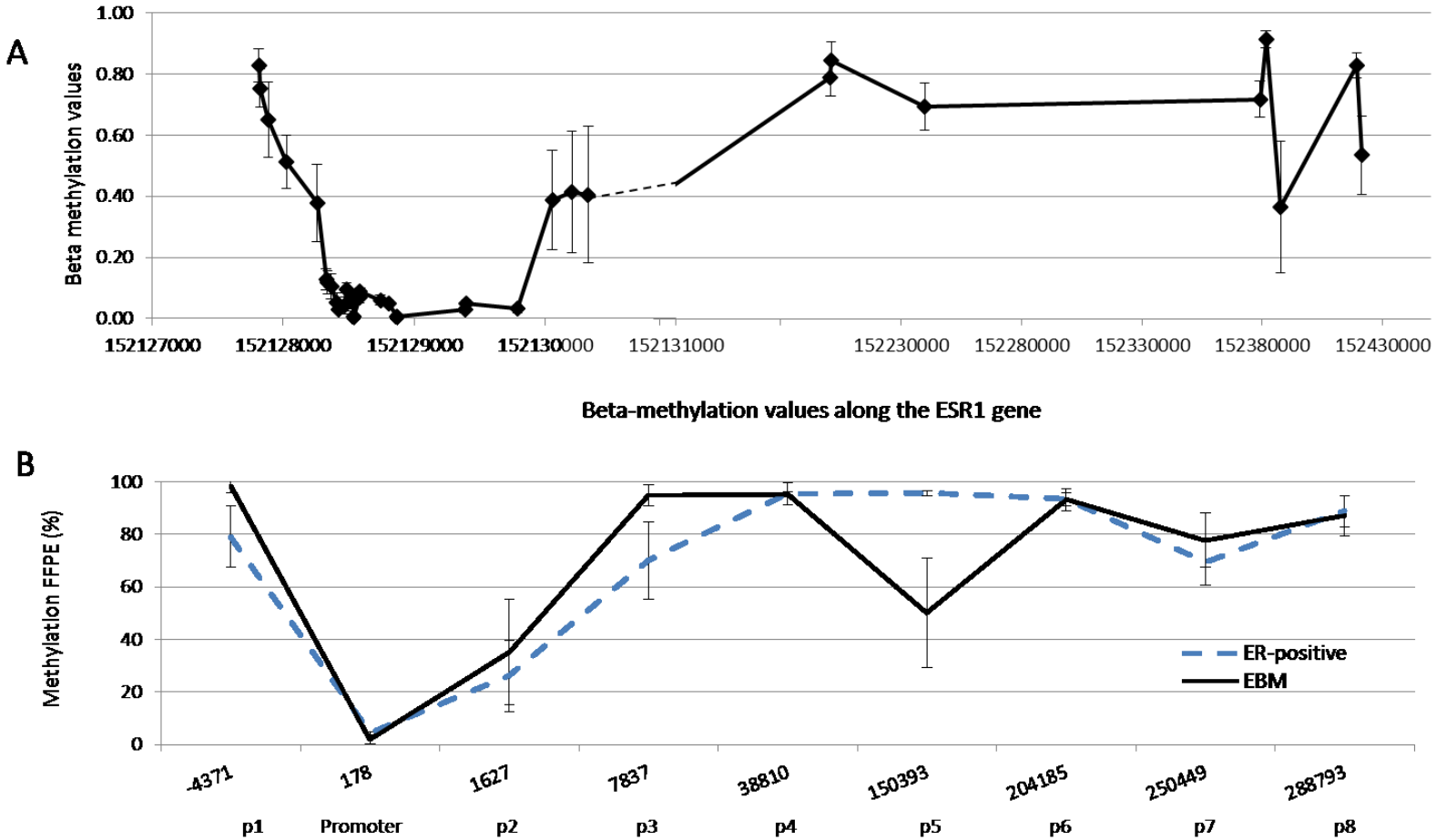


Figure 4.8: A) Methylation values from 450K analysis of 36 DNA samples extracted from breast epithelial cells from expressed breast milk. The overall pattern matched those of ER-positive cancer cell lines, and regions of variability were found at regions that mapped to p2 and p7. The promoter region is shown with a higher x-axis scale resolution than the intragenic region. B) Validation of the methylation levels using pyrosequencing in a separate set of 53 EBM samples (black solid line). The pattern of methylation is very similar to that of the 450K analysis of EBM samples and of ER-pos cancer cell lines (dashed line) with high levels of variability at p2 for both. Reproduced with appropriate permissions from (Shenker et al., 2015).

4.3.5 RNA analysis

According to the results of the Bioanalyser 2100 (RIN score >7, Table 4.4), high quality RNA was obtained from 10 of the 11 fresh RNA samples using a standard TRIzol® technique (Fig. 4.9).

Table 4.4: RNA Integrity Scores (RIN) for the 11 RNA samples from fresh EBM samples and one frozen sample.

Bioanalyser ID	Donor	Source	RIN
1	70	Frozen	2.4
2	70	Fresh	6.4
3	94	Fresh	7.3
4	112	Fresh	7.8
5	115	Fresh	7.3
6	131	Fresh	3.3
7	133	Fresh	7.0
8	135	Fresh	6.0
9	141	Fresh	8.2
10	147	Fresh	7.8
11	153	Fresh	7.9
12	138	Fresh	6.9

Furthermore, the OD260/280 was >1.8 for all 11 samples (data not shown). cDNA was prepared and RT-qPCR assays for ESR1 were performed according to the techniques and primers described above for the cell line analysis (Chapter 2). EBM samples were normalised against MCF7, and MDA-MB-231 RNA was used as the negative control, along with a negative reverse transcriptase sample. These samples were therefore suitable for transcriptome analysis, and at the time of writing funding was approved for this to take place for eight samples at the Oxford Genome Technology, but the results were not as yet available.

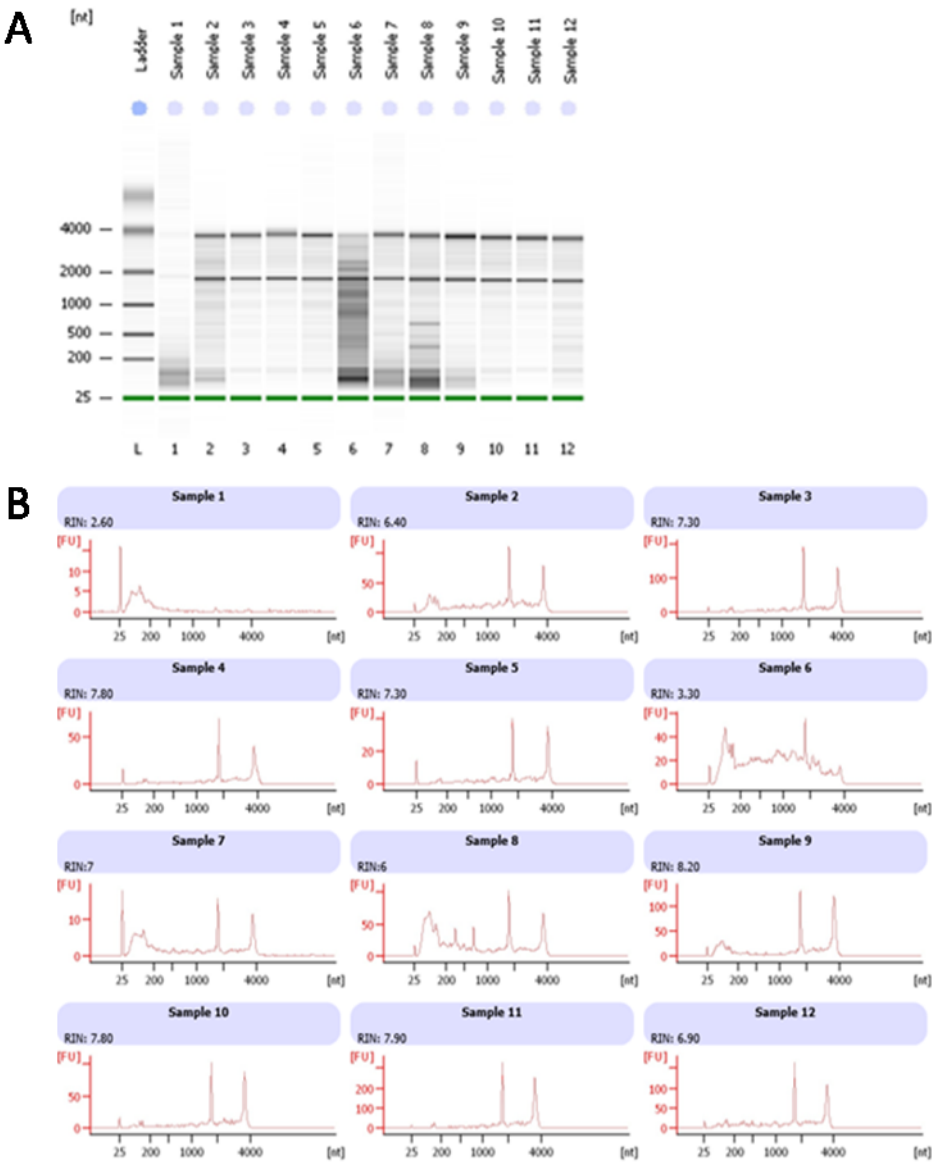


Figure 4.9: Quality control data from the Bioanalyser assays showing A) the image of a gel showing bands at the expected molecular weights for ribosomal RNA (18S and 28S at ≈2,000 and ≈4,000 nucleotides, respectively), and B) the same peaks at the expected loci.

ESR1 expression has not been characterised previously in breast epithelial cells from human expressed breast milk. The 11 samples underwent simultaneous RT-qPCR assay for ESR1 using the same primers for ESR1 and GAPDH as described in Chapter 2 of this thesis. MCF7 cell line-derived cDNA was taken as the baseline (fold change = 1), against which the EBM cDNA samples were normalised.

The results showed that each of the EBM samples had high levels of ESR1 expression, with an average increased fold change relative to MCF7 of 7.37 (range, 2.2- to 19-fold, Table 4.5). However, no significant correlation was found between ESR1 expression levels and the methylation variable region (p2) in this small subset of samples (R^2 = 0.0082, Fig. 4.10).

Table 4.5: Fold change results from the RT-qPCR assay for *ESR1* mRNA in 11 EBM samples, MCF7 and MDA-MB-231, relative to GAPDH.

Sample ID	Replicate 1	Replicate 2	Replicate 3	Average fold change
MCF7	1.27	0.92	0.86	**1.01**
#70	10.08	7.81	13.75	**10.55**
#94	6.97	7.20	7.11	**7.10**
#112	16.58	21.02	19.20	**18.93**
#115	2.34	3.62	3.63	**3.20**
#131	8.69	9.26	11.02	**9.66**
#133	3.15	3.93	2.91	**3.33**
#138	10.50	9.08	10.85	**10.15**
#135	5.95	6.34	4.12	**5.47**
#141	3.79	5.78	3.99	**4.52**
#147	2.95	1.69	1.96	**2.20**
#153	11.93	12.41	12.61	**12.32**
MDA-MB-231	0.01	0.06	0.06	**0.04**

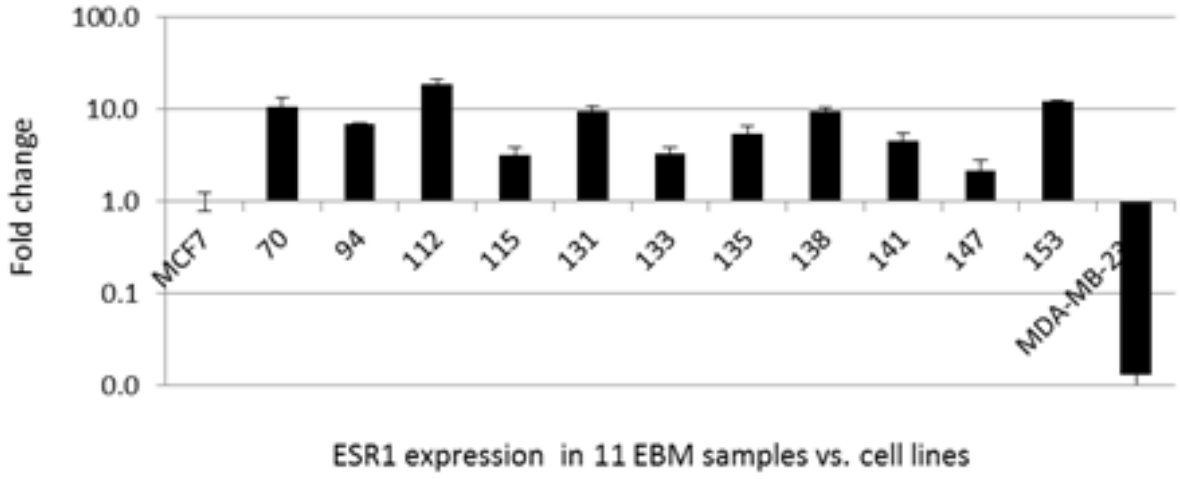


Figure 4.10: RT-qPCR analysis of 11 EBM epithelial cell RNA samples showing high levels of ESR1 expression normalised against ER-pos MCF7 and ER-neg MDA-MB-231 cell lines.

4.2.6 Genotyping results

Data from five SNPs that are associated with breast cancer risk were extracted from the array to check genotype frequencies within the EBM samples (Table 4.6). The failure rate was relatively high for one SNP in PAX9 (rs2236007) at 13% (n = 10), but only one other sample failed in one other SNP in these 400 assays (5 x 80; overall failure rate: 3.5%). As the majority of this population were from a Caucasian European background, genotype frequencies were checked against the reported literature for genotypes in European populations according to the HapMap Genome Browser (release #28) and the NCBI Short Genetic Variations database (http://www.ncbi.nlm.nih.gov/SNP/). Table 4.6 shows that there were no significant differences between the observed and expected genotypes.

The OncoArray data was validated by genotyping assays using PyroMark genotyping software on the pyrosequencer for five SNPs associated with breast cancer risk (Michailidou et al., 2015). The data broadly validated in the 27 individuals with data from both assays, as shown in Table 4.7, with only nine inconsistencies. Most of the inconsistencies between the two techniques occurred for PAX9. Although there were some technical challenges in the data generated by genotyping on the pyrosequencer, some degree of error may have been inherent in PAX9 on the OncoArray as 12 individual samples failed for this assay on the array.

Table 4.6: Genotype frequencies of the five breast cancer risk-associated SNPs selected from the OncoArray data in 80 EBM samples (minor allele underlined). Observed versus expected genotype frequencies were calculated using the Chi-squared test.

	Genotype	n	Observed freq.	Expected freq.	P-value
PAX9, rs2236007 (T/C)	CC	37	0.53	0.62	
	TC	30	0.43	0.34	0.157
	TT	3	0.04	0.05	
	Failed	**10**			
RAD23B, rs10759243 (A/C)	AA	10	0.12	0.10	
	AC	31	0.39	0.40	0.759
	CC	39	0.49	0.50	
	Failed	**0**			
FGFR2, rs2981579 (C/T)	CC	8	0.10	0.14	
	CT	41	0.51	0.42	0.162
	TT	31	0.39	0.44	
	Failed	**0**			
ESR1, rs3757318 (A/G)	AA	0	0.00	0.00	
	AG	8	0.10	0.13	0.672
	GG	72	0.90	0.87	
	Failed	**0**			
CCND1, rs614367 (C/T)	CC	61	0.76	0.78	
	TC	15	0.19	0.19	0.828
	TT	3	0.04	0.03	
	Failed	**1**			

Table 4.7: Validation of OncoArray SNP data at five breast cancer risk-associated SNPs with genotyping assays (italicised; the latter was mostly performed by Isobel Steer, an MSc student under my supervision). Inconsistencies are highlighted in red.

Donor ID	FGFR2 (T/C)		ESR1 (G/A)		CCND1 (G/A)		PAX9 (G/A)		RAD23B (G/T)	
	OA	G	OA	G	OA	G	OA	G	OA	G
1	CC	*T/C*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GG	*G/G*
18	CC	*C/C*	GG	*G/G*	GG	*G/G*	AA	*A/A*	GG	*G/G*
21	TC	*T/C*	GG	*G/G*		*G/A*	GA	*G/A*	GG	*G/G*
27	TC	*T/C*	GG	*G/G*	GG	*G/G*	GA	*G/G*	GG	*G/G*
29	TC	*T/C*	GG	*G/G*	GA	*G/A*	GG		GG	*G/G*
35	CC	*C/C*	GG	*G/G*	AA	*A/A*	GA	*G/A*	GG	*G/G*
36	TC	*T/C*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GG	*G/G*
39	TT	*T/T*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GT	*G/T*
40	TC	*T/C*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GG	*G/G*
42	CC	*C/C*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GT	*G/T*
46	TC	*T/C*	GG	*G/G*	GG				GG	*G/G*
50	TT	*T/T*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GT	*G/G*
51	TC	*T/C*	GG	*G/G*	GG	*G/G*	GA	*G/A*	GG	*G/G*
53	CC	*C/C*	GG	*G/G*	GG	*G/G*	GG	*G/A*	GG	*G/G*
54	TC	*T/C*	GG	*G/G*	GG	*G/G*	GA	*G/A*	TT	*T/T*
55	TC	*T/C*	GG	*G/G*	GG	*G/G*	GA	*G/A*	GT	*G/T*
59	CC	*C/C*	GG	*G/G*	GG	*G/G*	GA	*G/G*	GT	*G/T*
63	TC	*C/C*	GG	*G/G*	GG	*G/G*	GA	*G/A*	GG	*G/G*
64	CC	*C/C*	AG	*G/A*	GA	*G/G*	GG	*G/G*	GT	*G/T*
65	CC	*C/C*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GG	*G/G*
67	TC	*T/C*	GG	*G/G*	AA	*A/A*	GG	*G/G*	GT	*G/T*
69	TC	*T/C*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GG	*G/G*
70	TC		AG	*G/G*	AA		GG	*G/G*	GG	*G/G*
71	TC	*C/C*	AG	*G/A*	GG	*G/G*	GG	*G/G*	TT	*T/T*
72	TC	*T/C*	GG	*G/G*	GG		GA	*G/G*	GT	*G/T*
73	CC	*C/C*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GG	*G/G*
74	TT	*T/T*	GG	*G/G*	GG	*G/G*	GG	*G/G*	GG	*G/G*
Inconsistencies	**2**		**1**		**1**		**4**		**1**	

OA: OncoArray; G: genotyping assay.

When the 450K data was combined with the OncoArray data, with regards to *ESR1*, all donors with available data for both were homozygotes for the AA allele. As shown in the heat map in Figure 4.11, methylation levels were constant across the *ESR1* gene with some distinct regions of variability.

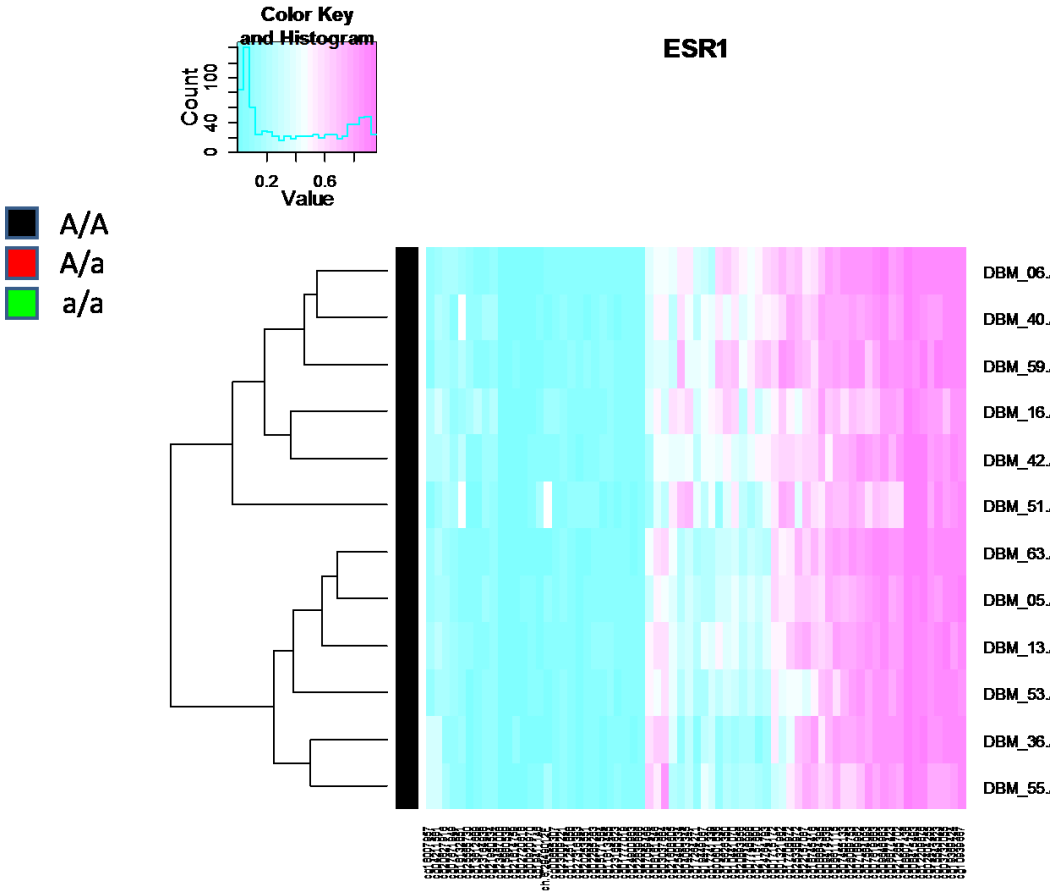


Figure 4.11: Heat map to show correlation between genotype and methylation levels across the *ESR1* gene in 12 DNA samples from EBM. X-axis probes, Y-axis, samples. Blue indicates relative hypomethylation, while pink shows methylated probes.

4.3.7 Combining meQTL data into the polygenic risk score for breast cancer

The analysis and figures were prepared by myself and Dr Flanagan using the data generated from the OncoArray and 450K assays that I prepared. Briefly, our aim was to identify novel methylation quantitative trait loci (meQTL) for breast cancer risk. Using a linear regression, over 254 billion tests (477,201 CpG sites x 533,631 SNPs)

were performed to determine altered methylation levels at CpG sites related to risk-associated SNPs. The Bonferroni correction was applied to account for multiple tests, and p-values <2 x 10^{-13} were taken as significant. A preliminary analysis has identified numerous SNP-methylation associations, which will require validation using genotyping and pyrosequencing assays (or another quantitative assay of methylation).One of the most significant, thus far, was variant rs2380205 (chr10:5,886,484) and cg12219469 (chr10:5,936,426, Fig. 4.12). These results show that it is possible to identify novel meth-QTLs in breast epithelial cells from breast milk, however, further validation of this data is required.

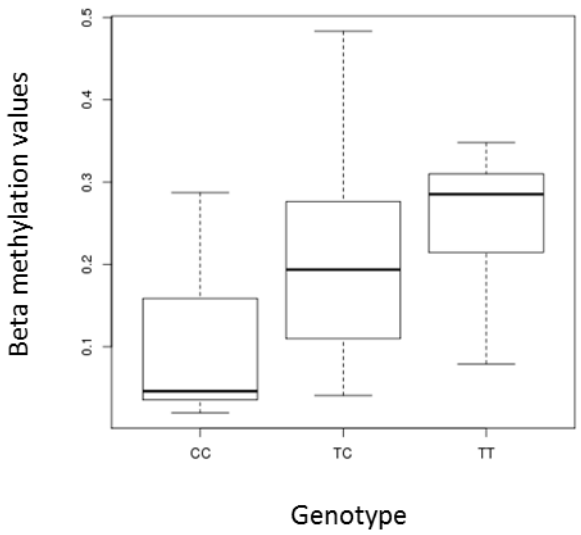


Figure 4.12: Differential methylation depending on genotype at a single CpG site, cg12219469.

4.3.8 The methylation signature of former smoking in donor samples

Current smokers are excluded from donating to the milk banks, and we did not have sufficient former smokers in the methylation array cohort (n = 1/36) to investigate known smoking markers using the acquired 450K data. Given the limited time and

resources available, I conducted pyrosequencing assays for the top marker associated with former smoking in blood DNA, the 2q37 region (Shenker et al., 2013a), using bisulphite pyrosequencing in 88 individuals (69 never smokers, 19 former smokers). Interestingly, this 2q37 marker was also the only smoking marker independently associated with breast cancer risk in Chapter 3. The results showed a significant reduction in methylation in former smokers at this locus (p = 0.026; Figure 4.13), which was the same trend as shown in PBMCs of former smokers. This is the first evidence that smoking has a direct epigenetic effect on breast epithelial cells of healthy premenopausal women.

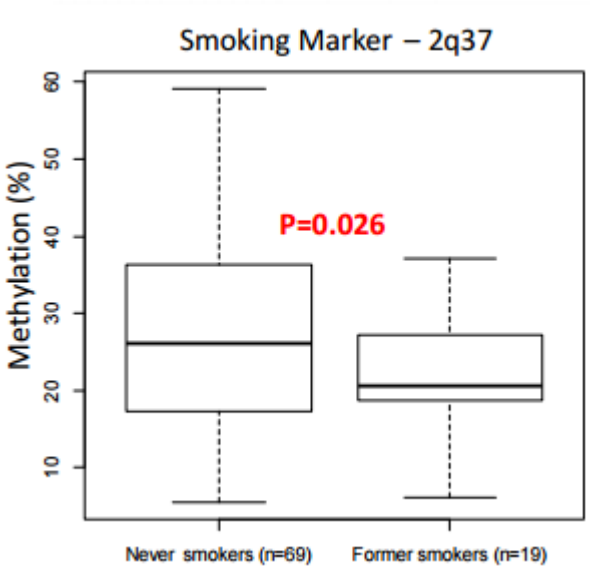


Figure 4.13: Results of pyrosequencing for the top marker previously associated with former smoking in white blood cell DNA, showing hypomethylation in former smokers (n = 19) compared to never smokers (n = 69).

4.4 DISCUSSION

4.4.1 Breast epithelial cells: their role in breast cancer

Although breast cancer has long been thought of as a single disease entity, it is a heterogeneous disease with several distinct subtypes, mainly comprising the oestrogen receptor positive, HER2 amplified and triple-negative groups. Initial gene expression studies with low numbers of tumours identified five broad subtypes (Sorlie et al., 2006, Sorlie et al., 2001), but these were not always reproducible (Gluz et al., 2013). More recent studies have capitalised on technological advancements to screen much large samples of a tumour genome (Jönsson et al., 2010). With the advent of similar technological advancements in epigenetics, specific gene methylation profiling (Muggerud et al., 2010, Perou, 2011, Wei et al., 2008) and genome-wide methylome analyses (Flanagan et al., 2010, Holm et al., 2010, Sproul et al., 2011) have added more information to each tumour subtype, creating the possibility of truly personalised diagnosis and treatment options (Ogino et al., 2013). Specific molecular biology profiles now exist for luminal A, luminal B, HER2 and basal or triple-negative breast cancer (Kamalakaran et al., 2010, Stefansson et al., 2011, Creighton, 2012, Geyer et al., 2012). Overarching studies then incorporate specific datasets on copy number, DNA methylation, sequencing, RNA, microRNA (miRNA) and protein expression for each subtype (Cancer_Genome_Atlas_Network, 2012). This new glut of information has rapidly revised the accepted understanding of how breast cancer arises, and informed treatment options, potential therapeutic targets and prognosis (Reis-Filho and Pusztai, 2011). This information has also yielded information that helps to explain how breast cancer develops, the cells involved, and who is at risk.

Breast epithelial cells form a single layer within the milk ducts (Fig. 4.14). Normal human breast epithelium expresses an array of markers that is indicative of only two epithelial lineages in the mature gland – the epithelial cells and myoepithelial cells. Only the former have the capacity to become lactocytes in the terminally differentiated cellular state, and it is these cells that are most commonly transformed to become tumour cells (Ince et al., 2007). The immature human and rodent breast contain luminal and basal cells from approximately 11 weeks of development with distinct phenotypes (Anbazhagan et al., 1991, Kordon and Smith, 1998, Anbazhagan et al., 1998).

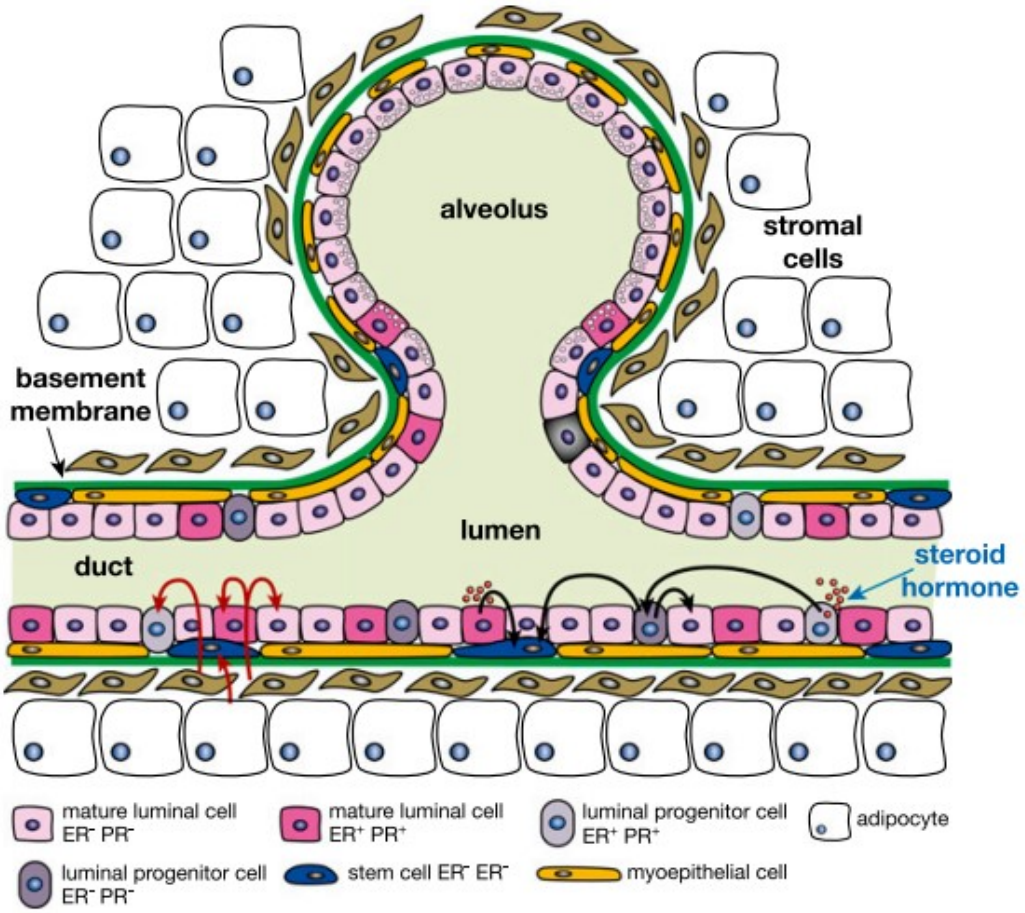


Figure 4.14: The cellular structure of human breast milk ducts and surrounding stroma, reproduced from (Visvader and Stingl, 2014), and reproduced in accordance with the copyright laws laid out in the following link: https://creativecommons.org/licenses/by-nc/4.0/legalcode.

Although a stem cell theory of breast cancer development was propounded for over a decade (Smalley and Ashworth, 2003), it is most likely that breast cancer results from both germline and somatic genetic mutations and epigenetic changes in luminal breast epithelial cells (BECs), causing both familial and sporadic disease (Visvader, 2011). Breast epithelial cells have been transformed in vitro into breast carcinomas (Ince et al., 2007). Surface markers of breast cancers were used to categorise tumours, and hormone receptor positive tumours were initially proposed to originate from luminal epithelial cells, while basal hormone receptor negative tumours were thought to derive from basal myoepithelial cells. This appears to have been a confusion of the nomenclature as, although myoepithelial cells are located on the basal membrane separating the milk ducts from the stroma (Fig. 4.14), true myoepithelial-derived tumours are actually very rare (Yoon and Chitale, 2013); expression analyses showed no evidence that the myoepithelial cell was the cell type of origin (Molyneux and Smalley, 2011, Prat and Perou, 2011). Further expression analyses have shown that basal tumours, which tend to be triple-negative and distinctly different from ER-positive luminal tumours, actually share luminal progenitors in preneoplastic tissue (Lim et al., 2009, Prat et al., 2013, Prat et al., 2014). The use of surface markers to determine the cell type of origin of tumours is therefore less valid, and it is likely that most breast cancers arise from luminal, ER-positive ductal epithelial cells that subsequently de-differentiate via a combination of genetic mutations and epigenetic changes. However, the precise single luminal progenitor remains unknown from the hierarchy of development of breast epithelial cells (Visvader and Stingl, 2014).

Results from our laboratory and others have identified candidate epigenetic markers of breast cancer risk in blood (Brennan et al., 2012, Snell et al., 2008),

which may represent an epigenetic history of lifetime cancer-related environmental exposures (Shenker et al., 2013a, Christensen et al., 2010). Although efforts are underway to understand the role of novel risk markers (Baird and Caldas, 2013) and to seek more candidate markers (Lawrence et al., 2013), their roles in terms of gene expression and function in normal BECs is largely unexplored. As evidenced by the lack of validation of cell line findings by tumour biopsy tissue in Chapter 2 of this thesis (Shenker et al., 2015), and the inherent issues of using both in epigenetic analyses, there is a clear need for a source of homogeneous breast epithelial cells from healthy women, as well as those with breast cancer. The latter had been addressed to some extent with the use of ductal lavage (Dooley et al., 2001, Evron et al., 2001, Fabian et al., 2005, Konstandiadou et al., 2012), but this invasive test is not tolerated unless there is a definite clinical need, and so would be unsuitable for the large-scale epidemiological studies required.

4.4.2 Initial sample characterisation, handling and troubleshooting

The initial aim of this pilot study was to develop a standard quality control protocol that determines the optimal quantity of EBM required to extract sufficient cell numbers, as previous studies have described a range from 10-200 ml (Browne et al., 2011, Densmore and Pflueger, 2006, Gorlewska-Roberts et al., 2002, Thompson et al., 1998, Wong et al., 2010). The successful handling of human milk samples so that low volumes of milk could yield DNA concentrations that were sufficient for experimental purposes entailed a steep learning curve, and relatively little information was available as a guide. A wide range of sample quality was evident, and initial attempts to spin down cell pellets yielded only very low cell numbers (<1 x 10^4 per mL). Using a combination of published protocols (refs) and trial and error.

Optimal speeds, spin durations, and wash numbers were determined from the literature to establish the final protocol.

Milk is an unusual biofluid; although the lipid, protein and sugar constituents are relatively stable once supply is established by 6 weeks postnatally (Thompson et al., 1998), breast milk is highly variable in its fluid consistency depending on the time of day, temperature, and humidity levels in order to provide the baby with nutrition and hydration optimised to its environment. The number of cells per millilitre did not correlate with the quantity and quality of the DNA extracted (data not shown). This observation is likely to be multifactorial, as a result of the variable storage conditions at home (e.g., belated freezing, repeated freeze-thaw cycles and the limitations of home freezing) and transportation issues into the Milk Bank and laboratory.

The cells were shown to be variably viable (11% - 79%), and many of these cells would have sloughed off from within milk ducts as a result of necrosis or apoptosis, and may therefore have fragmented DNA before processing. The cells were highly fragile, and overly vigorous centrifugation presumably resulted in rupture of the cell membrane as a result of the high kinetic energy. However, slow spin cycles (<2000 rpm) produced very small cell pellets; in this scenario, it appeared that the cells were unable to descend through the viscous lipid layers and long chain oligosaccharides.

Regarding DNA extraction, it was impossible to gain sufficient yields using kit techniques with mesh-based tubes and lysis/buffer treatments. Despite industry advice from both the Qiagen and Zymo technical team, this issue was never resolved. Neither company had previously tested these kits for the extraction of DNA from human breast milk, and the available DNA Food Kit [cat. no; Qiagen, Crawley, UK) had been optimised for bovine milk, and also did not work. It is most likely that

this was a consequence of contamination of the cell pellet during processing, and the incorporation of viscous long chain molecules that blocked the mesh. The use of further washes did not improve the success rate for these kits, and the modified phenol: chloroform technique remains the only viable technique, even though it is more labour intensive and required over 24 hours to process samples. Yields increased 4-fold (data not shown) when duplicate phenol and chloroform steps were introduced.

Very low concentrations of DNA were reported using the Picogreen assay both in our laboratory and at Genome Quebec. This is likely to have been due to the Picogreen assay being adversely affected by highly fragmented and single stranded DNA (Sedlackova et al., 2013), which was the case with the majority of these samples. This was proven by running genomic DNA on an agarose gel, which showed smears of DNA fragments of different sizes (data not shown). However, our data output and validation for a range of molecular biology techniques has demonstrated that, provided certain precautions are taken during the experimental processes, viable and useful data can be generated from exfoliated BECs.

The lack of correlation between cell counts and DNA concentration from each 25-ml sample may reflect inadequacies in the process of collection and storage of the milk before it reaches the Milk Bank and laboratory. For example, if the milk has undergone multiple freeze-thaw cycles in an inefficient home freezer or in a freezer that is repeatedly opened, the cells may rupture and the DNA would become more fragmented. This would also be the case if the milk was not frozen quickly after being expressed.

Anti-epithelial membrane antigen (CD227) was used to quantify the proportion of BECs in breast milk samples by flow cytometry. The literature on the characterisation

of breast milk-derived cells is divided between the use of EMA or epithelium-specific antigen (EpCAM), however, EMA was selected because it has been well characterised as present on the surface of luminal breast epithelial cells (Gomm et al., 1995, O'Hare et al., 1991, Gusterson and Stein, 2012), while the cell surface expression of EpCAM can be downregulated in cells that are stressed, necrotic or apoptotic. This could be one reason why the current results showed levels of 98% epithelial cells, whereas samples processed by Professor K Arcaro's group at the University of Massachusetts using EpCAM only had an epithelial proportion of 65%-70% (personal communication). Further work will be required to optimise other antigens that could further delineate the different epithelial, stem cell and monocyte populations within breast milk (Hassiotou et al., 2013).

4.4.3 Data generation

The observed intragenic *ESR1* methylation in epithelial cells extracted from human breast milk confirmed the in vitro cell line findings of high levels of IGM in ER-positive tumour cell lines, and low levels of promoter methylation. This finding suggests that, if a homogenous cell type is investigated, the epigenetic profile is also more homogenous compared to the tumour biopsy material, where the cell-specific signatures create a mixed signal (Heijmans and Mill, 2012). From the limited number of samples of RNA available from EBM (n = 11), no correlation was found between the highly variable methylated region at p2 in *ESR1* and ESR1 expression using RT-qPCR. However, further studies with a larger number of samples will be required to investigate such associations.

Of note, this is the first study to show the feasibility of using DNA extracted from cells in EBM for 450K DNA methylation analyses. All 36 samples passed the quality

control procedures, with a relatively low level of excluded probes. Although the DNA from such frozen milk samples is relatively fragmented (data not shown), sufficient quality is retained to enable both array and PCR-based assays to be performed. Given that these cells are in a highly proliferative state during lactation (Hassiotou et al., 2013), and represent the most oestrogen responsive breast epithelial cell type, it will be important to collect and characterise further samples of these cells in the future to gain a greater understanding of the normal biology of ductal epithelial cells, and how their epigenetic status differs from cancer cells (Faupel-Badger et al., 2013). They also represent an important resource in which the impact of environmental and intrinsic cancer risk factors can be assessed.

The donors that were former smokers (n = 28) therefore constituted an invaluable cohort. The methylation level at the loci affected by cigarette smoke over a long duration, even in former smokers as described in Chapter 3, was analysed. The direct epigenetic effect of cigarette smoking on breast epithelial cells has not been shown previously, and, excitingly, a significantly reduced methylation level was determined at the CpG site of interest in 2q37. This result should be interpreted with caution because of the relatively low sample size, but the result was validated by the former smoker donor samples analysed by the 450K array. Current smokers are not able to donate milk, because of the risk of the transfer of toxins to the premature babies who receive donor milk. Therefore, there were no current smokers in this cohort of donors. A future sample collection will aim to recruit both former and current smokers to further characterise the molecular biology of breast epithelial cells exposed to tobacco smoke. Given the evidence that smoking in young women (Gram et al., 2005), and in women before the birth of their first child (Hamajima et al.,

2002, Reynolds, 2013, Dossus et al., 2014), is a risk factor for breast cancer, this will be an exciting future study.

Finally, the first data from the OncoArray (n = 80 donors) and 450K analysis (n = 36 donors) was successfully generated and has been preliminarily interrogated using the crossover samples (n = 31) to identify novel meQTLs in this cell type. Future studies will aim to validate these and identify more candidate loci using the increased power of higher sample numbers. These meQTLs could be invaluable for both validating current epigenetic risk markers identified in blood, as well as to incorporate into the current risk prediction for breast cancer (Eccles et al., 2013, Engel and Fischer, 2015, Howell et al., 2014).

4.4.4 Limitations

Lactation is a specific physiological state for in the breast, which represents the terminally differentiated state of the breast epithelial cell. This could represent a significant limitation as the gene expression pattern would differ during lactation from the resting or post-lactation state. This would also have effects on the epigenetic state of these cells which must be taken into account during the data interpretation stages. However, lactating epithelial cells are in their most ER-responsive state, and therefore these cells are ideal for the study of normal ER binding. Both lactating cells and ER-positive breast cancer cells are highly responsive to oestrogen. Therefore, the genes and epigenetic patterns in the context of high levels of oestrogen responsiveness may be more relevant to ER-driven carcinogenesis than cells that are non-lactating and non-oestrogen responsive.

A further limitation of the current study was that the sample recruitment, via a hospital milk bank, is weighted towards a relatively homogeneous population of

women. Dr Stephanie Tempest's team of sociologists at Brunel University have determined that milk bank donors are predominantly professional, employed, white women with a median age of ≈33 years (unpublished data, presented at the UKAMB 2014 national conference). The demographic data from our pilot study were in accordance with these findings. This represents a limitation in terms of the spread of age and ethnicity, and a future nationwide community-based collection will strive to counteract this bias. However, there is also well-documented bias in the type of woman that breastfeeds (Ibanez et al., 2012, Labbok, 2013), with the higher socioeconomic groups predominating; only widespread attitudinal changes will enable breast milk to be collected from women in every part of society.

Despite attempts to produce all possible data before the end of this PhD, unavoidable gaps remain, not least a comprehensive analysis and validation of the 450K and OncoArray data from the limited number of samples processed to date, and the generation of the transcriptomic data. Furthermore, because of the nature of the sample collection and wastage of samples and DNA in optimisation steps, very few samples had overlapping data from each experiment, meaning that a complete picture of BECs from a single donor sample cannot at this stage be established. However, a great deal of groundwork has been laid by this pilot proof-of-principle study, not least that a relatively homogeneous population of BECs can be successfully isolated from this bioresource, and that an array of molecular biology assays can be performed that generate meaningful results. Future work will focus on optimising DNA and RNA extractions from smaller volumes of breast milk so that sufficient material is available to undergo 450K, transcriptomic, and SNP array analyses, in addition to ChIP-seq and mass spectrometry assays, with appropriate validation steps.

4.4.5 Conclusions

This chapter describes the work of a pilot study to establish the feasibility of the use of breast epithelial cells extracted from breast milk samples donated to a hospital milk bank. There has not been any investigation into breast milk-derived epithelial cells using advanced epigenetic tools to date, and this project has used and added to existing protocols to optimise DNA and RNA extraction procedures for each type of analysis. This body of work lays the groundwork for a future large-scale collection of BECs samples from breast milk across the UK and beyond, with the aim of generating methylation signatures that could enhance breast cancer risk prediction on an individualised basis.

CHAPTER 5

DISCUSSION, AVENUES FOR FUTURE RESEARCH AND CONCLUSIONS

5.1 GENERAL DISCUSSION

The original purpose of this body of work was to assess intragenic methylation levels in a single gene, *ESR1*, a gene that has profound implications for breast cancer aetiology, treatment and prognosis. During the first decade after the millennium, the majority of research into DNA methylation was focussed on the relationship between CpG methylation at the promoter regions of genes. Dogma held that DNA methylation was only functionally important at promoters associated with CpG islands (CGI) (Bird, 1986, Suzuki and Bird, 2008, Deaton and Bird, 2011, Jones, 1999), whereupon hypomethylation enabled gene transcription and hypermethylation blocked the formation of the transcriptional machinery by altering the conformation of the DNA double helix (Nathan and Crothers, 2002). Although array-based technology was enhancing the breadth of genes that could be investigated at a single point in time, the most widely available platform, the Illumina 27K methylation array, was designed solely to assess promoter-associated methylation. The functional role of intragenic methylation in humans was a relatively unknown quantity in 2010, especially at the level of a single gene.

The primary goal of my first doctoral studies was to establish the relative levels of IGM in a range of breast cancer cell lines to determine whether patterns of methylation varied according to expression levels as had previously been reported in plants (Zilberman et al., 2007) and lower animals (Zemach et al., 2010b, Zemach and Zilberman, 2010), using an array of epigenetic and molecular biology techniques, and to address whether intragenic methylation at the level of an individual gene was functionally important for gene transcription. The results of pyrosequencing from the six breast cancer cell lines showed clearly that different

patterns of intragenic methylation were present and ubiquitous, within a spectrum, between cells that were ER-positive or ER-negative. Expression analyses conducted after modifying methylation levels with the demethylating agent, decitabine, also indicated that reduced levels of methylation elsewhere from the promoter was functionally important. However, it was not clear from this work and the various mechanistic studies subsequently conducted whether this related to single CpG sites, methylation levels across the entire gene, or off-target effects that impacted ESR1 expression either in *trans* or *cis*.

One of the central questions raised by studies of intragenic methylation is what is the functional role? Functional studies have increasingly indicated that intragenic methylation status is a by-product of the transcription machinery, controlled by other epigenetic mechanisms. Importantly, knockout studies of the DNA methyltransferases in stem cell lines showed no effect on gene expression levels versus controls (Blattler et al., 2014). However, no similar studies have as yet shown in differentiated cell types, tumour cell lines or in vivo-derived models or biopsies. Therefore, intragenic methylation signatures may represent ‘effect’ rather than ‘cause’ in stem cells, but be more functionally important at different developmental stages and disease states including cancer (Maunakea et al., 2010, Maunakea et al., 2013, Malousi and Kouidou, 2012, Brenet et al., 2011).

During the course of this PhD, technological and methodological advances have led to increased insights into the question of the role of intragenic methylation. The results of this current work can now be assessed easily since the advent of 450K technology, which included CpG sites across the genome including intragenic and intergenic loci (Dedeurwaerder et al., 2011). The 450K array uses both Infinium I probes (two probes for the same locus, but one is specific for methylated CpG sites

and the other for unmethylated) and Infinium II probes (a single probe, with different colours for methylated [green] and unmethylated [red] CpG sites). Despite the well-recognised limitations of the Illumina HumanMethylation450K array, in terms of the crossover between probes and other genomic regions (6-10% of probes) or single nucleotide polymorphisms (9.4% of Infinium I probes, 15.5% Infinium II probes) (Chen et al., 2013), the potential for batch effects (Harper et al., 2013) and spurious analysis results (Wilhelm-Benartzi et al., 2013), this tool still represents a major advance in epigenetic analysis. The 450K array examines methylation at 485,577 CpG sites, across 99% of RefSeq genes and includes every genomic region, including intragenic and intergenic sites.

The finding of differential patterns of methylation in cells lines that correlated with ESR1 expression opened the doors for biomarker development and clinically relevant modulation. There was a strong rationale to validate the cell line observations in patient samples. However, as described in Chapter 2, major difficulties were encountered in handling tumour biopsy material for epigenetic research, not least the fragmentation of the DNA that led to PCR assay failures and inadequate pyrosequencing results. These technical issues were largely overcome by redesigning each primer set to produce amplicons that were <100 bp. After further optimisation steps of the PCR protocols for each assay, results were obtained from the FFPE samples that were consistent with fresh frozen samples, although with wider ranges within samples groups. However, the patterns of IGM between tumour types did not replicate the cell line data, and there was a large spread of methylation scores between individuals with similar disease profiles (Shenker et al., 2015).

In fact, these conflicting results highlighted one of the central issues in epigenetics: which research tool represents the closest epigenetic ‘truth’ – the cell

line or the tumour biopsy data? Neither is a complete surrogate for true biology as an experimental tool. Cell lines are immortalized through multiple passages in an in vitro environment that does not represent a good surrogate for the in vivo microenvironment of tumour cells. The interactions between the epithelial and mesenchymal layers, the stroma in the breast, are highly important during organogenesis (Sakakura, 1991), as well as during pathological states including cancer (Kang and Massague, 2004, Moreno-Bueno et al., 2008). Therefore, the in vitro analysis of a single cell type cannot possibly provide more than a snapshot of cellular function.

Furthermore, cells in culture rapidly lose certain epigenetic characteristics found in normal and cancer cells, including 5-hydroxymethylation, because of the loss of TET gene expression (Dahl et al., 2011). As 5-hydroxymethylation levels are recorded by the pyrosequencer during assays targeted at methylcytosines, relatively higher levels of 'methylation' would be observed in tumour biopsy samples compared to cell lines. To address this hypothesis for the difference in cell line and tumour results, we sought to analyse 5-hydroxymethylation levels at each locus across ESR1. However, to put the relative level of 5-hydroxymethylation in context, work presented this year at the American Association of Cancer Research by Dr Delgado-Cruzata showed that the percentage of 5-hydroxymethylation at single loci in cancer tissues constituted less than 1% of the overall contribution to pyrosequencing methylation levels in hepatocellular carcinoma (unpublished data). This small difference is even within the accepted range of variation in pyrosequencing results, and therefore cannot explain the large differences between tumour biopsies and cell lines shown in Chapter 2, but further work will be needed to assess the relative levels and function of 5-hydroxymethylation in normal breast tissue and tumours.

The most likely explanation for the lack of correlation between tumour biopsy and cell line methylation patterns was the heterogeneity of cell types within each slide. Heterogeneity has long been a major criticism of epigenetic studies, described as one of the 'plagues' (Heijmans and Mill, 2012). Within the epigenetics community, attempts have been and are continuing to be made to rectify this deficit, with the establishment of comprehensive atlases of cell-specific epigenetic data. The primary aim of these consortia is map the normal epigenetic state of each cell type, in order to provide a benchmark against which pathological findings can be compared. The largest of these is the Epigenome Roadmap (Kundaje et al., 2015). Tissue samples are also being analysed across an array of tumour types by The Cancer Genome Atlas (TCGA) consortium (http://cancergenome.nih.gov/), and the publically available nature of such vast datasets with comprehensive methodological and analysis oversight is accelerating the rate at which advances can be made.

With regards to our samples, heterogeneity was considered from the outset. Although areas of tumour were marked on each haemotoxylin and eosin slide by a pathologist in order that the tumour tissue could be macrodissected prior to DNA extraction, only >70% of that area could be described as tumour tissue. However, relatively recent work has shown that areas within discrete tumour blocks may contain different clonal variations that are markedly different in terms of their proteome and genomic characteristics. Theoretically, the clonal evolution model assumes a branched evolution of cancer and is a leading hypothesis to explain how heterogeneity within a tumour is generated (Turner and Reis-Filho, 2012, Meacham and Morrison, 2013). High levels of cellular heterogeneity have been determined within a range of tumour types, specifically pancreatic neoplasms (Zhu et al., 2007, Samuel and Hudson, 2011) and breast cancer (Ng et al., 2015, Sreekumar et al.,

2015, Baird and Caldas, 2013), amongst others. Even within a single cell type, such as infiltrating T lymphocytes, an astonishing amount of heterogeneity can be present within different regions of a single histological slide (Emerson et al., 2013, Gerlinger et al., 2013, Robins et al., 2013). Each of these areas of pathological and healthy cell types would contribute a different methylation pattern to the overall methylation data (Dedeurwaerder et al., 2011), with the mixed picture generating a completely different pattern of IGM to that generated from homogeneous cell lines, regardless of their limitations.

Epigenetic signatures are increasingly investigated in the context of how environmental exposures can influence cell function and genetic expression (Ladd-Acosta, 2015), with additional relevance as such signatures are unique to specific cell types. One of the strongest criticisms of cancer epigenetic association studies (EWAS) is that most results are generated from blood, which does not represent a homogeneous population of nucleated cells, but is not a reasonable surrogate for understanding cell-specific epigenetic changes (Brennan and Flanagan, 2012). Results from heterogeneous tissues should accordingly be interpreted with caution, and where possible, a homogeneous cell population of interest should be used for epigenetic analyses.

In the context of cancer research, one of the most active areas of research is the development of tools that enable clinicians to assess risk years before disease develop. With regards to epigenetics, such tools must be based on an understanding of cell-specific exposure to known cancer risk factors, and an analysis of the most appropriate cell types. Chapter 3 of this thesis describes how a methylation signature of exposure to cigarette smoke was established in blood cells. Consistently reduced levels of methylation were validated at several sites across the genome, many of

which were stable for decades after smoking cessation (Shenker et al., 2013a). Methylation assays at four of these sites were combined into a single index that was more predictive of current and former smoking that the current gold standard analysis used worldwide, cotinine (Shenker et al., 2013b). Although the work described here clearly demonstrated expression changes in a mouse model of smoking for affected genes, including AHRR and AHR, the functional relationship between discrete regions of decreased methylation in smokers and gene expression remains to be fully explored. As described above, it is not clear whether intragenic or intergenic DNA methylation levels are directly functionally relevant, individual CpG loci found within enhancers, promoters and other regulatory regions of the genome may play a highly significant role on expression at a single gene level. Mutations or methylation level changes at these sites relative to adjacent regions may have more significant functional impact than those observed after global genome methylation changes. The intergenic regions identified in the smoking work at 6p21 and 2q37 both map to regions of DNase hypersensitivity and regions of active chromatin.

Multiple environmental exposures signatures are now available, including methylation states of alcohol exposures (Bagnardi et al., 2015, Flanagan et al., 2015), aging (Horvath, 2013, Li et al., 2013, Mendelsohn and Larrick, 2013, Wnuk et al., 2013), and obesity (Horvath et al., 2014). The speed at which results can be generated reflects the ease at which epigenetic analyses can now be conducted, and the decreasing cost of such studies. For example, our validated smoking signature published in 2013 has since been cited over 100 times by a range of studies from different populations, ethnicities and smoke exposure levels. Broadly the same loci are shown repeatedly, but little work to date has indicated why the CpG sites at these loci are affected, and what relevance if any this phenomenon has

for cancer risk and disease development. There is also the need for mechanistic work to be conducted in cell types relevant to disease states, and in our group, specifically the breast epithelial cell.

This issue has been a constant background concern throughout the period of my thesis. Despite the success of validating 450K data regarding methylation signatures of current and former smoking exposures, the central questions raised by the failure of the cell line data to validate by tumour biopsy assays remained. This was particularly the case, as the samples had been the first to be provided by the newly established Breast Cancer Campaign Tissue Bank (renamed subsequently as the Breast Cancer Now Tissue Bank). Our study was the first recipient of samples. This precious resource should be used to its greatest effect, and the validation of cell line epigenetic results may not therefore be its best use.

It became apparent during the course of the work in this thesis that the analysis of tissue specimens for cell type-specific methylation signatures was not fit for our required purpose, and there was a profound need for a homogenous population of normal breast epithelial cells from healthy women for epigenetic analyses. This is a difficult issue, as cell lines from normal mammary epithelium obviously have the same issues that affect tumour cells lines after immortalisation. Given the single or double cell layer of epithelial cells lining milk ducts, these cells cannot be microdissected, and ductal lavage as an invasive technique is unsuitable and not tolerated for research purposes alone (Fabian et al., 2005, Dooley et al., 2001). Tools such as the use of bead technology to extract pure populations of epithelial cells are expensive and technically labour intensive (Gomm et al., 1995). Therefore, gaining access to sufficient numbers of samples from different individuals for population level research is impractical using currently available techniques. To

address this issue, during the first year of my PhD I spent 6 months working in collaboration with breast surgeons at the Royal Marsden Hospital and researchers at the Institute of Cancer Research to develop a method for extracting a pure population of breast epithelial cells from surgical specimens. However, this technique was labour intensive, susceptible to infection because of the large number of tissue culture steps, and yielded very low numbers of cells even after 3 weeks of passaging. Administrative blocks eventually rendered the project unsustainable, and this had led to a large gap in where I had hoped to take the initial data from the cell lines. The serendipitous discovery prior to the third and final year of my PhD that expressed breast milk reportedly contained high numbers of such cells was an excellent opportunity.

The characterisation of the cells derived from the donor breast milk samples obtained from the hospital milk bank showed that the population of epithelial membrane antigen (EMA)-positive cells, a marker of differentiated epithelial cells (O'Hare et al., 1991), was very high with a median of approximately 97%. This was substantially higher than had been found by Dr Kathleen Arcaro's group in the USA, who found only about 70% of cells to be epithelial, using the EPCAM antibody for flow cytometric analyses (personal communication). There is controversy about which epithelial marker is optimal for addressing cell lineage in lactating ducts (Stingl et al., 2005), and further work will be required to characterise these cell precisely, and separate out homogeneous and reliable populations of stem cells, myoepithelial cells and lymphocyte components.

Once extracted and pelleted, the milk-derived breast epithelial cells could be subjected to a range of molecular analyses, with the successful extraction after some optimisation of DNA and RNA that could undergo array-based techniques for

methylation (Illumina 450K), SNP analysis (OncoChip array), and transcriptomic analysis (Oxford Genome Technology, Oxford, UK). From the 36 samples that underwent 450K methylation analysis, the information of methylation across the *ESR1* gene was extracted and validated by pyrosequencing (n = 56). Interestingly, the pattern mapped almost exactly to that observed in the initial ER-positive breast cancer cell line data (Shenker et al., 2015), which leant weight to the hypothesis that the cell line data, despite its limitations, reflected the epigenetic state more accurately than the tumour biopsy data. Therefore, one of the most significant reflections from my work has been that results from tumours should be interpreted with great caution.

In an attempt to determine whether breast milk-derived breast epithelial cells could be used to validate the methylation signature of smoke exposure. Donors to the Milk Bank are screened prior to donating for exposure to compounds that would prevent them being able to donate milk to vulnerable premature infants. This includes a screen of whether they have recently (<4 months) received blood products, whether they are taking any medication (prescription or non-prescription), and whether they are smokers. Currently, active smokers are not allowed to donate milk. However, in our cohort, DNA samples from six former smokers underwent 450K methylation analysis. Their methylation patterns were significantly lower at the loci identified in Chapter 3 as susceptible to demethylation after exposure to smoke exposure. This finding is the first to demonstrate a direct effect of tobacco exposure on the breast epithelial cell. The data needs to be validated by studies with higher numbers of former and current smokers, but may represent an exciting advance in our understanding of smoke exposure and risk. As a side issue, it would also enable women who are currently smoking to donate their milk for research purposes, and

therefore not feel that their excess milk is otherwise wasted (Updegrove, 2013). These results also completed the circle of my PhD, and leave me with several future research directions to pursue.

5.2 FUTURE RESEARCH DIRECTIONS

5.2.1 Breast feeding and cancer risk

Milk is the primary source of nutrition for newborns, and the World Health Organization guidelines published in 2003 recommend exclusive breast feeding without supplementation until the infant is around 6 months of age (2010). This may represent the single most beneficial public health intervention for reducing mortality and improving health in children aged <5 yr (Jones et al., 2003). Trends in cultural practices of infant feeding have changed worldwide over the last century, largely due to the increased use of supplemental infant formula. A recent review of Red Book statistics from the National Infant Feeding Survey reported that the proportion of women in the UK who breastfed exclusively to the age of 4 months was 23%, falling to 1% by the time the infant is 6 months old (Figure 5.1) (HSCIC).

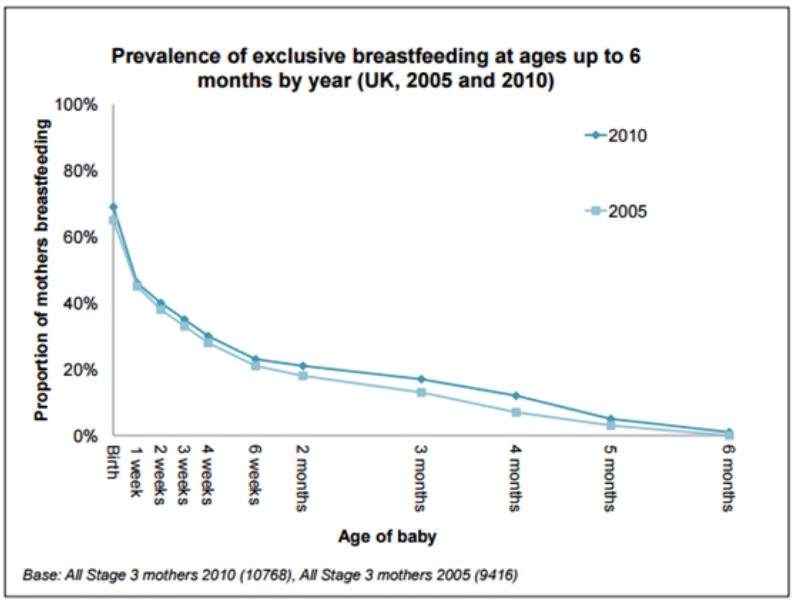


Figure 5.1: Prevalence of exclusive breastfeeding in the UK at ages up to 6 months in 2005 and 2010 (reproduced from the Infant Feeding Survey, http://www.hscic.gov.uk/catalogue/PUB08694/ifs-uk-2010-chap2-inc-prev-dur.pdf).

Efforts are becoming more focused to acquire information as to why the choice of infant nutrition has moved away from breast feeding, and how and whether public health policies can reverse this trend (Labbok, 2013).

Breast milk contains diverse components (Ballard and Morrow, 2013), comprising not only the proteins, sugars, fats and enzymes that provide complete nutrition (Bertino et al., 2013), but also fat globules that transport miRNAs and other signaling molecules, soluble proteins, immune cells, including macrophages and lymphocytes, secreted antibodies, cytokines, microRNAs, and antimicrobial peptides (Armogida et al., 2004). These latter factors protect the infant from infections (Kramer and Kakuma, 2012), allergies (Luccioli et al., 2014, Julia et al., 2015) and enable signaling between mother and influences with consequent influences on appetite and satiation (Ip et al., 2007). Breast milk may play a role in the induction of tolerance in the infant (Strobel, 2001), and contains growth factors that help appropriate intestinal development (Kosaka, 2010). The anti-inflammatory and immune-modulating substances, including cytokines and their soluble receptors, appear to protect the breastfed infant from immune-related diseases in later life, including type 1 diabetes, obesity, celiac disease and some forms of cancer, including lymphoma (Davis et al., 1988), leukaemia and breast cancer (Freudenheim et al., 1994, Nichols et al., 2008, Martin et al., 2005), although the latter finding has been disputed by a large case-control study (Wise et al., 2009). One source of data that I aim to access is the Hertfordshire Cohort Study managed by the MRC Population Health Sciences Institute in Southampton, who have not only the data from this prospective study of infant nutrition and long-term health, but stored blood samples from each participant. The Hertfordshire Cohort Study was established in the 1930s, and collected information about every baby born in a single

county in the UK over a 10-year period, and followed up each individual over their lifetime (Syddall et al., 2010). This is the cohort from which the Barker hypothesis was generated (Barker et al., 1989), and continues to be a source of information about early infant exposures and later health (Robinson et al., 2013). An analysis of this cohort may yield some clues regarding the effects of breastfeeding in breast cancer risk in both the mother and the infant, to supplement the paucity of data that addresses this question in the current published literature. As one explanatory hypothesis, breast fed infants tend to gain weight at a faster rate than formula fed infants , and weight gain >25 g per week is protective of later breast cancer (Lagiou et al., 2008). However, this field of research needs more carefully designed studies with larger cohorts to make firm conclusions about the protective effect of breastfeeding on breast cancer risk in the infant, and the number of possible confounders is huge.

What is more clear is that the risk of breast cancer is lowered in the breastfeeding mother (WCRF, 2007). Indeed, the very first epidemiological questionnaire case-control study ever was conducted on this issue by a female physician, almost unique for her time, Dr Janet Lane-Claypon (1877–1967). She collected information from 500 women with breast cancer matched against 500 broadly similar healthy women and assessed a carefully thought through set of risk factors, many of them reproductive. Indeed, many of the breast cancer risk factors that are accepted today originated from this study, published in 1926 (Winkelstein, 2004): lactation, parity, age at first pregnancy, and ages at both menarche and menopause. Moreover, she showed that breast cancer risk increased in women who did not breastfeed.

The total sum of literature that has investigated the association between ever-breastfeeding and never-breastfeeding currently runs to 37 case-control studies, with a further 55 case-control studies that have assessed the impact of breastfeeding duration on breast cancer risk reduction. Although some of the results have been conflicting, particularly from lower powered studies, several meta-analyses have demonstrated that, once methodological heterogeneity is removed and sufficiently large cohorts analysed, the effect of breastfeeding on breast cancer risk reduction is large. The first meta-analysis published in the Lancet in 2002 by the Collaborative Group on Hormonal Factors in Breast Cancer showed a relative risk reduction of 4.3% for every year of breastfeeding combined with 7% for every birth (i.e., episode of fresh lactation); the authors predicted that, if breastfeeding and parity rates in developed countries matched those in developing countries, the cumulative reduction in breast cancer incidence in developed countries would fall from 6.3% to 2.7% (<3 women in 100 <70 years of age) (2002).

In women with high risk genetic mutations, the data appears to be even clearer. In women who are carriers for BRCA1 or BRCA2 mutations, with a lifetime risk of 80% for developing breast cancer, three large case-control studies (Jernstrom et al., 2004, Kotsopoulos et al., 2012) and a meta-analysis of 10 studies (Pan et al., 2014) indicate that breastfeeding for more than 12 months reduces risk by between 32% and 50%, and that this is a time-dependent effect. Breastfeeding is also protective against developing ovarian cancer [ref], and this effect appears to be larger in BRCA1 mutation carriers (Kotsopoulos et al., 2014).

At present, the mechanism that explains how breastfeeding reduces cancer risk remains unclear. The widely taught explanation that breastfeeding suppresses ovarian cycles and resulting hormonal fluctuations, so reducing oestrogen exposure,

does not hold true given that the lactating breast is in its more oestrogen-responsive state. Our group aims to conduct further work to establish whether a component of the milk itself may be responsible, in the context of the BECS study described below.

5.2.2 The Breast Milk Epigenetic Cohort Study (BECS)

Breast Cancer Campaign (recently rebranded as Breast Cancer Now after merging with Breakthrough Breast Cancer), one of the leading charities in the UK for breast cancer research, brought together a panel of experts in a range of associated research fields to generate the so-called 'Gap analysis' (Thompson et al. 2008). This laid out the key areas in which resources and innovation needed to be focussed in order to overcome critical 'gaps' in the understanding of breast cancer biology, risk prediction, treatment and patient experience. From this, the BCC Tissue Bank was conceived and established, and the project described in Chapter 2 of this thesis studied the first assigned batch of tumour samples (Shenker et al. 2015). Unfortunately, the results of methylation analyses across the ESR1 gene yielded results that did not validate the cell line findings, and we were left with a conundrum of which experiment had given a 'true' result.

BCC produced a second Gap Analysis in 2013 (Eccles et al. 2013). This report described 10 gaps in breast cancer knowledge, which need innovation and focus to overcome them. This report specifically states under the heading of Epigenetics, "There is a gap in our understanding of cause or consequence between epigenetic traits and gene transcription." The Breast milk Epithelial Cohort Study (BECS) aims to contribute to filling in this gap, and further understanding the normal biology of the lactating breast epithelial cell.

Our primary future aim is therefore to develop a novel resource of a breast milk biobank, initially for the further investigation of epigenetic changes that can identify risk prospectively, with specific reference for the prediction of individuals who are at a higher risk of developing breast cancer. This study population will constitute the Breast milk Epigenetic Cohort Study (BECS), for which NREC ethics approval is provisionally granted pending minor changes to the protocol, and funding is actively being sought. The development of the BECS cohort will enable multiple aims to be fulfilled:

1. The basic biology of the lactating epithelial cell can be studied in more detail in a systematic manner.
2. The implications of genetic variation and environmental exposures can be understood in the context of breast cancer risk.

Given that epithelial cells from breast milk samples were found to represent over 95% of the cellular component of each sample, and these cells have been successfully subjected to a range of molecular biology techniques, future work will focus on establishing a large population-based study of matched milk and blood samples from women across the UK and beyond. We have already established a collaboration with the UK Association of Milk Banking (UKAMB), in order to collect samples from the nationwide network by rolling out the collection procedures optimised at the QCCH Milk Bank.

5.2.3 The UK Milk Bank Network: purpose and scope

In the premature infant, multiple benefits of providing breast milk to the infant have been shown over synthetic feed (Underwood, 2013). The most obvious of these is a

reduction in the incidence of necrotizing enterocolitis (NEC), an inflammatory condition of the gastrointestinal tract that carries the risk for multiple surgeries and bowel resections, sepsis, the lifelong need for total parenteral nutrition with its associated complications, and death. Human milk, whether maternal or donated from a third party through a Milk Bank, can reduce the incidence of NEC by up to 80% (Boyd et al., 2007, Quigley et al., 2007, McGuire, 2012, Quigley and McGuire, 2014). A reduction in the incidence of sepsis and of bronchopulmonary dysplasia as well as enhanced feeding tolerance has also been observed (Arslanoglu et al., 2010). In a comparison of outcomes between extremely premature babies born in 1995 and 2006, the increased use of breast milk was associated with improved psychomotor developmental index at 30 months (Vohr et al., 2007). In the 2006 cohort, 43% of all surviving babies were receiving breast milk at discharge. Part of the shortfall in infants receiving HM is the ability of the mother to produce milk after delivering so early into the pregnancy. The purpose of milk banking is to provide donor HM to premature and sick infants in place of maternal breast feeding.

The current UK network consists of 17 milk banks, most associated with level III neonatal intensive care units with surgical capacity. The provision of donor milk to premature babies has been greatly facilitated by the collaboration of the Milk Bank network with SERV (Service by Emergency Rider Volunteers), a member of the National Association of Blood Bikes. SERV is an organisation of volunteer motorcycle couriers, who collect donor milk from the homes of donors, transport this material in a frozen state, and enable deliveries to be made of processed milk from milk banks to affiliated NICUs within the catchment area of each milk bank. At present it is unclear whether we can work with SERV to collect fresh and frozen milk samples, or whether other methods of preserving DNA and RNA will need to be

optimised so that the UK postal service can be used instead (Allen-Hall and McNevin, 2012).

With regards to environmental exposures and cancer risk, excitingly, preliminary work in the few samples of breast epithelial cells derived from breast milk of former smokers showed that the methylation at loci identified in Chapter 3 in PBMCs were similarly affected. This is the first indication of a direct epigenetic effect of an environmental exposure, in this case cigarette smoke, on the epithelial lining of milk ducts. Smoking by the mother is known to have detrimental effects with regards to growth (Muraro et al., 2014), epigenetic and neurological effects in the developing fetus (Kupers et al., 2015, Lee et al., 2015, Sabbagh et al., 2015, Joubert et al., 2014, Nielsen et al., 2014, Herbstman et al., 2012), but surprisingly little work has been conducted on the direct effects of cigarette smoke on human breast milk (Baheiraei et al., 2014). Further work will aim to validate these results in larger sample sets of breast epithelial cells and expand our understanding of the methylomic effects on breast epithelial cells of other environmental exposures, including alcohol, obesity and oestrogen exposure.

The collection of large numbers of samples of breast milk would also help to inform the gaps in our current knowledge of the normal functioning of the cellular component of the lactating breast. A limited number of ChIP-Seq assays, validated by ChIP-PCR, have been performed by our collaborator, Dr Luca Magnani, and Miss Van Nyugen (Research Assistant). Three EBM samples have been processed to date for ER and H3K27ac (a marker of transcriptional activation), and these results have shown an interesting pattern of ESR1 binding that is quite different from that described already in ER-positive breast cancer cell lines and tumours (Figure 5.2).

Further work is ongoing to optimise the number of cells in the input, and the most appropriate markers for investigation.

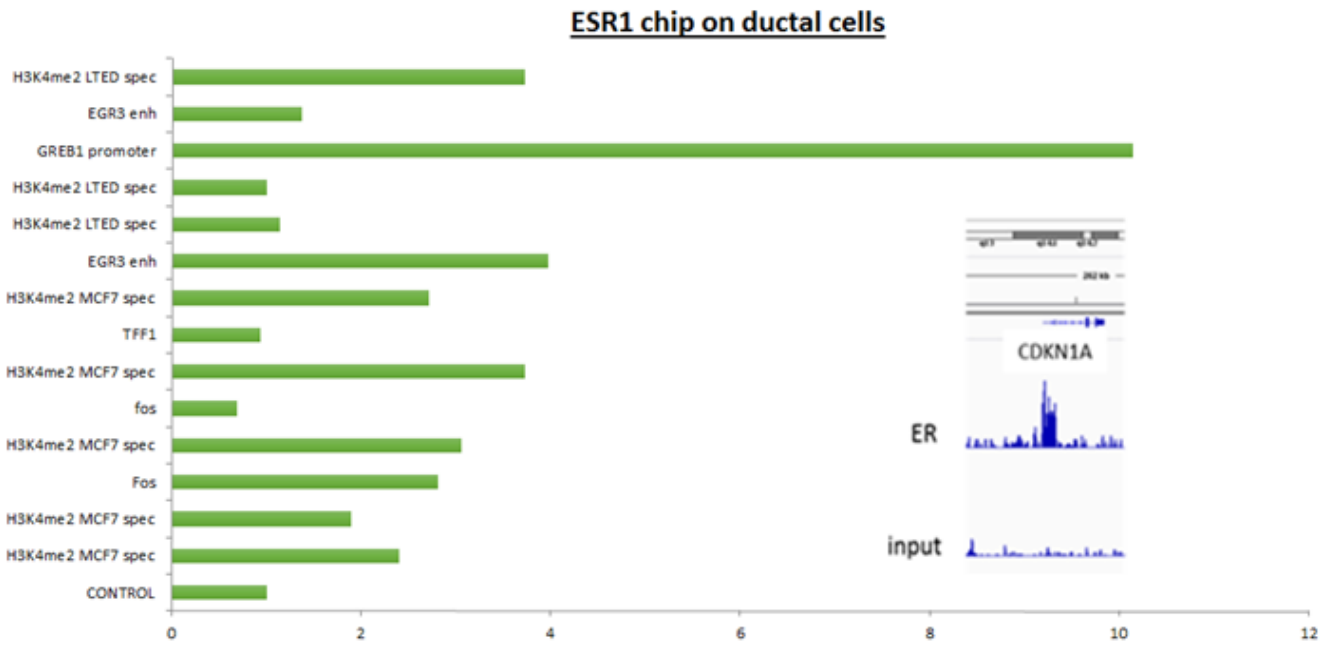


Figure 5.2: ChIP-PCR data for ESR1 in a single EBM sample, showing that the oestrogen receptor binds different targets in normal oestrogen responsive breast epithelial cells compared to breast tumours (input), with the inset showing a specific variable region at CDKN1A according to the ChIP-Seq data (experiments performed by Van Nyugen and Dr Luca Magnani on samples prepared and selected by myself).

Breast milk varies in terms of its constituents depending not only on the time since birth, but also the time of day, relative humidity, temperature and nutritional state of the mother (Koenig et al., 2005). Furthermore, the process of expressing and donating milk can lead to various differences between breastmilk components, in terms of technique, contamination and inter-individual factors (Tully et al., 2001). With the help of techniques such a mass spectrometry and nutritional analysis, further quality control protocols will be developed to standardise potential variable confounders between breast milk samples. For this aim, we aim to collaborate with the clinical mass spectrometry group at Imperial College, specifically to analyse breast cancer risk-associated factors in the lipid profile (Yee et al., 2010). This would

build on experimental protocols available in the literature (Schneider et al., 2014, Smilowitz et al., 2013), as well as recently acquired expertise in the CMS Group at Imperial College (Villasenor et al., 2014).

Other potential avenues of research include characterisation of the stem cell and myoepithelial cellular components of breast milk, and further understanding of the other components that make up the fluid phase of this under-researched biofluid.

5.2.4 HAMLET

As a biofluid of such importance as the primary source of nutrition to the developing infant, breast milk is surprisingly under-researched. Through my recent work, I have been able to meet some of the individual research groups working on different aspects of breast milk biology. The majority of these groups are limited by access to resources, both in terms of milk volumes and funding. A side aim of the establishment of the BECS collection will be to foster future relationships that aim to broaden the understanding of milk biology, infant nutrition and long-term health and cancer risk.

Professor Catharina Svanborg, a microbiologist at Lund University, leads the group that discovered and subsequently characterised the HAMLET protein complex. This potent anti-tumour agent is formed in human milk by the binding of a fatty acid, oleic acid, to the most common casein protein constituent in milk, α-lactalbumin (Fast et al., 2005). As a result, the protein becomes partially unfolded and stable, with unusual membrane uptake properties (Baumann et al., 2012, Gustafsson et al., 2009, Mossberg et al., 2010b). It is preferentially taken up by tumour or transformed cells across a range of tissue types, whereupon multiple

deleterious cellular effects occur that lead to rapid cell death. These effects include the binding of HAMLET to histone tails with nuclear disruption, mitochondrial interference and proteasome degradation (Duringer et al., 2003, Svanborg et al., 2003, Svensson et al., 2003, Baltzer et al., 2004, Gustafsson et al., 2005, Hallgren et al., 2006, Aits et al., 2009, Ho et al., 2012, Xie et al., 2013). While HAMLET also has strong anti-tumour properties, normal adjacent cells are left intact (Mossberg et al., 2010b, Svensson et al., 2002, Mossberg et al., 2010a). Mouse studies and preliminary human studies of papilloma (Gustafsson et al., 2004), colorectal (Puthia et al., 2013), bladder (Xiao et al., 2013, Mossberg et al., 2007), and central nervous system tumours (Fischer et al., 2004) have shown promising clinical results, and funding is now being sought to upscale the production of recombinant HAMLET and conduct appropriately powered early phase clinical trials.

We have already collaborated with Dr Svanborg's group by sending samples of expressed breast milk to her laboratory to undergo testing for HAMLET using column-based extraction systems. She is currently working towards establishing an ELISA-based assay, and I will be working with her group to bring such expertise back to Imperial to assay our samples in house. A method that would enable us to quantitate relative levels of HAMLET, inter-individual and intra-individual variability in terms of both HAMLET levels and anti-tumoricidal capabilities, would lead into numerous future research avenues.

5.3 LIMITATIONS AND STRENGTHS

There were a number of limitations to this body of work. With regards to the original cell line work, I was unable to pyrosequence the entire ESR1 promoter, largely as a result of the technical difficulties inherent in designing primer sets that can amplify such a large (995 bases) CpG-rich region. Although data was obtained from a 200-base region of the proximal promoter, methylation data from the middle and distal promoter region of *ESR1* could have provided relevant information regarding the transcriptional state of the gene in each tumour cell line and biopsy sample.

The experiments that were designed and conducted to test the various hypotheses to explain how intragenic methylation changes mediated by decitabine (DAC) led to the altered levels of expression in ER-positive cell lines were inconclusive. Further optimisation will be required to assess the presence of antisense transcript expression and cell cycle stages of the cells treated with DAC. Indeed, the 'off-target' effects of DAC on genes that may regulate ESR1 transcription were not investigated, and represent a large possible confounder. Furthermore, although efforts were made to limit the cytotoxic effects of DAC on the tumour cell lines via the MTT assay, non-specific cytotoxic effects of this treatment could have affected the results obtained from this initial body of work. One of the further limitations of this study was that I was unable to establish an assay to analyse 5-hydroxymethylation levels. I attempted to establish in our lab a quantitative pyrosequencing-based method using potassium permanganate (Booth et al., 2013, Booth et al., 2012). However, I lacked sufficient time to develop this assay satisfactorily. This highlights the relative difficulty in assessing other chemical modifications to DNA other than DNA methylation, at least on a large scale. The RT-

qPCR analyses were conducted using an inappropriate housekeeping gene according to the MIQE guidelines, and greater care will be taken in the future to ensure that the correct reference genes are selected for each cell type.

Similar issues for validation using pyrosequencing presented during the smoking study, whereby certain regions flagged from the 450K data as being demethylated in smokers could not be amplified for a range of technical and genomic reasons. Further work should be performed using different techniques to investigate these regions, in order to gain a greater understanding of whether and how their demethylation upon smoke exposure contributes to cancer risk. In addition, only small numbers of mice were available that had been exposed to smoke, as well as human lung tissue samples. Increased sample numbers and a broader range of smoke-exposed and non-exposed tissues will yield a greater understanding of the range of epigenetic responses to this environmental exposure.

With regards to the BECS study, the sample recruitment was limited in scope because of the nature of the collection, through a hospital Milk Bank. Currently, no samples are available from current smokers, and the ethnic and socioeconomic background of the donors is heavily weighted towards professional Caucasians. The development of the nationwide study, both through the Milk Bank network and the community-based recruitment through a web-based system, will aim to broaden the donor population.

In terms of strengths, the work presented in the PhD produced a range of novel findings, which led to four first author publications (full texts in Appendix 9). The EWAS studies in particular generated novel data regarding a methylation signature of smoking exposure that I validated using pyrosequencing, which has

since been replicated by numerous groups, were conducted in accordance with the published recommendations for appropriate design and analyses (Michels et al., 2013).

Furthermore, despite the initial failure of the Marsden mammoplasty epithelial cell extraction technique, I was able to develop the standard operating protocols to extract DNA and RNA from epithelial cells from expressed breast milk. This study extended the understanding of epigenetic profiling for the breast into three different types of cell or tissue: tumour cell lines, breast tumour tissue biopsies and healthy lactating epithelial cells. The data generated from this novel bioresource not only validated the cell line intragenic methylation data, but preliminary data indicated that smoking led to a discernible change in methylation in the breast epithelial cell. Further work will continue to confirm these findings in a larger dataset, and extend our knowledge of the epigenetic changes that occur in the breast epithelial cell in response to environmental exposures, aging and genetic background, including breast cancer risk-associated mutations. Through the establishment of a nationwide collection, both national and international networks would increase the numbers of donors recruited into this population-based study. The BECS study will bring together epigenetic data from the cell types of most relevance to breast cancer research, along with genetic and risk factor data, to generate an overall biological understanding of breast that can inform current risk models (Mill and Heijmans, 2013, Almouzni et al., 2014).

5.4 CONCLUSIONS

Epigenetics is still a relatively young field of research, and the mechanistic insights into intragenic methylation are a tiny part of the whole story, and highlighted the methodological deficiencies that should be considered when conducting research into DNA methylation. The work in this thesis regarding environmental exposures, specifically looking at signatures of smoking, represents a new potential gold-standard for the detection of former smoking. Finally, the characterisation of a homogeneous population of breast epithelial cells in expressed breast milk is an exciting advance with potential for the establishment of a large-scale biobank and wide-ranging collaborations.

Although a relatively new field in biology, advances in our understanding of the interplay between early life exposures, nutrition, environmental exposures, genetics and risk are changing the face of cancer research. The work in this PhD lays the foundations for a study that could enable a cancer epigenetic risk signature to be established in breast milk samples from healthy women and validated in blood samples. The increasing numbers of samples that can be investigated ever more cheaply means that population-based studies are within our grasp. Furthermore, increased exposure of breast milk research to the cancer research community will inevitably lead to increased exposure in the general population of the risk reducing benefits of breastfeeding, which is one of the primary aims of UNICEF and the WHO to improve global health. It is rare that so much biology is not understood within a single field, but the composition of breast milk is one such field. Given the interesting components such as HAMLET and other secreted proteins with anti-tumour effects and biomarker potential, this is truly an exciting area of research.

6. REFERENCES

Adomas, A. B., Grimm, S. A., Malone, C., Takaku, M., Sims, J. K. & Wade, P. A. 2014. Breast tumor specific mutation in GATA3 affects physiological mechanisms regulating transcription factor turnover. *BMC Cancer,* 14**,** 278.

Aits, S., Gustafsson, L., Hallgren, O., Brest, P., Gustafsson, M., Trulsson, M., Mossberg, A. K., Simon, H. U., Mograbi, B. & Svanborg, C. 2009. HAMLET (human alpha-lactalbumin made lethal to tumor cells) triggers autophagic tumor cell death. *Int J Cancer,* 124**,** 1008-19.

Aizer, A. A., Wilhite, T. J., Chen, M. H., Graham, P. L., Choueiri, T. K., Hoffman, K. E., Martin, N. E., Trinh, Q. D., Hu, J. C. & Nguyen, P. L. 2014. Lack of reduction in racial disparities in cancer-specific mortality over a 20-year period. *Cancer,* 120**,** 1532-9.

Akinyemiju T, Moore JX, Altekruse SF. 2015. Breast cancer survival in African-American women by hormone receptor subtypes. *Breast Cancer Res Treat,* 153(1):211-8.

Al-Salihi, M., Yu, M., Burnett, D. M., Alexander, A., Samlowski, W. E. & Fitzpatrick, F. A. 2011. The depletion of DNA Methyltransferase-1 and the epigenetic effects of 5-Aza-2'Deoxycytidine (Decitabine) are differentially regulated by cell cycle progression. *Epigenetics.*

Alberg, A. J., Daudt, A., Huang, H. Y., Hoffman, S. C., Comstock, G. W., Helzlsouer, K. J., Strickland, P. T. & Bell, D. A. 2004. N-acetyltransferase 2 (NAT2) genotypes, cigarette smoking, and the risk of breast cancer. *Cancer Detect Prev,* 28**,** 187-93.

Allen-Hall, A. & McNevin, D. 2012. Human tissue preservation for disaster victim identification (DVI) in tropical climates. *Forensic Sci Int Genet,* 6**,** 653-7.

Almouzni, G., Altucci, L., Amati, B., Ashley, N., Baulcombe, D., Beaujean, N., Bock, C., Bongcam-Rudloff, E., Bousquet, J., Braun, S., Paillerets, B., Bussemakers, M., Clarke, L., Conesa, A., Estivill, X., Fazeli, A., Grgurević, N., Gut, I., Heijmans, B. T., Hermouet, S., Houwing–Duistermaat, J., Iacobucci, I., Ilaš, J., Kandimalla, R., Krauss-Etschmann, S., Lasko, P., Lehmann, S., Lindroth, A., Majdič, G., Marcotte, E., Martinelli, G., Martinet, N., Meyer, E., Miceli, C., Mills, K., Moreno-Villanueva, M., Morvan, G., Nickel, D., Niesler, B., Nowacki, M., Nowak, J., Ossowski, S., Pelizzola, M., Pochet, R., Potočnik, U., Radwanska, M., Raes, J., Rattray, M., Robinson, M. D., Roelen, B., Sauer, S., Schinzer, D., Slagboom, E., Spector, T., Stunnenberg, H. G., Tiligada, E., Torres-Padilla, M. E., Tsonaka, R., Soom, A. V., Vidaković, M. & Widschwendter, M. 2014. Relationship between genome and epigenome - challenges and requirements for future research. *BMC Genomics,* 15.

Aloraifi, F., Boland, M. R., Green, A. J. & Geraghty, J. G. 2015. Gene analysis techniques and susceptibility gene discovery in non-BRCA1/BRCA2 familial breast cancer. *Surg Oncol*.

Amaral, S., Schroth, W., Kugler, S., Fritz, P., Simon, W. & Brauch, H. 2009. The promoter C specific ERalpha isoform is associated with tamoxifen outcome in breast cancer. *Breast Cancer Res Treat,* 118**,** 323-31.

Ambrosone, C. B., Abrams, S. M., Gorlewska-Roberts, K. & Kadlubar, F. F. 2007. Hair dye use, meat intake, and tobacco exposure and presence of carcinogen-DNA adducts in exfoliated breast ductal epithelial cells. *Arch Biochem Biophys,* 464**,** 169-75.

Amir, E., Freedman, O. C., Seruga, B. & Evans, D. G. 2010. Assessing women at high risk of breast cancer: a review of risk assessment models. *J Natl Cancer Inst.* United States.

Amos, C., Antoniou, A., Berchuck, A., Chenevix-Trench, G., Couch, F., Eeles, R., Esserman, L., Gayther, S., Goh, C., Goldgar, D., Gruber, S., Haiman, C., Hall, P., Hunter, D., Kote-Jarai, Z., Lepage, P., Lindstrom, S., McKay, J., Milne, R., Peters, U., Pharoah, P., Phelan, C., Schumacher, F., Sellers, T., Simard, J., Wand, Z., Seminara, D., Chanock, S., Easton, D. & Henderson, B. A comprehensive genetic analysis of common cancer risk through the development of the Oncochip. American Society of Human Genetics, 2013 Boston.

Anbazhagan, R., Bartek, J., Monaghan, P. & Gusterson, B. A. 1991. Growth and development of the human infant breast. *Am J Anat,* 192**,** 407-17.

Anbazhagan, R., Osin, P. P., Bartkova, J., Nathan, B., Lane, E. B. & Gusterson, B. A. 1998. The development of epithelial phenotypes in the human fetal and infant breast. *J Pathol,* 184**,** 197-206.

Anderson, W. F., Rosenberg, P. S., Petito, L., Katki, H. A., Ejlertsen, B., Ewertz, M., Rasmussen, B. B., Jensen, M. B. & Kroman, N. 2013. Divergent estrogen receptor-positive and -negative breast cancer trends and etiologic heterogeneity in Denmark. *Int J Cancer,* 133**,** 2201-6.

Anothaisintawee, T., Wiratkapun, C., Lerdsitthichai, P., Kasamesup, V., Wongwaisayawan, S., Srinakarin, J., Hirunpat, S., Woodtichartpreecha, P., Boonlikit, S., Teerawattananon, Y. & Thakkinstian, A. 2013. Risk factors of breast cancer: a systematic review and meta-analysis. *Asia Pac J Public Health,* 25**,** 368-87.

Antoine, C., Ameye, L., Paesmans, M. & Rozenberg, S. 2014. Systematic review about breast cancer incidence in relation to hormone replacement therapy use. *Climacteric,* 17**,** 116-32.

Antoniou, A. C., Hardy, R., Walker, L., Evans, D. G., Shenton, A., Eeles, R., Shanley, S., Pichert, G., Izatt, L., Rose, S., Douglas, F., Eccles, D., Morrison, P. J., Scott, J., Zimmern, R. L., Easton, D. F. & Pharoah, P. D. 2008. Predicting the likelihood of carrying a BRCA1 or BRCA2 mutation: validation of BOADICEA, BRCAPRO, IBIS, Myriad and the Manchester scoring system using data from UK genetics clinics. *J Med Genet.* England.

Appanah, R., Dickerson, D. R., Goyal, P., Groudine, M. & Lorincz, M. C. 2007. An unmethylated 3' promoter-proximal region is required for efficient transcription initiation. *PLoS Genet,* 3**,** e27.

Appleton, K., Mackay, H. J., Judson, I., Plumb, J. A., McCormick, C., Strathdee, G., Lee, C., Barrett, S., Reade, S., Jadayel, D., Tang, A., Bellenger, K., Mackay, L., Setanoians, A., Schatzlein, A., Twelves, C., Kaye, S. B. & Brown, R. 2007. Phase I and pharmacodynamic trial of the DNA methyltransferase inhibitor decitabine and carboplatin in solid tumors. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 25**,** 4603-4609.

Aran, D., Toperoff, G., Rosenberg, M. & Hellman, A. 2011. Replication timing-related and gene body-specific methylation of active human genes. *Hum Mol Genet.* England.

Arcaro, K. F. & Anderton, D. L. 2008. Potential of using breast milk as a tool to study breast cancer and breast cancer risk. *Future Oncol,* 4**,** 595-7.

Archey, W. B., McEachern, K. A., Robson, M., Offit, K., Vaziri, S. A., Casey, G., Borg, A. & Arrick, B. A. 2002. Increased CpG methylation of the estrogen receptor gene in BRCA1-linked estrogen receptor-negative breast cancers. *Oncogene,* 21**,** 7034-41.

Armogida, S. A., Yannaras, N. M., Melton, A. L. & Srivastava, M. D. 2004. Identification and quantification of innate immune system mediators in human breast milk. *Allergy Asthma Proc,* 25**,** 297-304.

Arslanoglu, S., Ziegler, E. E. & Moro, G. E. 2010. Donor human milk in preterm infant feeding: evidence and recommendations. *J Perinat Med,* 38**,** 347-51.

Baba, T., Mimura, J., Gradin, K., Kuroiwa, A., Watanabe, T., Matsuda, Y., Inazawa, J., Sogawa, K. & Fujii-Kuriyama, Y. 2001. Structure and expression of the Ah receptor repressor gene. *J Biol Chem.* United States.

Badal, S. & Delgoda, R. 2014. Role of the modulation of CYP1A1 expression and activity in chemoprevention. *J Appl Toxicol,* 34**,** 743-53.

Bagnardi, V., Rota, M., Botteri, E., Tramacere, I., Islami, F., Fedirko, V., Scotti, L., Jenab, M., Turati, F., Pasquali, E., Pelucchi, C., Galeone, C., Bellocco, R., Negri, E., Corrao, G., Boffetta, P. & La Vecchia, C. 2015. Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta-analysis. *Br J Cancer,* 112**,** 580-93.

Baheiraei, A., Shamsi, A., Khaghani, S., Shams, S., Chamari, M., Boushehri, H. & Khedri, A. 2014. The effects of maternal passive smoking on maternal milk lipid. *Acta Med Iran,* 52**,** 280-5.

Baird, R. D. & Caldas, C. 2013. Genetic heterogeneity in breast cancer: the road to personalized medicine? *BMC Med,* 11**,** 151.

Baker, J. C., Jr., Ostrander, J. H., Lem, S., Broadwater, G., Bean, G. R., D'Amato, N. C., Goldenberg, V. K., Rowell, C., Ibarra-Drendall, C., Grant, T., Pilie, P. G., Vasilatos, S. N., Troch, M. M., Scott, V., Wilke, L. G., Paisie, C., Rabiner, S. M., Torres-Hernandez, A., Zalles, C. M. & Seewaldt, V. L. 2008. ESR1 promoter hypermethylation does not predict atypia in RPFNA nor persistent atypia after 12 months tamoxifen chemoprevention. *Cancer Epidemiol Biomarkers Prev,* 17**,** 1884-90.

Ball, M. P., Li, J. B., Gao, Y., Lee, J. H., LeProust, E. M., Park, I. H., Xie, B., Daley, G. Q. & Church, G. M. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol,* 27**,** 361-8.

Ballard, O. & Morrow, A. L. 2013. Human milk composition: nutrients and bioactive factors. *Pediatr Clin North Am,* 60**,** 49-74.

Baltzer, A., Svanborg, C. & Jaggi, R. 2004. Apoptotic cell death in the lactating mammary gland is enhanced by a folding variant of alpha-lactalbumin. *Cell Mol Life Sci,* 61**,** 1221-8.

Barker, D. J. P., Osmond, C., Winter, P. D., Margetts, B. & Simmonds, S. J. 1989. WEIGHT IN INFANCY AND DEATH FROM ISCHAEMIC HEART DISEASE. *The Lancet,* 334**,** 577-580.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. & Zhao, K. 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell,* 129**,** 823-837.

Baumann, A., Gjerde, A. U., Ying, M., Svanborg, C., Holmsen, H., Glomm, W. R., Martinez, A. & Halskau, O. 2012. HAMLET forms annular oligomers when deposited with phospholipid monolayers. *J Mol Biol,* 418**,** 90-102.

Baylin, S., Herman, J., Graff, J., Vertino, P. & Issa, J. 1998. Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv Cancer Res,* 72**,** 141-96.

Baylin, S. B., Esteller, M., Rountree, M. R., Bachman, K. E., Schuebel, K. & Herman, J. G. 2001. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum Mol Genet,* 10**,** 687-92.

Beatson, G. 1896. On the treatment of inoperable cases of carcinoma of the mamma: suggestions for a new method of treatment, with illustrative cases. *The Lancet,* 148**,** 162-165.

Bell, E. 2013. *Investigation of the epigenetic effects of intragenic methylation in breast cancer.* MSc, Imperial College.

Bennett, C. M., Coleman, H. G., Veal, P. G., Cantwell, M. M., Lau, C. C. & Murray, L. J. 2015. Lifestyle factors and small intestine adenocarcinoma risk: A systematic review and meta-analysis. *Cancer Epidemiol*.

Benowitz, N. L. 2008. Clinical pharmacology of nicotine: implications for understanding, preventing, and treating tobacco addiction. *Clin Pharmacol Ther,* 83**,** 531-41.

Benowitz, N. L., Hukkanen, J. & Jacob, P., 3rd 2009. Nicotine chemistry, metabolism, kinetics and biomarkers. *Handb Exp Pharmacol***,** 29-60.

Bentz, J. 2009. The role of nipple-aspirate fluid cytology in breast-cancer risk screening. *MLO Med Lab Obs,* 41**,** 8, 10, 12-7.

Beral, V. 2003. Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet.* England.

Berry, D. A. & Ravdin, P. M. 2007. Breast cancer trends: a marriage between clinical trial evidence and epidemiology. *J Natl Cancer Inst.* United States.

Bertino, E., Giuliani, F., Baricco, M., Di Nicola, P., Peila, C., Vassia, C., Chiale, F., Pirra, A., Cresi, F., Martano, C. & Coscia, A. 2013. Benefits of donor milk in the feeding of preterm infants. *Early Hum Dev*.

Bird, A. P. 1986. CpG-rich islands and the function of DNA methylation. *Nature,* 321**,** 209-13.

Blattler, A., Yao, L., Witt, H., Guo, Y., Nicolet, C. M., Berman, B. P. & Farnham, P. J. 2014. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biol,* 15**,** 469.

Bombard, Y., Rozmovits, L., Trudeau, M., Leighl, N. B., Deal, K. & Marshall, D. A. 2015. The value of personalizing medicine: medical oncologists' views on gene expression profiling in breast cancer treatment. *Oncologist,* 20**,** 351-6.

Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W. & Balasubramanian, S. 2012. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science,* 336**,** 934-7.

Booth, M. J., Ost, T. W., Beraldi, D., Bell, N. M., Branco, M. R., Reik, W. & Balasubramanian, S. 2013. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc,* 8**,** 1841-51.

Boyd, C. A., Quigley, M. A. & Brocklehurst, P. 2007. Donor breast milk versus infant formula for preterm infants: systematic review and meta-analysis. *Arch Dis Child Fetal Neonatal Ed,* 92**,** F169-75.

Boyd, S. 1897. On Oophorectomy in the Treatment of Cancer. *Br Med J,* 2**,** 890-896.

Boyd, S. 1899. Remarks on Oophorectomy in the Treatment of Cancer of the Breast. *Br Med J,* 1**,** 257-262.

Brand, J. S., Humphreys, K., Thompson, D. J., Li, J., Eriksson, M., Hall, P. & Czene, K. 2014. Volumetric mammographic density: heritability and association with breast cancer susceptibility loci. *J Natl Cancer Inst,* 106.

Brawley, O. W. 2013. Health disparities in breast cancer. *Obstet Gynecol Clin North Am,* 40**,** 513-23.

Breitling, L. P., Yang, R., Korn, B., Burwinkel, B. & Brenner, H. 2011. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet,* 88**,** 450-7.

Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A. J., Socci, N. D. & Scandura, J. M. 2011. DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One,* 6**,** e14524.

Brennan, K. & Flanagan, J. M. 2012. Is there a link between genome-wide hypomethylation in blood and cancer risk? *Cancer Prev Res (Phila),* 5**,** 1345-57.

Brennan, K., Garcia-Closas, M., Orr, N., Fletcher, O., Jones, M., Ashworth, A., Swerdlow, A., Thorne, H., Riboli, E., Vineis, P., Dorronsoro, M., Clavel-Chapelon, F., Panico, S., Onland-Moret, N. C., Trichopoulos, D., Kaaks, R., Khaw, K. T., Brown, R. & Flanagan, J. M. 2012. Intragenic ATM methylation in peripheral blood DNA as a biomarker of breast cancer risk. *Cancer Res,* 72**,** 2304-13.

Breslow, R. A., Chen, C. M., Graubard, B. I. & Mukamal, K. J. 2011. Prospective study of alcohol consumption quantity and frequency and cancer-specific mortality in the US population. *Am J Epidemiol,* 174**,** 1044-53.

Brinton, L. A., Hoover, R. & Fraumeni, J. F., Jr. 1983. Reproductive factors in the aetiology of breast cancer. *Br J Cancer,* 47**,** 757-62.

Browne, E. P., Punska, E. C., Lenington, S., Otis, C. N., Anderton, D. L. & Arcaro, K. F. 2011. Increased promoter methylation in exfoliated breast epithelial cells in women with a previous breast biopsy. *Epigenetics,* 6**,** 1425-35.

Brzozowski, A. M., Pike, A. C., Dauter, Z., Hubbard, R. E., Bonn, T., Engstrom, O., Ohman, L., Greene, G. L., Gustafsson, J. A. & Carlquist, M. 1997. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature,* 389**,** 753-8.

Bustin, S. A. 2010. Why the need for qPCR publication guidelines?--The case for MIQE. *Methods,* 50**,** 217-26.

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J. & Wittwer, C. T. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem,* 55**,** 611-22.

Buyru, N., Altinisik, J., Ozdemir, F., Demokan, S. & Dalay, N. 2009. Methylation profiles in breast cancer. *Cancer Invest,* 27**,** 307-12.

Cancer, C. G. o. H. F. i. B. 2002. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. *Lancet,* 360**,** 187-95.

Cancer_Genome_Atlas_Network 2012. Comprehensive molecular portraits of human breast tumours. *Nature,* 490**,** 61-70.

Carbone, D. 1992. Smoking and cancer. *Am J Med,* 93**,** 13S-17S.

Carey, L. A. 2010. Through a glass darkly: advances in understanding breast cancer biology, 2000-2010. *Clin Breast Cancer,* 10**,** 188-95.

Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoute, J., Brodsky, A. S., Keeton, E. K., Fertuck, K. C., Hall, G. F., Wang, Q., Bekiranov, S., Sementchenko, V., Fox, E. A., Silver, P. A., Gingeras, T. R., Liu, X. S. & Brown, M. 2006. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet,* 38**,** 1289-97.

Cedar, H. & Bergman, Y. 2008. Epigenetic silencing during early lineage commitment.

Cedar, H. & Bergman, Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet,* 10**,** 295-304.

Chan, T. A., Glockner, S., Yi, J. M., Chen, W., Van Neste, L., Cope, L., Herman, J. G., Velculescu, V., Schuebel, K. E., Ahuja, N. & Baylin, S. B. 2008. Convergence of mutation and epigenetic alterations identifies common genes in cancer that predict for poor prognosis. *PLoS Med.* United States.

Chen, Y. A., Lemire, M., Choufani, S., Butcher, D. T., Grafodatskaya, D., Zanke, B. W., Gallinger, S., Hudson, T. J. & Weksberg, R. 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics,* 8**,** 203-9.

Chiba, T., Uchi, H., Yasukawa, F. & Furue, M. 2011. Role of the arylhydrocarbon receptor in lung disease. *Int Arch Allergy Immunol.* Switzerland: Basel.

Christensen, B. C., Kelsey, K. T., Zheng, S., Houseman, E. A., Marsit, C. J., Wrensch, M. R., Wiemels, J. L., Nelson, H. H., Karagas, M. R., Kushi, L. H., Kwan, M. L. & Wiencke, J. K. 2010. Breast cancer DNA methylation profiles are associated with tumor size and alcohol and folate intake. *PLoS Genet,* 6**,** e1001043.

Cleator, S. J., Ahamed, E., Coombes, R. C. & Palmieri, C. 2009. A 2009 update on the treatment of patients with hormone receptor-positive breast cancer. *Clin Breast Cancer,* 9 Suppl 1**,** S6-S17.

Cole, P. & MacMahon, B. 1969. Oestrogen fractions during early reproductive life in the aetiology of breast cancer. *Lancet,* 1**,** 604-6.

Consortium, T. E. P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

Cotterchio, M., Nadalin, V. & Sauer, M. 2002. Human breast cancer and lymphomas may share a common aetiology involving Mouse Mammary Tumour Virus (MMTV). *Med Hypotheses,* 59**,** 492-4.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G. & Collins, F. S. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* United States.

Creighton, C. J. 2012. The molecular profile of luminal B breast cancer. *Biologics,* 6**,** 289-97.

Dahl, C., Gronbaek, K. & Guldberg, P. 2011. Advances in DNA methylation: 5-hydroxymethylcytosine revisited. *Clin Chim Acta.* 2010. Published by Elsevier B.V.

Davis, M. K., Savitz, D. A. & Graubard, B. I. 1988. Infant feeding and childhood cancer. *Lancet,* 2**,** 365-8.

De Summa, S., Pinto, R., Sambiasi, D., Petriella, D., Paradiso, V., Paradiso, A. & Tommasi, S. 2013. BRCAness: a deeper insight into basal-like breast tumors. *Ann Oncol,* 24 Suppl 8**,** viii13-viii21.

Deaton, A. M. & Bird, A. 2011. CpG islands and the regulation of transcription. *Genes Dev.* United States.

Dedeurwaerder, S., Desmedt, C., Calonne, E., Singhal, S. K., Haibe-Kains, B., Defrance, M., Michiels, S., Volkmar, M., Deplus, R., Luciani, J., Lallemand, F., Larsimont, D., Toussaint, J., Haussy, S., Rothe, F., Rouas, G., Metzger, O., Majjaj, S., Saini, K., Putmans, P., Hames, G., van Baren, N., Coulie, P. G., Piccart, M., Sotiriou, C. & Fuks, F. 2011. DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol Med,* 3**,** 726-41.

Densmore, L. & Pflueger, S. 2006. Genetic analysis of human milk by fluorescence in situ hybridization. *Cancer Genet Cytogenet,* 164**,** 148-51.

DiFranza, J. R. & Lew, R. A. 1995. Effect of maternal cigarette smoking on pregnancy complications and sudden infant death syndrome. *J Fam Pract,* 40**,** 385-94.

Di Santi A, Cernera G, Migliaccio A, Perillo B. 2014. Analysis of histone posttranslational modifications in the control of chromatin plasticity observed at estrogen-responsive sites in human breast cancer cells. *Methods Mol Biol,* 1204:59-69.

Doerfler, W. 2008. In pursuit of the first recognized epigenetic signal--DNA methylation: a 1976 to 2008 synopsis. *Epigenetics : official journal of the DNA Methylation Society,* 3**,** 125-133.

Dogan, M. V., Shields, B., Cutrona, C., Gao, L., Gibbons, F. X., Simons, R., Monick, M., Brody, G. H., Tan, K., Beach, S. R. & Philibert, R. A. 2014. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. *BMC Genomics.* England.

Doisneau-Sixou, S., Sergio, C., Carroll, J., Hui, R., Musgrove, E. & Sutherland, R. 2003. Estrogen and antiestrogen regulation of cell cycle progression in breast cancer cells. *Endocr Relat Cancer,* 10**,** 179-86.

Doll, R. & Hill, A. B. 1954. The mortality of doctors in relation to their smoking habits; a preliminary report. *Br Med J,* 1**,** 1451-5.

Dooley, W. C., Ljung, B. M., Veronesi, U., Cazzaniga, M., Elledge, R. M., O'Shaughnessy, J. A., Kuerer, H. M., Hung, D. T., Khan, S. A., Phillips, R. F., Ganz, P. A., Euhus, D. M., Esserman, L. J., Haffty, B. G., King, B. L., Kelley, M. C., Anderson, M. M., Schmit, P. J., Clark, R. R., Kass, F. C., Anderson, B. O., Troyan, S. L., Arias, R. D., Quiring, J. N., Love, S. M., Page, D. L. & King, E. B. 2001. Ductal lavage for detection of cellular atypia in women at high risk for breast cancer. *J Natl Cancer Inst,* 93**,** 1624-32.

Dossus, L., Boutron-Ruault, M. C., Kaaks, R., Gram, I. T., Vilier, A., Fervers, B., Manjer, J., Tjonneland, A., Olsen, A., Overvad, K., Chang-Claude, J., Boeing, H., Steffen, A., Trichopoulou, A., Lagiou, P., Sarantopoulou, M., Palli, D., Berrino, F., Tumino, R., Vineis, P., Mattiello, A., Bueno-de-Mesquita, H. B., van Duijnhoven, F. J., Bakker, M. F., Peeters, P. H., Weiderpass, E., Bjerkaas, E., Braaten, T., Menéndez, V., Agudo, A., Sanchez, M. J., Amiano, P., Tormo, M. J., Barricarte, A., Butt, S., Khaw, K. T., Wareham, N., Key, T. J., Travis, R. C., Rinaldi, S., McCormack, V., Romieu, I., Cox, D. G., Norat, T., Riboli, E. & Clavel-Chapelon, F. 2014. Active and passive cigarette smoking and breast cancer risk: results from the EPIC cohort. *Int J Cancer,* 134**,** 1871-88.

Duringer, C., Hamiche, A., Gustafsson, L., Kimura, H. & Svanborg, C. 2003. HAMLET interacts with histones and chromatin in tumor cell nuclei. *J Biol Chem,* 278**,** 42131-5.

Eccles, S. A., Aboagye, E. O., Ali, S., Anderson, A. S., Armes, J., Berditchevski, F., Blaydes, J. P., Brennan, K., Brown, N. J., Bryant, H. E., Bundred, N. J., Burchell, J. M., Campbell, A. M., Carroll, J. S., Clarke, R. B., Coles, C. E., Cook, G. J., Cox, A., Curtin, N. J., Dekker, L. V., Silva, I. o. S., Duffy, S. W., Easton, D. F., Eccles, D. M., Edwards, D. R., Edwards, J., Evans, D., Fenlon, D. F., Flanagan, J. M., Foster, C., Gallagher, W. M., Garcia-Closas, M., Gee, J. M., Gescher, A. J., Goh, V., Groves, A. M., Harvey, A. J., Harvie, M., Hennessy, B. T., Hiscox, S., Holen, I., Howell, S. J., Howell, A., Hubbard, G., Hulbert-Williams, N., Hunter, M. S., Jasani, B., Jones, L. J., Key, T. J., Kirwan, C. C., Kong, A., Kunkler, I. H., Langdon, S. P., Leach, M. O., Mann, D. J., Marshall, J. F., Martin, L., Martin, S. G., Macdougall, J. E., Miles, D. W., Miller, W. R., Morris, J. R., Moss, S. M., Mullan, P., Natrajan, R., O'Connor, J. P., O'Connor, R., Palmieri, C., Pharoah, P. D., Rakha, E. A., Reed, E., Robinson, S. P., Sahai, E., Saxton, J. M., Schmid, P., Smalley, M. J., Speirs, V., Stein, R., Stingl, J., Streuli, C. H., Tutt, A. N., Velikova, G., Walker, R.

A., Watson, C. J., Williams, K. J., Young, L. S. & Thompson, A. M. 2013. Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. *Breast Cancer Res,* 15**,** R92.

Ekmektzoglou, K. A., Xanthos, T., German, V. & Zografos, G. C. 2009. Breast cancer: from the earliest times through to the end of the 20th century. *Eur J Obstet Gynecol Reprod Biol,* 145**,** 3-8.

El Saghir, N. S. 2008. Responding to the challenges of breast cancer in egypt and other arab countries. *J Egypt Natl Canc Inst,* 20**,** 309-12.

El Saghir, N. S., Khalil, M. K., Eid, T., El Kinge, A. R., Charafeddine, M., Geara, F., Seoud, M. & Shamseddine, A. I. 2007. Trends in epidemiology and management of breast cancer in developing Arab countries: a literature and registry analysis. *Int J Surg,* 5**,** 225-33.

Elliott, H. R., Tillin, T., McArdle, W. L., Ho, K., Duggirala, A., Frayling, T. M., Davey Smith, G., Hughes, A. D., Chaturvedi, N. & Relton, C. L. 2014. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin Epigenetics,* 6**,** 4.

Eltom, S., Stevenson, C. S., Rastrick, J., Dale, N., Raemdonck, K., Wong, S., Catley, M. C., Belvisi, M. G. & Birrell, M. A. 2011. P2X7 receptor and caspase 1 activation are central to airway inflammation observed after exposure to tobacco smoke. *PLoS One,* 6**,** e24097.

Emerson, R. O., Sherwood, A. M., Rieder, M. J., Guenthoer, J., Williamson, D. W., Carlson, C. S., Drescher, C. W., Tewari, M., Bielas, J. H. & Robins, H. S. 2013. High-throughput sequencing of T-cell receptors reveals a homogeneous repertoire of tumour-infiltrating lymphocytes in ovarian cancer. *J Pathol,* 231**,** 433-40.

Enerly, E., Sheng, Z. & Li, K. B. 2005. Natural antisense as potential regulator of alternative initiation, splicing and termination. *In Silico Biol,* 5**,** 367-77.

Engel, C. & Fischer, C. 2015. Breast cancer risks and risk prediction models. *Breast Care (Basel),* 10**,** 7-12.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. & Bernstein, B. E. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* 473**,** 43-9.

Evans, B. R., Karchner, S. I., Allan, L. L., Pollenz, R. S., Tanguay, R. L., Jenny, M. J., Sherr, D. H. & Hahn, M. E. 2008. Repression of aryl hydrocarbon receptor (AHR) signaling by AHR repressor: role of DNA binding and competition for AHR nuclear translocator. *Mol Pharmacol.* United States.

Evans, R. M., Mangelsdorf, D. J. 2014. Nuclear Receptors, RXR, and the Big Bang. *Cell*, 157; 1: 255-6.

Evron, E., Dooley, W. C., Umbricht, C. B., Rosenthal, D., Sacchi, N., Gabrielson, E., Soito, A. B., Hung, D. T., Ljung, B., Davidson, N. E. & Sukumar, S. 2001. Detection of breast cancer cells in ductal lavage fluid by methylation-specific PCR. *Lancet,* 357**,** 1335-6.

Fabian, C. J., Kimler, B. F., Mayo, M. S. & Khan, S. A. 2005. Breast-tissue sampling for risk assessment and prevention. *Endocr Relat Cancer.* England.

Fan, M., Yan, P. S., Hartman-Frey, C., Chen, L., Paik, H., Oyer, S. L., Salisbury, J. D., Cheng, A. S., Li, L., Abbosh, P. H., Huang, T. H. & Nephew, K. P. 2006. Diverse gene expression and DNA methylation profiles correlate with differential adaptation of breast cancer cells to the antiestrogens tamoxifen and fulvestrant. *Cancer Res,* 66**,** 11954-66.

Fandy, T. E., Carraway, H. & Gore, S. D. 2007. DNA demethylating agents and histone deacetylase inhibitors in hematologic malignancies. *Cancer J.* United States.

Fast, J., Mossberg, A. K., Svanborg, C. & Linse, S. 2005. Stability of HAMLET--a kinetically trapped alpha-lactalbumin oleic acid complex. *Protein Sci,* 14**,** 329-40.

Faupel-Badger, J. M., Arcaro, K. F., Balkam, J. J., Eliassen, A. H., Hassiotou, F., Lebrilla, C. B., Michels, K. B., Palmer, J. R., Schedin, P., Stuebe, A. M., Watson, C. J. & Sherman, M. E. 2013. Postpartum remodeling, lactation, and breast cancer risk: summary of a National Cancer Institute-sponsored workshop. *J Natl Cancer Inst,* 105**,** 166-74.

Feinberg, A. & Vogelstein, B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature,* 301**,** 89-92.

Feinberg, A. P. & Tycko, B. 2004. The history of cancer epigenetics. *Nat Rev Cancer,* 4**,** 143-53.
Feng, S., Cokus, S. J., Zhang, X., Chen, P. Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., Ukomadu, C., Sadler, K. C., Pradhan, S., Pellegrini, M. & Jacobsen, S. E. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A,* 107**,** 8689-94.
Feng, W., Shen, L., Wen, S., Rosen, D. G., Jelinek, J., Hu, X., Huan, S., Huang, M., Liu, J., Sahin, A. A., Hunt, K. K., Bast, R. C., Jr., Shen, Y., Issa, J. P. & Yu, Y. 2007. Correlation between CpG methylation profiles and hormone receptor status in breast cancers. *Breast Cancer Res,* 9**,** R57.
Ferguson, A. T., Lapidus, R. G., Baylin, S. B. & Davidson, N. E. 1995. Demethylation of the estrogen receptor gene in estrogen receptor-negative breast cancer cells can reactivate estrogen receptor gene expression. *Cancer Res,* 55**,** 2279-83.
Ferguson, A. T., Vertino, P. M., Spitzner, J. R., Baylin, S. B., Muller, M. T. & Davidson, N. E. 1997. Role of estrogen receptor gene demethylation and DNA methyltransferase.DNA adduct formation in 5-aza-2'deoxycytidine-induced cytotoxicity in human breast cancer cells. *J Biol Chem,* 272**,** 32260-6.
Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J. W., Comber, H., Forman, D. & Bray, F. 2013. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer,* 49**,** 1374-403.
Fischer, W., Gustafsson, L., Mossberg, A. K., Gronli, J., Mork, S., Bjerkvig, R. & Svanborg, C. 2004. Human alpha-lactalbumin made lethal to tumor cells (HAMLET) kills human glioblastoma cells in brain xenografts by an apoptosis-like mechanism and prolongs survival. *Cancer Res,* 64**,** 2105-12.
Flanagan, J. M. 2011. DNA methylation profiling to identify known and novel tumour subtypes. *In:* Watanabe, H. S. (ed.) *Horizons in Cancer Research.* Nova Science Publishers.
Flanagan, J. M., Brook, M. N., Orr, N., Tomczyk, K., Coulson, P., Fletcher, O., Jones, M. E., Schoemaker, M. J., Ashworth, A., Swerdlow, A., Brown, R. & Garcia-Closas, M. 2015. Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study. *Cancer Epidemiol Biomarkers Prev,* 24**,** 221-9.
Flanagan, J. M., Cocciardi, S., Waddell, N., Johnstone, C. N., Marsh, A., Henderson, S., Simpson, P., da Silva, L., Khanna, K., Lakhani, S., Boshoff, C. & Chenevix-Trench, G. 2010. DNA methylome of familial breast cancer identifies distinct profiles defined by mutation status. *Am J Hum Genet,* 86**,** 420-33.
Flanagan, J. M., Munoz-Alegre, M., Henderson, S., Tang, T., Sun, P., Johnson, N., Fletcher, O., Dos Santos Silva, I., Peto, J., Boshoff, C., Narod, S. & Petronis, A. 2009. Gene-body hypermethylation of ATM in peripheral blood DNA of bilateral breast cancer patients. *Human molecular genetics,* 18**,** 1332-1342.
Flanagan, J. M., Popendikyte, V., Pozdniakovaite, N., Sobolev, M., Assadzadeh, A., Schumacher, A., Zangeneh, M., Lau, L., Virtanen, C., Wang, S. C. & Petronis, A. 2006. Intra- and interindividual epigenetic variation in human germ cells. *American Journal of Human Genetics,* 79**,** 67-84.
Florescu, A., Ferrence, R., Einarson, T., Selby, P., Soldin, O. & Koren, G. 2009. Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus on developmental toxicology. *Ther Drug Monit,* 31**,** 14-30.
Ford, D., Easton, D. F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D. T., Weber, B., Lenoir, G., Chang-Claude, J., Sobol, H., Teare, M. D., Struewing, J., Arason, A., Scherneck, S., Peto, J., Rebbeck, T. R., Tonin, P., Neuhausen, S., Barkardottir, R., Eyfjord, J., Lynch, H., Ponder, B. A., Gayther, S. A. & Zelada-Hedman, M. 1998. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet,* 62**,** 676-89.
Foulkes, W. D. & Shuen, A. Y. 2013. In brief: BRCA1 and BRCA2. *J Pathol,* 230**,** 347-9.

Freudenheim, J. L., Marshall, J. R., Graham, S., Laughlin, R., Vena, J. E., Bandera, E., Muti, P., Swanson, M. & Nemoto, T. 1994. Exposure to breastmilk in infancy and the risk of breast cancer. *Epidemiology,* 5**,** 324-31.

Fu, S., Hu, W., Iyer, R., Kavanagh, J. J., Coleman, R. L., Levenback, C. F., Sood, A. K., Wolf, J. K., Gershenson, D. M., Markman, M., Hennessy, B. T., Kurzrock, R. & Bast, R. C., Jr. 2011. Phase 1b-2a study to reverse platinum resistance through use of a hypomethylating agent, azacitidine, in patients with platinum-resistant or platinum-refractory epithelial ovarian cancer. *Cancer,* 117**,** 1661-9.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. & Mulvihill, J. J. 1989. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute,* 81**,** 1879-1886.

Gamaleddin, I. H., Trigo, J. M., Gueye, A. B., Zvonok, A., Makriyannis, A., Goldberg, S. R. & Le Foll, B. 2015. Role of the endogenous cannabinoid system in nicotine addiction: novel insights. *Front Psychiatry,* 6**,** 41.

Gardiner-Garden, M. & Frommer, M. 1987. CpG islands in vertebrate genomes. *J Mol Biol,* 196**,** 261-82.

Gately, I. 2001. Tobacco: A Cultural History of How an Exotic Plant Seduced Civilization. London: Simon and Schuster.

Gathani, T., Ali, R., Balkwill, A., Green, J., Reeves, G., Beral, V. & Moser, K. A. 2014. Ethnic differences in breast cancer incidence in England are due to differences in known risk factors for the disease: prospective study. *Br J Cancer,* 110**,** 224-9.

Gaudet, M. M., Campan, M., Figueroa, J. D., Yang, X. R., Lissowska, J., Peplonska, B., Brinton, L. A., Rimm, D. L., Laird, P. W., Garcia-Closas, M. & Sherman, M. E. 2009. DNA hypermethylation of ESR1 and PGR in breast cancer: pathologic and epidemiologic associations. *Cancer Epidemiol Biomarkers Prev,* 18**,** 3036-43.

Gerlinger, M., Quezada, S. A., Peggs, K. S., Furness, A. J., Fisher, R., Marafioti, T., Shende, V. H., McGranahan, N., Rowan, A. J., Hazell, S., Hamm, D., Robins, H. S., Pickering, L., Gore, M., Nicol, D. L., Larkin, J. & Swanton, C. 2013. Ultra-deep T cell receptor sequencing reveals the complexity and intratumour heterogeneity of T cell clones in renal cell carcinomas. *J Pathol,* 231**,** 424-32.

Geyer, F. C., Rodrigues, D. N., Weigelt, B. & Reis-Filho, J. S. 2012. Molecular classification of estrogen receptor-positive/luminal breast cancers. *Adv Anat Pathol,* 19**,** 39-53.

Glantz, S. A. & Johnson, K. C. 2014. The surgeon general report on smoking and health 50 years later: breast cancer and the cost of increasing caution. *Cancer Epidemiol Biomarkers Prev,* 23**,** 37-46.

Glass, A. G., Lacey, J. V., Carreon, J. D. & Hoover, R. N. 2007. Breast cancer incidence, 1980-2006: combined roles of menopausal hormone therapy, screening mammography, and estrogen receptor status. *J Natl Cancer Inst,* 99**,** 1152-61.

Glasspool, R. M., Brown, R., Gore, M. E., Rustin, G. J., McNeish, I. A., Wilson, R. H., Pledge, S., Paul, J., Mackean, M., Hall, G. D., Gabra, H., Halford, S. E., Walker, J., Appleton, K., Ullah, R., Kaye, S. & Group, S. G. T. 2014. A randomised, phase II trial of the DNA-hypomethylating agent 5-aza-2'-deoxycytidine (decitabine) in combination with carboplatin vs carboplatin alone in patients with recurrent, partially platinum-sensitive ovarian cancer. *Br J Cancer,* 110**,** 1923-9.

Gluz, O., Hofmann, D., Wurstlein, R., Liedtke, C., Nitz, U. & Harbeck, N. 2013. Genomic profiling in luminal breast cancer. *Breast Care (Basel),* 8**,** 414-22.

Go, R. E., Hwang, K. A. & Choi, K. C. 2015. Cytochrome P450 1 family and cancers. *J Steroid Biochem Mol Biol,* 147**,** 24-30.

Goldberg, A. D., Allis, C. D. & Bernstein, E. 2007. Epigenetics: a landscape takes shape. *Cell,* 128**,** 635-8.

Goll, M. G. & Bestor, T. H. 2005. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem,* 74**,** 481-514.

Gomm, J. J., Browne, P. J., Coope, R. C., Liu, Q. Y., Buluwela, L. & Coombes, R. C. 1995. Isolation of pure populations of epithelial and myoepithelial cells from the normal human mammary gland using immunomagnetic separation with Dynabeads. *Anal Biochem,* 226**,** 91-9.

Gonzalez, C. A. & Riboli, E. 2010. Diet and cancer prevention: Contributions from the European Prospective Investigation into Cancer and Nutrition (EPIC) study. *Eur J Cancer.* England: 2010 Elsevier Ltd.

Gorlewska-Roberts, K., Green, B., Fares, M., Ambrosone, C. B. & Kadlubar, F. F. 2002. Carcinogen-DNA adducts in human breast epithelial cells. *Environ Mol Mutagen,* 39**,** 184-92.

Gram, I. T., Braaten, T., Terry, P. D., Sasco, A. J., Adami, H. O., Lund, E. & Weiderpass, E. 2005. Breast cancer risk among women who start smoking as teenagers. *Cancer Epidemiol Biomarkers Prev,* 14**,** 61-6.

Green, S., Walter, P., Kumar, V., Krust, A., Bornert, JM., Argos, P., Chambon, P. 1986. Human oestrogen receptor cDNA: sequence, expression and homology to v-erb-A. *Nature*, 320(6058):134-9.

Groth, A., Rocha, W., Verreault, A. & Almouzni, G. 2007. Chromatin challenges during DNA replication and repair. *Cell,* 128**,** 721-33.

Guida, F., Sandanger, T. M., Castagne, R., Campanella, G., Polidoro, S., Palli, D., Krogh, V., Tumino, R., Sacerdote, C., Panico, S., Severi, G., Kyrtopoulos, S. A., Georgiadis, P., Vermeulen, R. C., Lund, E., Vineis, P. & Chadeau-Hyam, M. 2015. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Genet,* 24**,** 2349-59.

Gur-Dedeoglu, B., Konu, O., Bozkurt, B., Ergul, G., Seckin, S. & Yulug, I. G. 2009. Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. *Oncol Res,* 17**,** 353-65.

Gustafsson, L., Aits, S., Onnerfjord, P., Trulsson, M., Storm, P. & Svanborg, C. 2009. Changes in proteasome structure and function caused by HAMLET in tumor cells. *PLoS One,* 4**,** e5229.

Gustafsson, L., Hallgren, O., Mossberg, A. K., Pettersson, J., Fischer, W., Aronsson, A. & Svanborg, C. 2005. HAMLET kills tumor cells by apoptosis: structure, cellular mechanisms, and therapy. *J Nutr,* 135**,** 1299-303.

Gustafsson, L., Leijonhufvud, I., Aronsson, A., Mossberg, A. K. & Svanborg, C. 2004. Treatment of skin papillomas with topical alpha-lactalbumin-oleic acid. *N Engl J Med,* 350**,** 2663-72.

Gusterson, B. A. & Stein, T. 2012. Human breast development. *Semin Cell Dev Biol,* 23**,** 567-73.

Hall, P. & Easton, D. 2013. Breast cancer screening: time to target women at risk. *Br J Cancer,* 108**,** 2202-4.

Hallgren, O., Gustafsson, L., Irjala, H., Selivanova, G., Orrenius, S. & Svanborg, C. 2006. HAMLET triggers apoptosis but tumor cell death is independent of caspases, Bcl-2 and p53. *Apoptosis,* 11**,** 221-33.

Hamajima, N., Hirose, K., Tajima, K., Rohan, T., Calle, E. E., Heath, C. W., Coates, R. J., Liff, J. M., Talamini, R., Chantarakul, N., Koetsawang, S., Rachawat, D., Morabia, A., Schuman, L., Stewart, W., Szklo, M., Bain, C., Schofield, F., Siskind, V., Band, P., Coldman, A. J., Gallagher, R. P., Hislop, T. G., Yang, P., Kolonel, L. M., Nomura, A. M., Hu, J., Johnson, K. C., Mao, Y., De Sanjosé, S., Lee, N., Marchbanks, P., Ory, H. W., Peterson, H. B., Wilson, H. G., Wingo, P. A., Ebeling, K., Kunde, D., Nishan, P., Hopper, J. L., Colditz, G., Gajalanski, V., Martin, N., Pardthaisong, T., Silpisornkosol, S., Theetranont, C., Boosiri, B., Chutivongse, S., Jimakorn, P., Virutamasen, P., Wongsrichanalai, C., Ewertz, M., Adami, H. O., Bergkvist, L., Magnusson, C., Persson, I., Chang-Claude, J., Paul, C., Skegg, D. C., Spears, G. F., Boyle, P., Evstifeeva, T., Daling, J. R., Hutchinson, W. B., Malone, K., Noonan, E. A., Stanford, J. L., Thomas, D. B., Weiss, N. S., White, E., Andrieu, N., Brêmond, A., Clavel, F., Gairard, B., Lansac, J., Piana, L., Renaud, R., Izquierdo, A., Viladiu, P., Cuevas, H. R., Ontiveros, P., Palet, A., Salazar, S. B., Aristizabel, N., Cuadros, A., Tryggvadottir, L., Tulinius, H., Bachelot, A., Lê, M. G., Peto, J., Franceschi, S., Lubin, F., Modan, B., Ron, E., Wax, Y., Friedman, G. D., Hiatt, R. A., Levi, F., Bishop, T., Kosmelj, K., et al. 2002. Alcohol, tobacco and breast cancer--collaborative

reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. *Br J Cancer,* 87**,** 1234-45.
Hanahan, D. & Weinberg, R. A. 2000. The hallmarks of cancer. *Cell,* 100**,** 57-70.
Hannan, L. M., Jacobs, E. J. & Thun, M. J. 2009. The association between cigarette smoking and risk of colorectal cancer in a large prospective cohort from the United States. *Cancer Epidemiol Biomarkers Prev.* United States.
Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A. & Feinberg, A. P. 2011. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet.*
Harper, K. N., Peters, B. A. & Gamble, M. V. 2013. Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiol Biomarkers Prev,* 22**,** 1052-60.
Hassiotou, F., Beltran, A., Chetwynd, E., Stuebe, A. M., Twigger, A. J., Metzger, P., Trengove, N., Lai, C. T., Filgueira, L., Blancafort, P. & Hartmann, P. E. 2012. Breastmilk is a novel source of stem cells with multilineage differentiation potential. *Stem Cells,* 30**,** 2164-74.
Hassiotou, F. & Geddes, D. 2013. Anatomy of the human mammary gland: Current status of knowledge. *Clin Anat,* 26**,** 29-48.
Hassiotou, F., Geddes, D. T. & Hartmann, P. E. 2013. Cells in human milk: state of the science. *J Hum Lact,* 29**,** 171-82.
Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E., Kuan, S., Luu, Y., Klugman, S., Antosiewicz-Bourget, J., Ye, Z., Espinoza, C., Agarwahl, S., Shen, L., Ruotti, V., Wang, W., Stewart, R., Thomson, J. A., Ecker, J. R. & Ren, B. 2010. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell,* 6**,** 479-91.
Hegaard, H. K., Kjaergaard, H., Møller, L. F., Wachmann, H. & Ottesen, B. 2006. The effect of environmental tobacco smoke during pregnancy on birth weight. *Acta Obstet Gynecol Scand,* 85**,** 675-81.
Heijmans, B. T. & Mill, J. 2012. Commentary: The seven plagues of epigenetic epidemiology. *Int J Epidemiol,* 41**,** 74-8.
Heijmans, B. T., Tobi, E. W., Stein, A. D., Putter, H., Blauw, G. J., Susser, E. S., Slagboom, P. E. & Lumey, L. H. 2008. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 17046-17049.
Hein, D. W. 2000. N-Acetyltransferase genetics and their role in predisposition to aromatic and heterocyclic amine-induced carcinogenesis. *Toxicol Lett.* Netherlands.
Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M. & Ren, B. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature,* 459**,** 108-12.
Heldring, N., Pike, A., Andersson, S., Matthews, J., Cheng, G., Hartman, J., Tujague, M., Strom, A., Treuter, E., Warner, M. & Gustafsson, J. A. 2007. Estrogen receptors: how do they signal and what are their targets. *Physiol Rev,* 87**,** 905-31.
Hellman, A. & Chess, A. 2007. Gene body-specific methylation on the active X chromosome. *Science,* 315**,** 1141-3.
Herbstman, J. B., Tang, D., Zhu, D., Qu, L., Sjodin, A., Li, Z., Camann, D. & Perera, F. P. 2012. Prenatal Exposure to Polycyclic Aromatic Hydrocarbons, Benzo[a]Pyrene-DNA Adducts and Genomic DNA Methylation in Cord Blood. *Environ Health Perspect*.
Herceg, Z. & Paliwal, A. 2011. Epigenetic mechanisms in hepatocellular carcinoma: How environmental factors influence the epigenome. *Mutat Res.* 2011. Published by Elsevier B.V.
Herman, A. I., DeVito, E. E., Jensen, K. P. & Sofuoglu, M. 2014. Pharmacogenetics of nicotine addiction: role of dopamine. *Pharmacogenomics,* 15**,** 221-34.

Herrmann, M., King, K. & Weitzman, M. 2008. Prenatal tobacco smoke and postnatal secondhand smoke exposure and child neurodevelopment. *Curr Opin Pediatr,* 20**,** 184-90.

Hilton, H. N., Graham, J. D., Kantimm, S., Santucci, N., Cloosterman, D., Huschtscha, L. I., Mote, P. A. & Clarke, C. L. 2012. Progesterone and estrogen receptors segregate into different cell subpopulations in the normal human breast. *Mol Cell Endocrinol,* 361**,** 191-201.

Hilton, H. N., Kantimm, S., Graham, J. D. & Clarke, C. L. 2013. Changed lineage composition is an early event in breast carcinogenesis. *Histol Histopathol,* 28**,** 1197-204.

Ho, C. S. J., Rydstrom, A., Trulsson, M., Balfors, J., Storm, P., Puthia, M., Nadeem, A. & Svanborg, C. 2012. HAMLET: functional properties and therapeutic potential. *Future Oncol,* 8**,** 1301-13.

Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., Park, J., Butler, J., Rafii, S., McCombie, W. R., Smith, A. D. & Hannon, G. J. 2011. Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell.* United States: 2011 Elsevier Inc.

Holliday, R. & Pugh, J. E. 1975. DNA modification mechanisms and gene activity during development. *Science,* 187**,** 226-32.

Holm, K., Hegardt, C., Staaf, J., Vallon-Christersson, J., Jonsson, G., Olsson, H., Borg, A. & Ringner, M. 2010. Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Res.* England.

Hornberger, J., Chen, S. C., Li, Q., Kakad, P. & Quay, S. C. 2015. Proliferative epithelial disease identified in nipple aspirate fluid and risk of developing breast cancer: a systematic review. *Curr Med Res Opin,* 31**,** 253-62.

Horvath, S. 2013. DNA methylation age of human tissues and cell types. *Genome Biol,* 14**,** R115.

Horvath, S., Erhart, W., Brosch, M., Ammerpohl, O., von Schonfels, W., Ahrens, M., Heits, N., Bell, J. T., Tsai, P. C., Spector, T. D., Deloukas, P., Siebert, R., Sipos, B., Becker, T., Rocken, C., Schafmayer, C. & Hampe, J. 2014. Obesity accelerates epigenetic aging of human liver. *Proc Natl Acad Sci U S A*.

Hosoya, T., Harada, N., Mimura, J., Motohashi, H., Takahashi, S., Nakajima, O., Morita, M., Kawauchi, S., Yamamoto, M. & Fujii-Kuriyama, Y. 2008. Inducibility of cytochrome P450 1A1 and chemical carcinogenesis by benzo[a]pyrene in AhR repressor-deficient mice. *Biochem Biophys Res Commun.* United States.

Hou, N. & Huo, D. 2013. A trend analysis of breast cancer incidence rates in the United States from 2000 to 2009 shows a recent increase. *Breast Cancer Res Treat,* 138**,** 633-41.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K. & Kelsey, K. T. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics,* 13**,** 86.

Howell, A., Anderson, A. S., Clarke, R. B., Duffy, S. W., Evans, D. G., Garcia-Closas, M., Gescher, A. J., Key, T. J., Saxton, J. M. & Harvie, M. N. 2014. Risk determination and prevention of breast cancer. *Breast Cancer Res,* 16**,** 446.

HSCIC, H. a. S. C. I. C. *Infant Feeding Survey 2010* [Online]. Available: http://www.hscic.gov.uk/catalogue/PUB08694 [Accessed 28th July 2015].

Huang, Z. & Murphy, S. K. 2013. Increased Intragenic IGF2 Methylation is Associated with Repression of Insulator Activity and Elevated Expression in Serous Ovarian Carcinoma. *Front Oncol,* 3**,** 131.

IARC 2004. Tobacco smoke and involuntary smoking. *IARC Monogr Eval Carcinog Risks Hum,* 83**,** 1-1438.

Ibanez, G., Martin, N., Denantes, M., Saurel-Cubizolles, M. J., Ringa, V. & Magnier, A. M. 2012. Prevalence of breastfeeding in industrialized countries. *Rev Epidemiol Sante Publique,* 60**,** 305-20.

Ince, T. A., Richardson, A. L., Bell, G. W., Saitoh, M., Godar, S., Karnoub, A. E., Iglehart, J. D. & Weinberg, R. A. 2007. Transformation of different human breast epithelial cell types leads to distinct tumor phenotypes. *Cancer Cell,* 12**,** 160-70.

Ip, S., Chung, M., Raman, G., Chew, P., Magula, N., DeVine, D., Trikalinos, T. & Lau, J. 2007. Breastfeeding and maternal and infant health outcomes in developed countries. *Evid Rep Technol Assess (Full Rep)*, 1-186.

Iwase, H., Omoto, Y., Iwata, H., Toyama, T., Hara, Y., Ando, Y., Ito, Y., Fujii, Y. & Kobayashi, S. 1999. DNA methylation analysis at distal and proximal promoter regions of the oestrogen receptor gene in breast cancers. *Br J Cancer,* 80, 1982-6.

Jackson, K. J., Muldoon, P. P., De Biasi, M. & Damaj, M. I. 2014. New mechanisms and perspectives in nicotine withdrawal. *Neuropharmacology*.

Jacobi, C. E., de Bock, G. H., Siegerink, B. & van Asperen, C. J. 2009. Differences and similarities in breast cancer risk assessment models in clinical practice: which model to choose? *Breast Cancer Res Treat,* 115, 381-90.

Jacquemier, J., Ginestier, C., Rougemont, J., Bardou, V. J., Charafe-Jauffret, E., Geneix, J., Adelaide, J., Koki, A., Houvenaeghel, G., Hassoun, J., Maraninchi, D., Viens, P., Birnbaum, D. & Bertucci, F. 2005. Protein expression profiling identifies subclasses of breast cancer and predicts prognosis. *Cancer Res,* 65, 767-79.

Jensen, E. V. 1962. On the mechanism of estrogen action. *Perspect Biol Med,* 6, 47-59.

Jensen, E. V. & Jordan, V. C. 2003. The estrogen receptor: a model for molecular medicine. *Clin Cancer Res,* 9, 1980-9.

Jensen, E. V., Suzuki, T., Kawashima, T., Stumpf, W. E., Jungblut, P. W. & DeSombre, E. R. 1968. A two-step mechanism for the interaction of estradiol with rat uterus. *Proc Natl Acad Sci U S A,* 59, 632-8.

Jernstrom, H., Lubinski, J., Lynch, H. T., Ghadirian, P., Neuhausen, S., Isaacs, C., Weber, B. L., Horsman, D., Rosen, B., Foulkes, W. D., Friedman, E., Gershoni-Baruch, R., Ainsworth, P., Daly, M., Garber, J., Olsson, H., Sun, P. & Narod, S. A. 2004. Breast-feeding and the risk of breast cancer in BRCA1 and BRCA2 mutation carriers. *J Natl Cancer Inst,* 96, 1094-8.

Johnson, K. C. & Glantz, S. A. 2008. Evidence secondhand smoke causes breast cancer in 2005 stronger than for lung cancer in 1986. *Prev Med,* 46, 492-6.

Johnson, K. C., Miller, A. B., Collishaw, N. E., Palmer, J. R., Hammond, S. K., Salmon, A. G., Cantor, K. P., Miller, M. D., Boyd, N. F., Millar, J. & Turcotte, F. 2011. Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk (2009). *Tob Control.* England.

Jones, G., Steketee, R. W., Black, R. E., Bhutta, Z. A. & Morris, S. S. 2003. How many child deaths can we prevent this year? *Lancet,* 362, 65-71.

Jones, P. A. 1999. The DNA methylation paradox. *Trends Genet,* 15, 34-7.

Jones, P. A. & Baylin, S. B. 2002. The fundamental role of epigenetic events in cancer. *Nat Rev Genet,* 3, 415-28.

Jones, P. A. & Baylin, S. B. 2007. The epigenomics of cancer. *Cell,* 128, 683-92.

Joubert, B. R., Håberg, S. E., Bell, D. A., Nilsen, R. M., Vollset, S. E., Midttun, O., Ueland, P. M., Wu, M. C., Nystad, W., Peddada, S. D. & London, S. J. 2014. Maternal smoking and DNA methylation in newborns: in utero effect or epigenetic inheritance? *Cancer Epidemiol Biomarkers Prev,* 23, 1007-17.

Joubert, B. R., Håberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., Huang, Z., Hoyo, C., Midttun, Ø., Cupul-Uicab, L. A., Ueland, P. M., Wu, M. C., Nystad, W., Bell, D. A., Peddada, S. D. & London, S. J. 2012. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect,* 120, 1425-31.

Julia, V., Macia, L. & Dombrowicz, D. 2015. The impact of diet on asthma and allergic diseases. *Nat Rev Immunol,* 15, 308-22.

Jönsson, G., Staaf, J., Vallon-Christersson, J., Ringnér, M., Holm, K., Hegardt, C., Gunnarsson, H., Fagerholm, R., Strand, C., Agnarsson, B. A., Kilpivaara, O., Luts, L., Heikkilä, P., Aittomäki, K., Blomqvist, C., Loman, N., Malmström, P., Olsson, H., Johannsson, O. T., Arason, A.,

Nevanlinna, H., Barkardottir, R. B. & Borg, A. 2010. Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res,* 12**,** R42.

Kamalakaran, S., Varadan, V., Giercksky Russnes, H. E., Levy, D., Kendall, J., Janevski, A., Riggs, M., Banerjee, N., Synnestvedt, M., Schlichting, E., Karesen, R., Shama Prasada, K., Rotti, H., Rao, R., Rao, L., Eric Tang, M. H., Satyamoorthy, K., Lucito, R., Wigler, M., Dimitrova, N., Naume, B., Borresen-Dale, A. L. & Hicks, J. B. 2010. DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables. *Mol Oncol.* 2010 Federation of European Biochemical Societies. Published by Elsevier B.V.

Kanai, Y. 2010. Genome-wide DNA methylation profiles in precancerous conditions and cancers. *Cancer Sci,* 101**,** 36-45.

Kang, Y. & Massague, J. 2004. Epithelial-mesenchymal transitions: twist in development and metastasis. *Cell,* 118**,** 277-9.

Kangaspeska, S., Stride, B., Metivier, R., Polycarpou-Schwarz, M., Ibberson, D., Carmouche, R. P., Benes, V., Gannon, F. & Reid, G. 2008. Transient cyclical methylation of promoter DNA. *Nature,* 452**,** 112-5.

Kanno, Y., Takane, Y., Izawa, T., Nakahama, T. & Inouye, Y. 2006. The inhibitory effect of aryl hydrocarbon receptor repressor (AhRR) on the growth of human breast cancer MCF-7 cells. *Biol Pharm Bull.* Japan.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., Boyle, A. P., Zhang, Q. C., Zakharia, F., Spacek, D. V., Li, J., Xie, D., Olarerin-George, A., Steinmetz, L. M., Hogenesch, J. B., Kellis, M., Batzoglou, S. & Snyder, M. 2013. Extensive variation in chromatin states across humans. *Science,* 342**,** 750-2.

Kastl, L., Brown, I. & Schofield, A. C. 2010. Effects of decitabine on the expression of selected endogenous control genes in human breast cancer cells. *Mol Cell Probes,* 24**,** 87-92.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K. C., Hallinan, J., Mattick, J., Hume, D. A., Lipovich, L., Batalov, S., Engström, P. G., Mizuno, Y., Faghihi, M. A., Sandelin, A., Chalk, A. M., Mottagui-Tabar, S., Liang, Z., Lenhard, B., Wahlestedt, C., Group, R. G. E. R., Group), G. S. G. G. N. P. C. & Consortium, F. 2005. Antisense transcription in the mammalian transcriptome. *Science,* 309**,** 1564-6.

Kawajiri, K. & Fujii-Kuriyama, Y. 2007. Cytochrome P450 gene regulation and physiological functions mediated by the aryl hydrocarbon receptor. *Arch Biochem Biophys,* 464**,** 207-12.

Keen, J. & Davidson, N. 2003. The biology of breast carcinoma. *Cancer,* 97**,** 825-33.

Kendall, A., Folkerd, E. J. & Dowsett, M. 2007. Influences on circulating oestrogens in postmenopausal women: relationship with breast cancer. *J Steroid Biochem Mol Biol,* 103**,** 99-109.

Khoury, M. J., Gomez-Farias, M. & Mulinare, J. 1989. Does maternal cigarette smoking during pregnancy cause cleft lip and palate in offspring? *Am J Dis Child,* 143**,** 333-7.

Khuder, S. A. & Mutgi, A. B. 2000. Reproductive factors are crucial in the aetiology of breast cancer. *Br J Cancer.* Scotland.

Kleinman, J. C. & Madans, J. H. 1985. The effects of maternal smoking, physical stature, and educational attainment on the incidence of low birth weight. *Am J Epidemiol,* 121**,** 843-55.

Kline, J., Stein, Z. & Hutzler, M. 1987. Cigarettes, alcohol and marijuana: varying associations with birthweight. *Int J Epidemiol,* 16**,** 44-51.

Klinge, C. M. 2001. Estrogen receptor interaction with estrogen response elements. *Nucleic Acids Res,* 29**,** 2905-19.

Knight, C. H. & Peaker, M. 1982. Development of the mammary gland. *J Reprod Fertil,* 65**,** 521-36.

Koenig, A., de Albuquerque Diniz, E. M., Barbosa, S. F. & Vaz, F. A. 2005. Immunologic factors in human milk: the effects of gestational age and pasteurization. *J Hum Lact,* 21**,** 439-43.

Komashko, V. M., Farnham, P. J., Acevedo, L. G., Squazzo, S. L., Iyengar, S. S., Rabinovich, A., O'Geen, H. & Green, R. 2010. 5-azacytidine treatment reorganizes genomic histone modification patterns

Using ChIP-chip technology to reveal common principles of transcriptional repression in normal and cancer cells. *Epigenetics.* United States.

Konstandiadou, I., Mastoraki, A., Kotsilianou, O., Karakitsos, P., Athanasas, G., Smyrniotis, V. & Arkadopoulos, N. 2012. Does ductal lavage assert its role as a noninvasive diagnostic modality to identify women at low risk of breast cancer development? *J Gynecol Oncol,* 23**,** 110-4.

Kordon, E. C. & Smith, G. H. 1998. An entire functional mammary gland may comprise the progeny from a single cell. *Development,* 125**,** 1921-30.

Kotsopoulos, J., Lubinski, J., Gronwald, J., Cybulski, C., Demsky, R., Neuhausen, S. L., Kim-Sing, C., Tung, N., Friedman, S., Senter, L., Weitzel, J., Karlan, B., Moller, P., Sun, P. & Narod, S. A. 2014. Factors influencing ovulation and the risk of ovarian cancer in BRCA1 and BRCA2 mutation carriers. *Int J Cancer*.

Kotsopoulos, J., Lubinski, J., Salmena, L., Lynch, H. T., Kim-Sing, C., Foulkes, W. D., Ghadirian, P., Neuhausen, S. L., Demsky, R., Tung, N., Ainsworth, P., Senter, L., Eisen, A., Eng, C., Singer, C., Ginsburg, O., Blum, J., Huzarski, T., Poll, A., Sun, P. & Narod, S. A. 2012. Breastfeeding and the risk of breast cancer in BRCA1 and BRCA2 mutation carriers. *Breast Cancer Res,* 14**,** R42.

Kramer, M. S. & Kakuma, R. 2012. Optimal duration of exclusive breastfeeding. *Cochrane Database Syst Rev,* 8**,** Cd003517.

Kulis, M., Queiros, A. C., Beekman, R. & Martin-Subero, J. I. 2013. Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochim Biophys Acta*.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y. C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K. H., Feizi, S., Karlic, R., Kim, A. R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L. H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., Kellis, M. & Consortium, R. E. 2015. Integrative analysis of 111 reference human epigenomes. *Nature,* 518**,** 317-30.

Kuo, L. C., Cheng, L. C., Lin, C. J. & Li, L. A. 2013. Dioxin and estrogen signaling in lung adenocarcinoma cells with different aryl hydrocarbon receptor/estrogen receptor α phenotypes. *Am J Respir Cell Mol Biol,* 49**,** 1064-73.

Kupers, L. K., Xu, X., Jankipersadsing, S. A., Vaez, A., la Bastide-van Gemert, S., Scholtens, S., Nolte, I. M., Richmond, R. C., Relton, C. L., Felix, J. F., Duijts, L., van Meurs, J. B., Tiemeier, H., Jaddoe, V. W., Wang, X., Corpeleijn, E. & Snieder, H. 2015. DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *Int J Epidemiol*.

Labbok, M. H. 2013. Breastfeeding: population-based perspectives. *Pediatr Clin North Am,* 60**,** 11-30.

Ladd-Acosta, C. 2015. Epigenetic Signatures as Biomarkers of Exposure. *Current Environmental Health Reports,* 2**,** 117-125.

Lagiou, P., Hsieh, C. C., Trichopoulos, D., Adami, H. O., Hall, P., Chie, L. & Ekbom, A. 2008. Neonatal growth and breast cancer risk in adulthood. *Br J Cancer.*

Lancaster, J. M., Wooster, R., Mangion, J., Phelan, C. M., Cochran, C., Gumbs, C., Seal, S., Barfoot, R., Collins, N., Bignell, G., Patel, S., Hamoudi, R., Larsson, C., Wiseman, R. W., Berchuck, A., Iglehart, J. D., Marks, J. R., Ashworth, A., Stratton, M. R. & Futreal, P. A. 1996. BRCA2 mutations in primary breast and ovarian cancers. *Nat Genet,* 13**,** 238-40.

Lapidus, R. G., Ferguson, A. T., Ottaviano, Y. L., Parl, F. F., Smith, H. S., Weitzman, S. A., Baylin, S. B., Issa, J. P. & Davidson, N. E. 1996. Methylation of estrogen and progesterone receptor gene 5' CpG islands correlates with lack of estrogen and progesterone receptor gene expression in breast tumors. *Clin Cancer Res,* 2**,** 805-10.

Lapidus, R. G., Nass, S. J., Butash, K. A., Parl, F. F., Weitzman, S. A., Graff, J. G., Herman, J. G. & Davidson, N. E. 1998. Mapping of ER gene CpG island methylation-specific polymerase chain reaction. *Cancer Res,* 58**,** 2515-9.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D. A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S. & Getz, G. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* 499**,** 214-8.

Lawson, J. S., Glenn, W. K., Salmons, B., Ye, Y., Heng, B., Moody, P., Johal, H., Rawlinson, W. D., Delprado, W., Lutze-Mann, L. & Whitaker, N. J. 2010. Mouse mammary tumor virus-like sequences in human breast cancer. *Cancer Res,* 70**,** 3576-85.

Lawson, J. S., Gunzburg, W. H. & Whitaker, N. J. 2006a. Viruses and human breast cancer. *Future Microbiol,* 1**,** 33-51.

Lawson, J. S., Tran, D. D., Carpenter, E., Ford, C. E., Rawlinson, W. D., Whitaker, N. J. & Delprado, W. 2006b. Presence of mouse mammary tumour-like virus gene sequences may be associated with morphology of specific human breast cancer. *J Clin Pathol,* 59**,** 1287-92.

Leader, J. E., Wang, C., Fu, M. & Pestell, R. G. 2006a. Epigenetic regulation of nuclear steroid receptors. *Biochemical pharmacology,* 72**,** 1589-1596.

Leader, J. E., Wang, C., Popov, V. M., Fu, M. & Pestell, R. G. 2006b. Epigenetics and the estrogen receptor. *Annals of the New York Academy of Sciences,* 1089**,** 73-87.

Leake, R. 1996. 100 years of the endocrine battle against breast cancer. *Lancet,* 347**,** 1780-1.

Lee, J. S., Fackler, M. J., Teo, W. W., Lee, J. H., Choi, C., Park, M. H., Yoon, J. H., Zhang, Z., Argani, P. & Sukumar, S. 2008. Quantitative promoter hypermethylation profiles of ductal carcinoma in situ in North American and Korean women: Potential applications for diagnosis. *Cancer Biol Ther,* 7**,** 1398-406.

Lee, J. S., Lo, P. K., Fackler, M. J., Argani, P., Zhang, Z., Garrett-Meyer, E. & Sukumar, S. 2007. A comparative study of Korean with Caucasian breast cancer reveals frequency of methylation in multiple genes correlates with breast cancer in young, ER, PR-negative breast cancer in Korean women. *Cancer Biol Ther,* 6**,** 1114-20.

Lee, K. W., Richmond, R., Hu, P., French, L., Shin, J., Bourdon, C., Reischl, E., Waldenberger, M., Zeilinger, S., Gaunt, T., McArdle, W., Ring, S., Woodward, G., Bouchard, L., Gaudet, D., Smith, G. D., Relton, C., Paus, T. & Pausova, Z. 2015. Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. *Environ Health Perspect,* 123**,** 193-9.

Li, J., Harris, R. A., Cheung, S. W., Coarfa, C., Jeong, M., Goodell, M. A., White, L. D., Patel, A., Kang, S. H., Shaw, C., Chinault, A. C., Gambin, T., Gambin, A., Lupski, J. R. & Milosavljevic, A. 2012.

Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. *PLoS Genet,* 8**,** e1002692.
Li, M., Liu, W., Yuan, T., Bai, R., Liu, G. H., Zhang, W. & Qu, J. 2013. DNA methylome: Unveiling your biological age. *Protein Cell,* 4**,** 723-5.
Li, S., Rong, M. & Iacopetta, B. 2006. DNA hypermethylation in breast cancer and its association with clinicopathological features. *Cancer Lett,* 237**,** 272-80.
Li, Y., Lambert, M. H. & Xu, H. E. 2003. Activation of nuclear receptors: a perspective from structural genomics. *Structure,* 11**,** 741-6.
Liedtke, C. & Rody, A. 2015. New treatment strategies for patients with triple-negative breast cancer. *Curr Opin Obstet Gynecol,* 27**,** 77-84.
Lim, E., Vaillant, F., Wu, D., Forrest, N. C., Pal, B., Hart, A. H., Asselin-Labat, M. L., Gyorki, D. E., Ward, T., Partanen, A., Feleppa, F., Huschtscha, L. I., Thorne, H. J., Fox, S. B., Yan, M., French, J. D., Brown, M. A., Smyth, G. K., Visvader, J. E., Lindeman, G. J. & kConFab 2009. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med,* 15**,** 907-13.
Limsui, D., Vierkant, R. A., Tillmans, L. S., Wang, A. H., Weisenberger, D. J., Laird, P. W., Lynch, C. F., Anderson, K. E., French, A. J., Haile, R. W., Harnack, L. J., Potter, J. D., Slager, S. L., Smyrk, T. C., Thibodeau, S. N., Cerhan, J. R. & Limburg, P. J. 2010. Cigarette smoking and colorectal cancer risk by molecularly defined subtypes. *J Natl Cancer Inst,* 102**,** 1012-22.
Lindstrom, S., Thompson, D. J., Paterson, A. D., Li, J., Gierach, G. L., Scott, C., Stone, J., Douglas, J. A., dos-Santos-Silva, I., Fernandez-Navarro, P., Verghase, J., Smith, P., Brown, J., Luben, R., Wareham, N. J., Loos, R. J., Heit, J. A., Pankratz, V. S., Norman, A., Goode, E. L., Cunningham, J. M., deAndrade, M., Vierkant, R. A., Czene, K., Fasching, P. A., Baglietto, L., Southey, M. C., Giles, G. G., Shah, K. P., Chan, H. P., Helvie, M. A., Beck, A. H., Knoblauch, N. W., Hazra, A., Hunter, D. J., Kraft, P., Pollan, M., Figueroa, J. D., Couch, F. J., Hopper, J. L., Hall, P., Easton, D. F., Boyd, N. F., Vachon, C. M. & Tamimi, R. M. 2014. Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. *Nat Commun,* 5**,** 5303.
Lissowska, J., Brinton, L. A., Zatonski, W., Blair, A., Bardin-Mikolajczak, A., Peplonska, B., Sherman, M. E., Szeszenia-Dabrowska, N., Chanock, S. & García-Closas, M. 2006. Tobacco smoking, NAT2 acetylation genotype and breast cancer risk. *Int J Cancer,* 119**,** 1961-9.
Lister, R., Pelizzola, M., Dowen, R., Hawkins, R., Hon, G., Tonti-Filippini, J., Nery, J., Lee, L., Ye, Z., Ngo, Q., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A., Thomson, J., Ren, B. & Ecker, J. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature,* 462**,** 315-22.
Lorincz, M., Dickerson, D., Schmitt, M. & Groudine, M. 2004. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol,* 11**,** 1068-75.
Lou S, Lee HM, Qin H, Li JW, Gao Z, Liu X, Chan LL, Kl Lam V, So WY, Wang Y, Lok S, Wang J, Ma RC, Tsui SK, Chan JC, Chan TF, Yip KY. 2014. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biol*, 15(7):408.
Love, R. R. & Philips, J. 2002. Oophorectomy for breast cancer: history revisited. *J Natl Cancer Inst,* 94**,** 1433-4.
Luccioli, S., Zhang, Y., Verrill, L., Ramos-Valle, M. & Kwegyir-Afful, E. 2014. Infant feeding practices and reported food allergies at 6 years of age. *Pediatrics,* 134 Suppl 1**,** S21-8.
Lyratzopoulos, G., Barbiere, J. M., Rachet, B., Baum, M., Thompson, M. R. & Coleman, M. P. 2011. Changes over time in socioeconomic inequalities in breast and rectal cancer survival in England and Wales during a 32-year period (1973-2004): the potential role of health care. *Ann Oncol,* 22**,** 1661-6.

Madureira, P. A., Varshochi, R., Constantinidou, D., Francis, R. E., Coombes, R. C., Yao, K. M. & Lam, E. W. 2006. The Forkhead box M1 protein regulates the transcription of the estrogen receptor alpha in breast cancer cells. *J Biol Chem,* 281**,** 25167-76.

Magnusson, C. M., Persson, I. R., Baron, J. A., Ekbom, A., Bergstrom, R. & Adami, H. O. 1999. The role of reproductive factors and use of oral contraceptives in the aetiology of breast cancer in women aged 50 to 74 years. *Int J Cancer,* 80**,** 231-6.

Malousi, A. & Kouidou, S. 2012. DNA hypermethylation of alternatively spliced and repeat sequences in humans. *Mol Genet Genomics*.

Manavathi, B., Samanthapudi, V. S. & Gajulapalli, V. N. 2014. Estrogen receptor coregulators and pioneer factors: the orchestrators of mammary gland cell fate and development. *Front Cell Dev Biol,* 2**,** 34.

Marmot, M. G., Altman, D. G., Cameron, D. A., Dewar, J. A., Thompson, S. G. & Wilcox, M. 2013. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer,* 108**,** 2205-40.

Martens, J. W., Margossian, A. L., Schmitt, M., Foekens, J. & Harbeck, N. 2009. DNA methylation as a biomarker in breast cancer. *Future Oncol,* 5**,** 1245-56.

Martin, F. L., Cole, K. J., Weaver, G., Hong, G. S., Lam, B. C., Balaram, P., Grover, P. L. & Phillips, D. H. 2001. Genotoxicity of human breast milk from different countries. *Mutagenesis,* 16**,** 401-6.

Martin, F. L., Cole, K. J., Weaver, G., Williams, J. A., Millar, B. C., Grover, P. L. & Phillips, D. H. 1999. Genotoxicity of human milk extracts and detection of DNA damage in exfoliated cells recovered from breast milk. *Biochem Biophys Res Commun,* 259**,** 319-26.

Martin, F. L., Cole, K. J., Williams, J. A., Millar, B. C., Harvey, D., Weaver, G., Grover, P. L. & Phillips, D. H. 2000. Activation of genotoxins to DNA-damaging species in exfoliated breast milk cells. *Mutat Res,* 470**,** 115-24.

Martin, R. M., Middleton, N., Gunnell, D., Owen, C. G. & Smith, G. D. 2005. Breast-feeding and cancer: the Boyd Orr cohort and a systematic review with meta-analysis. *J Natl Cancer Inst,* 97**,** 1446-57.

Martinez-Galan, J., Torres, B., Del Moral, R., Munoz-Gamez, J. A., Martin-Oliva, D., Villalobos, M., Nunez, M. I., Luna Jde, D., Oliver, F. J. & Ruiz de Almodovar, J. M. 2008. Quantitative detection of methylated ESR1 and 14-3-3-sigma gene promoters in serum as candidate biomarkers for diagnosis of breast cancer and evaluation of treatment efficacy. *Cancer Biol Ther,* 7**,** 958-65.

Matuschek, C., Bolke, E., Lammering, G., Gerber, P. A., Peiper, M., Budach, W., Taskin, H., Prisack, H. B., Schieren, G., Orth, K. & Bojar, H. 2010. Methylated APC and GSTP1 Genes in Serum DNA Correlate with the Presence of Circulating Blood Tumor Cells and are Associated with a More Aggressive and Advanced Breast Cancer Disease. *Eur J Med Res,* 15**,** 277-86.

Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. 2013. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res*.

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S. J., Haussler, D., Marra, M. A., Hirst, M., Wang, T. & Costello, J. F. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature,* 466**,** 253-7.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R. & Stamatoyannopoulos, J. A. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science,* 337**,** 1190-5.

Mavaddat, N., Antoniou, A. C., Easton, D. F. & Garcia-Closas, M. 2010. Genetic susceptibility to breast cancer. *Mol Oncol,* 4**,** 174-91.

Mavaddat, N., Pharoah, P. D., Michailidou, K., Tyrer, J., Brook, M. N., Bolla, M. K., Wang, Q., Dennis, J., Dunning, A. M., Shah, M., Luben, R., Brown, J., Bojesen, S. E., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Czene, K., Darabi, H., Eriksson, M., Peto, J., Dos-Santos-Silva, I., Dudbridge, F., Johnson, N., Schmidt, M. K., Broeks, A., Verhoef, S., Rutgers, E. J., Swerdlow, A., Ashworth, A., Orr, N., Schoemaker, M. J., Figueroa, J., Chanock, S. J., Brinton, L., Lissowska, J., Couch, F. J., Olson, J. E., Vachon, C., Pankratz, V. S., Lambrechts, D., Wildiers, H., Van Ongeval, C., van Limbergen, E., Kristensen, V., Grenaker Alnæs, G., Nord, S., Borresen-Dale, A. L., Nevanlinna, H., Muranen, T. A., Aittomäki, K., Blomqvist, C., Chang-Claude, J., Rudolph, A., Seibold, P., Flesch-Janys, D., Fasching, P. A., Haeberle, L., Ekici, A. B., Beckmann, M. W., Burwinkel, B., Marme, F., Schneeweiss, A., Sohn, C., Trentham-Dietz, A., Newcomb, P., Titus, L., Egan, K. M., Hunter, D. J., Lindstrom, S., Tamimi, R. M., Kraft, P., Rahman, N., Turnbull, C., Renwick, A., Seal, S., Li, J., Liu, J., Humphreys, K., Benitez, J., Pilar Zamora, M., Arias Perez, J. I., Menéndez, P., Jakubowska, A., Lubinski, J., Jaworska-Bieniek, K., Durda, K., Bogdanova, N. V., Antonenkova, N. N., Dörk, T., Anton-Culver, H., Neuhausen, S. L., Ziogas, A., Bernstein, L., Devilee, P., Tollenaar, R. A., Seynaeve, C., van Asperen, C. J., Cox, A., Cross, S. S., Reed, M. W., et al. 2015. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst,* 107.

McGuire, W. 2012. Donor human milk for preterm infants. *Pediatrics,* 130**,** e462; author reply 465-6.

Meacham, C. E. & Morrison, S. J. 2013. Tumour heterogeneity and cancer cell plasticity. *Nature,* 501**,** 328-337.

Mehrotra, J., Ganpat, M. M., Kanaan, Y., Fackler, M. J., McVeigh, M., Lahti-Domenici, J., Polyak, K., Argani, P., Naab, T., Garrett, E., Parmigiani, G., Broome, C. & Sukumar, S. 2004. Estrogen receptor/progesterone receptor-negative breast cancers of young African-American women have a higher frequency of methylation of multiple genes than those of Caucasian women. *Clin Cancer Res,* 10**,** 2052-7.

Mendelsohn, A. R. & Larrick, J. W. 2013. The DNA methylome as a biomarker for epigenetic instability and human aging. *Rejuvenation Res,* 16**,** 74-7.

Metivier, R., Gallais, R., Tiffoche, C., Le Peron, C., Jurkowska, R. Z., Carmouche, R. P., Ibberson, D., Barath, P., Demay, F., Reid, G., Benes, V., Jeltsch, A., Gannon, F. & Salbert, G. 2008. Cyclical DNA methylation of a transcriptionally active promoter. *Nature,* 452**,** 45-50.

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., Perkins, B. J., Czene, K., Eriksson, M., Darabi, H., Brand, J. S., Bojesen, S. E., Nordestgaard, B. G., Flyger, H., Nielsen, S. F., Rahman, N., Turnbull, C., Fletcher, O., Peto, J., Gibson, L., dos-Santos-Silva, I., Chang-Claude, J., Flesch-Janys, D., Rudolph, A., Eilber, U., Behrens, S., Nevanlinna, H., Muranen, T. A., Aittomäki, K., Blomqvist, C., Khan, S., Aaltonen, K., Ahsan, H., Kibriya, M. G., Whittemore, A. S., John, E. M., Malone, K. E., Gammon, M. D., Santella, R. M., Ursin, G., Makalic, E., Schmidt, D. F., Casey, G., Hunter, D. J., Gapstur, S. M., Gaudet, M. M., Diver, W. R., Haiman, C. A., Schumacher, F., Henderson, B. E., Le Marchand, L., Berg, C. D., Chanock, S. J., Figueroa, J., Hoover, R. N., Lambrechts, D., Neven, P., Wildiers, H., van Limbergen, E., Schmidt, M. K., Broeks, A., Verhoef, S., Cornelissen, S., Couch, F. J., Olson, J. E., Hallberg, E., Vachon, C., Waisfisz, Q., Meijers-Heijboer, H., Adank, M. A., van der Luijt, R. B., Li, J., Liu, J., Humphreys, K., Kang, D., Choi, J. Y., Park, S. K., Yoo, K. Y., Matsuo, K., Ito, H., Iwata, H., Tajima, K., Guénel, P., Truong, T., Mulot, C., Sanchez, M., Burwinkel, B., Marme, F., Surowy, H., Sohn, C., Wu, A. H., Tseng, C. C., Van Den Berg, D., Stram, D. O., González-Neira, A., Benitez, J., et al. 2015. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet,* 47**,** 373-80.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K., Wang, Q., Dicks, E., Lee, A., Turnbull, C., Rahman, N., Fletcher, O., Peto, J., Gibson, L., Dos Santos Silva, I., Nevanlinna, H., Muranen, T. A., Aittomaki, K., Blomqvist, C., Czene, K., Irwanto, A., Liu, J., Waisfisz, Q., Meijers-Heijboer,

H., Adank, M., van der Luijt, R. B., Hein, R., Dahmen, N., Beckman, L., Meindl, A., Schmutzler, R. K., Muller-Myhsok, B., Lichtner, P., Hopper, J. L., Southey, M. C., Makalic, E., Schmidt, D. F., Uitterlinden, A. G., Hofman, A., Hunter, D. J., Chanock, S. J., Vincent, D., Bacot, F., Tessier, D. C., Canisius, S., Wessels, L. F., Haiman, C. A., Shah, M., Luben, R., Brown, J., Luccarini, C., Schoof, N., Humphreys, K., Li, J., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Couch, F. J., Wang, X., Vachon, C., Stevens, K. N., Lambrechts, D., Moisse, M., Paridaens, R., Christiaens, M. R., Rudolph, A., Nickels, S., Flesch-Janys, D., Johnson, N., Aitken, Z., Aaltonen, K., Heikkinen, T., Broeks, A., Veer, L. J., van der Schoot, C. E., Guenel, P., Truong, T., Laurent-Puig, P., Menegaux, F., Marme, F., Schneeweiss, A., Sohn, C., Burwinkel, B., Zamora, M. P., Perez, J. I., Pita, G., Alonso, M. R., Cox, A., Brock, I. W., Cross, S. S., Reed, M. W., Sawyer, E. J., Tomlinson, I., Kerin, M. J., Miller, N., Henderson, B. E., et al. 2013. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet,* 45**,** 353-61, 361e1-2.

Michels, K. B. 2010. The promises and challenges of epigenetic epidemiology. *Experimental gerontology,* 45**,** 297-301.

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I., Houseman, E. A., Izzi, B., Kelsey, K. T., Meissner, A., Milosavljevic, A., Siegmund, K. D., Bock, C. & Irizarry, R. A. 2013. Recommendations for the design and analysis of epigenome-wide association studies. *Nat Methods,* 10**,** 949-55.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M. & Ding, W. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science,* 266**,** 66-71.

Mill, J. & Heijmans, B. T. 2013. From promises to practical strategies in epigenetic epidemiology. *Nature Reviews Genetics,* 14**,** 585-594.

Millikan, R. C., Pittman, G. S., Newman, B., Tse, C. K., Selmin, O., Rockhill, B., Savitz, D., Moorman, P. G. & Bell, D. A. 1998. Cigarette smoking, N-acetyltransferases 1 and 2, and breast cancer risk. *Cancer Epidemiol Biomarkers Prev,* 7**,** 371-8.

Miranda, T. B., Cortez, C. C., Yoo, C. B., Liang, G., Abe, M., Kelly, T. K., Marquez, V. E. & Jones, P. A. 2009. DZNep is a global histone methylation inhibitor that reactivates developmental genes not silenced by DNA methylation. *Mol Cancer Ther.* United States.

Mirza, S., Sharma, G., Prasad, C. P., Parshad, R., Srivastava, A., Gupta, S. D. & Ralhan, R. 2007. Promoter hypermethylation of TMS1, BRCA1, ERalpha and PRB in serum and tumor DNA of invasive ductal breast carcinoma patients. *Life Sci,* 81**,** 280-7.

Moennikes, O., Loeppen, S., Buchmann, A., Andersson, P., Ittrich, C., Poellinger, L. & Schwarz, M. 2004. A constitutively active dioxin/aryl hydrocarbon receptor promotes hepatocarcinogenesis in mice. *Cancer Res.* United States.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W. R., Hannon, G. J. & Smith, A. D. 2011. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell,* 146**,** 1029-41.

Molyneux, G. & Smalley, M. J. 2011. The cell of origin of BRCA1 mutation-associated breast cancer: a cautionary tale of gene expression profiling. *J Mammary Gland Biol Neoplasia,* 16**,** 51-5.

Monick, M. M., Beach, S. R., Plume, J., Sears, R., Gerrard, M., Brody, G. H. & Philibert, R. A. 2012. Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. *Am J Med Genet B Neuropsychiatr Genet,* 159B**,** 141-51.

Morabia, A., Bernstein, M. S., Bouchardy, I., Kurtz, J. & Morris, M. A. 2000. Breast cancer and active and passive smoking: the role of the N-acetyltransferase 2 genotype. *Am J Epidemiol,* 152**,** 226-32.

Moreno-Bueno, G., Portillo, F. & Cano, A. 2008. Transcriptional regulation of cell polarity in EMT and cancer. *Oncogene.* England.

Morris, T. J. & Beck, S. 2015. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods,* 72**,** 3-8.

Mossberg, A. K., Hun Mok, K., Morozova-Roche, L. A. & Svanborg, C. 2010a. Structure and function of human alpha-lactalbumin made lethal to tumor cells (HAMLET)-type complexes. *Febs j,* 277**,** 4614-25.

Mossberg, A. K., Puchades, M., Halskau, O., Baumann, A., Lanekoff, I., Chao, Y., Martinez, A., Svanborg, C. & Karlsson, R. 2010b. HAMLET interacts with lipid membranes and perturbs their structure and integrity. *PLoS One,* 5**,** e9384.

Mossberg, A. K., Wullt, B., Gustafsson, L., Mansson, W., Ljunggren, E. & Svanborg, C. 2007. Bladder cancers respond to intravesical instillation of HAMLET (human alpha-lactalbumin made lethal to tumor cells). *Int J Cancer,* 121**,** 1352-9.

Muggerud, A. A., Ronneberg, J. A., Warnberg, F., Botling, J., Busato, F., Jovanovic, J., Solvang, H., Bukholm, I., Borresen-Dale, A. L., Kristensen, V. N., Sorlie, T. & Tost, J. 2010. Frequent aberrant DNA methylation of ABCB1, FOXC1, PPP2R2B and PTEN in ductal carcinoma in situ and early invasive breast cancer. *Breast Cancer Res,* 12**,** R3.

Mukhtar, T. K., Yeates, D. R. & Goldacre, M. J. 2013. Breast cancer mortality trends in England and the assessment of the effectiveness of mammography screening: population-based study. *J R Soc Med,* 106**,** 234-42.

Mulero-Navarro, S. & Esteller, M. 2008. Epigenetic biomarkers for human cancer: the time is now. *Crit Rev Oncol Hematol,* 68**,** 1-11.

Muller, H. M., Widschwendter, A., Fiegl, H., Ivarsson, L., Goebel, G., Perkmann, E., Marth, C. & Widschwendter, M. 2003. DNA methylation in serum of breast cancer patients: an independent prognostic marker. *Cancer Res,* 63**,** 7641-5.

Muraro, A. P., Goncalves-Silva, R. M., Moreira, N. F., Ferreira, M. G., Nunes-Freitas, A. L., Abreu-Villaca, Y. & Sichieri, R. 2014. Effect of tobacco smoke exposure during pregnancy and preschool age on growth from birth to adolescence: a cohort study. *BMC Pediatr,* 14**,** 99.

Narod, S. A. & Foulkes, W. D. 2004. BRCA1 and BRCA2: 1994 and beyond. *Nature reviews.Cancer,* 4**,** 665-676.

Nathan, D. & Crothers, D. M. 2002. Bending and flexibility of methylated and unmethylated EcoRI DNA. *J Mol Biol,* 316**,** 7-17.

NCI. 2014. *PDQ® Breast Cancer Treatment. Available at:http://cancer.gov/cancertopics/pdq/treatment/breast/Patient.* [Online]. Bethesda, MD: National Cancer Institute. [Accessed 16th March 2015].

Nelson, H. D., Pappas, M., Zakher, B., Mitchell, J. P., Okinaka-Hu, L. & Fu, R. 2014. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: a systematic review to update the U.S. Preventive Services Task Force recommendation. *Ann Intern Med,* 160**,** 255-66.

Nestor, C. E., Ottaviano, R., Reinhardt, D., Cruickshanks, H. A., Mjoseng, H. K., McPherson, R. C., Lentini, A., Thomson, J. P., Dunican, D. S., Pennings, S., Anderton, S. M., Benson, M. & Meehan, R. R. 2015. Rapid reprogramming of epigenetic and transcriptional profiles in mammalian culture systems. *Genome Biol,* 16**,** 11.

Newcomb, P. A., Nichols, H. B., Beasley, J. M., Egan, K., Titus-Ernstoff, L., Hampton, J. M. & Trentham-Dietz, A. 2009. No difference between red wine or white wine consumption and breast cancer risk. *Cancer Epidemiol Biomarkers Prev,* 18**,** 1007-10.

Newman, L. A. 2014. Breast Cancer Disparities: High-Risk Breast Cancer and African Ancestry. *Surg Oncol Clin N Am,* 23**,** 579-592.

Ng, C. K., Schultheis, A. M., Bidard, F. C., Weigelt, B. & Reis-Filho, J. S. 2015. Breast cancer genomics from microarrays to massively parallel sequencing: paradigms and new insights. *J Natl Cancer Inst,* 107.

Ng, H. H. & Bird, A. 1999. DNA methylation and chromatin modification. *Curr Opin Genet Dev.* England.

NICE. 2010. *Donor milk banks: the operation of donor milk bank services* [Online]. Available: http://www.nice.org.uk/guidance/cg93 [Accessed 26th May 2015].

Nichols, H. B., Trentham-Dietz, A., Love, R. R., Hampton, J. M., Hoang Anh, P. T., Allred, D. C., Mohsin, S. K. & Newcomb, P. A. 2005. Differences in breast cancer risk factors by tumor marker subtypes among premenopausal Vietnamese and Chinese women. *Cancer Epidemiol Biomarkers Prev,* 14**,** 41-7.

Nichols, H. B., Trentham-Dietz, A., Sprague, B. L., Hampton, J. M., Titus-Ernstoff, L. & Newcomb, P. A. 2008. Effects of birth order and maternal age on breast cancer risk: modification by whether women had been breast-fed. *Epidemiology,* 19**,** 417-23.

Nielsen, C. H., Larsen, A. & Nielsen, A. L. 2014. DNA methylation alterations in response to prenatal exposure of maternal cigarette smoking: A persistent epigenetic impact on health from maternal lifestyle? *Arch Toxicol*.

Nilsen, T. I. & Vatten, L. J. 2001. Adult height and risk of breast cancer: a possible effect of early nutrition. *Br J Cancer,* 85**,** 959-61.

.

Novakovic, B., Ryan, J., Pereira, N., Boughton, B., Craig, J. M. & Saffery, R. 2014. Postnatal stability, tissue, and time specific effects of AHRR methylation change in response to maternal smoking in pregnancy. *Epigenetics.* United States.

O'Hare, M. J., Ormerod, M. G., Monaghan, P., Lane, E. B. & Gusterson, B. A. 1991. Characterization in vitro of luminal and myoepithelial cells isolated from the human mammary gland by cell sorting. *Differentiation,* 46**,** 209-21.

Ochs-Balcom, H. M., Wiesner, G. & Elston, R. C. 2007. A meta-analysis of the association of N-acetyltransferase 2 gene (NAT2) variants with breast cancer. *Am J Epidemiol.* United States.

Office, S. G. s. 2004. The 2004 United States Surgeon General's Report: The Health Consequences of Smoking. *N S W Public Health Bull,* 15**,** 107.

Ogino, S., Lochhead, P., Chan, A. T., Nishihara, R., Cho, E., Wolpin, B. M., Meyerhardt, J. A., Meissner, A., Schernhammer, E. S., Fuchs, C. S. & Giovannucci, E. 2013. Molecular pathological epidemiology of epigenetics: emerging integrative science to analyze environment, host, and disease. *Mod Pathol,* 26**,** 465-84.

Oh, A. Y. & Kacker, A. 2014. Do electronic cigarettes impart a lower potential disease burden than conventional tobacco cigarettes? Review on E-cigarette vapor versus tobacco smoke. *Laryngoscope,* 124**,** 2702-6.

Ooi, S. K. & Bestor, T. H. 2008. The colorful history of active DNA demethylation. *Cell,* 133**,** 1145-8.

Ottaviano, Y. L., Issa, J. P., Parl, F. F., Smith, H. S., Baylin, S. B. & Davidson, N. E. 1994. Methylation of the estrogen receptor gene CpG island marks loss of estrogen receptor expression in human breast cancer cells. *Cancer Res,* 54**,** 2552-5.

Palmieri, C., Cheng, G. J., Saji, S., Zelada-Hedman, M., Warri, A., Weihua, Z., Van Noorden, S., Wahlstrom, T., Coombes, R. C., Warner, M. & Gustafsson, J. A. 2002. Estrogen receptor beta in breast cancer. *Endocr Relat Cancer,* 9**,** 1-13.

Pan, H., He, Z., Ling, L., Ding, Q., Chen, L., Zha, X., Zhou, W., Liu, X. & Wang, S. 2014. Reproductive factors and breast cancer risk among BRCA1 or BRCA2 mutation carriers: results from ten studies. *Cancer Epidemiol,* 38**,** 1-8.

Perillo, B., Ombra, M., Bertoni, A., Cuozzo, C., Sacchetti, S., Sasso, A., Chiariotti, L., Malorni, A., Abbondanza, C. & Avvedimento, E. 2008. DNA oxidation as triggered by H3K9me2 demethylation drives estrogen-induced gene expression. *Science,* 319**,** 202-6.

Perillo, B., Sasso, A., Abbondanza, C. & Palumbo, G. 2000. 17beta-estradiol inhibits apoptosis in MCF-7 cells, inducing bcl-2 expression via two estrogen-responsive elements present in the coding sequence. *Mol Cell Biol,* 20**,** 2890-901.

Perou, C. M. 2011. Molecular stratification of triple-negative breast cancers. *Oncologist,* 16 Suppl 1**,** 61-70.

Philibert, R. A., Beach, S. R. & Brody, G. H. 2012. Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. *Epigenetics.* United States.

Philibert, R. A., Beach, S. R., Lei, M. K. & Brody, G. H. 2013. Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. *Clin Epigenetics,* 5**,** 19.
Pirie, K., Peto, R., Reeves, G. K., Green, J., Beral, V. & Collaborators, M. W. S. 2013. The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. *Lancet,* 381**,** 133-41.
Pleyer, L. & Greil, R. 2015. Digging deep into "dirty" drugs - modulation of the methylation machinery. *Drug Metab Rev***,** 1-28.
Pogo, B. G. & Holland, J. F. 1997. Possibilities of a viral etiology for human breast cancer. A review. *Biol Trace Elem Res,* 56**,** 131-42.
Prat, A., Adamo, B., Fan, C., Peg, V., Vidal, M., Galván, P., Vivancos, A., Nuciforo, P., Palmer, H. G., Dawood, S., Rodón, J., Ramony Cajal, S., Del Campo, J. M., Felip, E., Tabernero, J. & Cortés, J. 2013. Genomic analyses across six cancer types identify basal-like breast cancer as a unique molecular entity. *Sci Rep,* 3**,** 3544.
Prat, A., Cruz, C., Hoadley, K. A., Díez, O., Perou, C. M. & Balmaña, J. 2014. Molecular features of the basal-like breast cancer subtype based on BRCA1 mutation status. *Breast Cancer Res Treat,* 147**,** 185-91.
Prat, A. & Perou, C. M. 2011. Deconstructing the molecular portraits of breast cancer. *Mol Oncol,* 5**,** 5-23.
Puthia, M., Storm, P., Nadeem, A., Hsiung, S. & Svanborg, C. 2013. Prevention and treatment of colon cancer by peroral administration of HAMLET (human alpha-lactalbumin made lethal to tumour cells). *Gut*.
Quigley, M. & McGuire, W. 2014. Formula versus donor breast milk for feeding preterm or low birth weight infants. *Cochrane Database Syst Rev,* 4**,** CD002971.
Quigley, M. A., Henderson, G., Anthony, M. Y. & McGuire, W. 2007. Formula milk versus donor breast milk for feeding preterm or low birth weight infants. *Cochrane Database Syst Rev***,** Cd002971.
Ramos, E. A., Camargo, A. A., Braun, K., Slowik, R., Cavalli, I. J., Ribeiro, E. M., Pedrosa Fde, O., de Souza, E. M., Costa, F. F. & Klassen, G. 2010. Simultaneous CXCL12 and ESR1 CpG island hypermethylation correlates with poor prognosis in sporadic breast cancer. *BMC Cancer,* 10**,** 23.
Rauch, T. A., Wu, X., Zhong, X., Riggs, A. D. & Pfeifer, G. P. 2009. A human B cell methylome at 100-base pair resolution. *Proc Natl Acad Sci U S A.* United States.
Ravdin, P. M. 2008. Hormone replacement therapy and the increase in the incidence of invasive lobular cancer. *Breast Dis,* 30**,** 3-8.
Reeves, H. M. 2006. Sahagún's "Florentine codex," a little known Aztecan natural history of the Valley of Mexico. *Arch Nat Hist,* 33**,** 302-21.
Reik, W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature,* 447**,** 425-432.
Reis-Filho, J. S. & Pusztai, L. 2011. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet,* 378**,** 1812-23.
Report, W. C. 2008. *http://www.iarc.fr/en/publications/pdfs-online/wcr/2008/wcr_2008.pdf* [Online].
Reynolds, P. 2013. Smoking and breast cancer. *J Mammary Gland Biol Neoplasia,* 18**,** 15-23.
Riboli, E. 2001. The European Prospective Investigation into Cancer and Nutrition (EPIC): plans and progress. *J Nutr,* 131**,** 170S-175S.
Riboli, E. & Kaaks, R. 1997. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol,* 26 Suppl 1**,** S6-14.
Richmond, R. C., Simpkin, A. J., Woodward, G., Gaunt, T. R., Lyttleton, O., McArdle, W. L., Ring, S. M., Smith, A. D., Timpson, N. J., Tilling, K., Davey Smith, G. & Relton, C. L. 2015. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings

from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum Mol Genet.* England: The Author 2014. Published by Oxford University Press.

Riggs, A. D. 1975. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet,* 14**,** 9-25.

Rignall, B., Grote, K., Gavrilov, A., Weimer, M., Kopp-Schneider, A., Krause, E., Appel, K. E., Buchmann, A., Robertson, L. W., Lehmler, H. J., Kania-Korwel, I., Chahoud, I. & Schwarz, M. 2013. Biological and tumor-promoting effects of dioxin-like and non-dioxin-like polychlorinated biphenyls in mouse liver after single or combined treatment. *Toxicol Sci,* 133**,** 29-41.

Robertson, A. 1977. Conrad Hal Waddington, 8 November 1905--26 September 1975. *Biogr Mem Fellows R Soc,* 23**,** 575-622.

Robins, H. S., Ericson, N. G., Guenthoer, J., O'Briant, K. C., Tewari, M., Drescher, C. W. & Bielas, J. H. 2013. Digital genomic quantification of tumor-infiltrating lymphocytes. *Sci Transl Med,* 5**,** 214ra169.

Robinson, S., Ntani, G., Simmonds, S., Syddall, H., Dennison, E., Sayer, A. A., Barker, D. & Cooper, C. 2013. Type of milk feeding in infancy and health behaviours in adult life: findings from the Hertfordshire Cohort Study. *Br J Nutr,* 109**,** 1114-22.

Robinson DR1, Wu YM, Vats P, Su F, Lonigro RJ, Cao X, Kalyana-Sundaram S, Wang R, Ning Y, Hodges L, Gursky A, Siddiqui J, Tomlins SA, Roychowdhury S, Pienta KJ, Kim SY, Roberts JS, Rae JM, Van Poznak CH, Hayes DF, Chugh R, Kunju LP, Talpaz M, Schott AF, Chinnaiyan AM. 2013. Activating ESR1 mutations in hormone-resistant metastatic breast cancer. Nat Genet. 45(12):1446-51.

Royo, J. L., Hidalgo, M. & Ruiz, A. 2007. Pyrosequencing protocol using a universal biotinylated primer for mutation detection and SNP genotyping. *Nature protocols,* 2**,** 1734-1739.

Sabbagh, H. J., Hassan, M. H., Innes, N. P., Elkodary, H. M., Little, J. & Mossey, P. A. 2015. Passive smoking in the etiology of non-syndromic orofacial clefts: a systematic review and meta-analysis. *PLoS One,* 10**,** e0116963.

Sakakura, T. 1991. New aspects of stroma-parenchyma relations in mammary gland differentiation. *Int Rev Cytol,* 125**,** 165-202.

Sakorafas, G. H. & Safioleas, M. 2009. Breast cancer surgery: an historical narrative. Part I. From prehistoric times to Renaissance. *Eur J Cancer Care (Engl),* 18**,** 530-44.

Samuel, N. & Hudson, T. J. 2011. The molecular and cellular heterogeneity of pancreatic ductal adenocarcinoma. *Nature Reviews Gastroenterology and Hepatology,* 9**,** 77-87.

Santagata, S. & Ince, T. A. 2014. Normal cell phenotypes of breast epithelial cells provide the foundation of a breast cancer taxonomy. *Expert Rev Anticancer Ther,* 14**,** 1385-9.

Sasco, A. J., Secretan, M. B. & Straif, K. 2004. Tobacco smoking and cancer: a brief review of recent epidemiological evidence. *Lung Cancer.* Ireland.

Sasieni, P. D., Shelton, J., Ormiston-Smith, N., Thomson, C. S. & Silcocks, P. B. 2011. What is the lifetime risk of developing cancer?: the effect of adjusting for multiple primaries. *Br J Cancer.* England.

Sauter, E. R., Klein-Szanto, A., Macgibbon, B. & Ehya, H. 2010. Nipple aspirate fluid and ductoscopy to detect breast cancer. *Diagn Cytopathol,* 38**,** 244-51.

Schneider, S. S., Aslebagh, R., Wetie, A. G., Sturgeon, S. R., Darie, C. C. & Arcaro, K. F. 2014. Using breast milk to assess breast cancer risk: the role of mass spectrometry-based proteomics. *Adv Exp Med Biol,* 806**,** 399-408.

Schnitt, S. J. 2010. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Mod Pathol,* 23 Suppl 2**,** S60-4.

Schor, I. E., Fiszbein, A., Petrillo, E. & Kornblihtt, A. R. 2013. Intragenic epigenetic changes modulate NCAM alternative splicing in neuronal differentiation. *Embo j,* 32**,** 2264-74.

Schulz, W. A. & Goering, W. 2011. Eagles report: Developing cancer biomarkers from genome-wide DNA methylation analyses. *World J Clin Oncol,* 2**,** 1-7.

Screening, I. U. P. o. B. C. 2012. The benefits and harms of breast cancer screening: an independent review. *Lancet,* 380**,** 1778-86.
Sedlackova, T., Repiska, G., Celec, P., Szemes, T. & Minarik, G. 2013. Fragmentation of DNA affects the accuracy of the DNA quantitation by the commonly used methods. *Biol Proced Online,* 15**,** 5.
Seitz, H. K., Pelucchi, C., Bagnardi, V. & La Vecchia, C. 2012. Epidemiology and pathophysiology of alcohol and breast cancer: Update 2012. *Alcohol Alcohol,* 47**,** 204-12.
Shann, Y. J., Cheng, C., Chiao, C. H., Chen, D. T., Li, P. H. & Hsu, M. T. 2008. Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines. *Genome Res.* United States.
Sharma, G., Mirza, S., Yang, Y. H., Parshad, R., Hazrah, P., Datta Gupta, S. & Ralhan, R. 2009. Prognostic relevance of promoter hypermethylation of multiple genes in breast cancer patients. *Cell Oncol,* 31**,** 487-500.
Shenker, N. 2010. *Intragenic methylation of the oestrogen receptor alpha gene in human breast cancer: implications for gene expression.* MSC in Reproductive and Developmental Biology, Imperial College.
Shenker, N. & Flanagan, J. M. 2012. Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research. *Br J Cancer,* 106**,** 248-53.
Shenker, N. S., Flower, K. J., Wilhelm-Benartzi, C. S., Dai, W., Bell, E., Gore, E., El Bahrawy, M., Weaver, G., Brown, R. & Flanagan, J. M. 2015. Transcriptional implications of intragenic DNA methylation in the oestrogen receptor alpha gene in breast cancer cells and tissues. *BMC Cancer,* 15**,** 337.
Shenker, N. S., Polidoro, S., van Veldhoven, K., Sacerdote, C., Ricceri, F., Birrell, M. A., Belvisi, M. G., Brown, R., Vineis, P. & Flanagan, J. M. 2013a. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Hum Mol Genet,* 22**,** 843-51.
Shenker, N. S., Ueland, P. M., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., Flanagan, J. M. & Vineis, P. 2013b. DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology,* 24**,** 712-6.
Shi L, Sun L, Li Q, Liang J, Yu W, Yi X, Yang X, Li Y, Han X, Zhang Y, Xuan C, Yao Z, Shang Y. 2011. Histone demethylase JMJD2B coordinates H3K4/H3K9 methylation and promotes hormonally responsive breast carcinogenesis. *PNAS*, 108(18):7541-6.
Shiels, M. S., Gibson, T., Sampson, J., Albanes, D., Andreotti, G., Beane Freeman, L., Berrington de Gonzalez, A., Caporaso, N., Curtis, R. E., Elena, J., Freedman, N. D., Robien, K., Black, A. & Morton, L. M. 2014. Cigarette smoking prior to first cancer and risk of second smoking-associated cancers among survivors of bladder, kidney, head and neck, and stage I lung cancers. *J Clin Oncol,* 32**,** 3989-95.
Shimizu, Y., Nakatsuru, Y., Ichinose, M., Takahashi, Y., Kume, H., Mimura, J., Fujii-Kuriyama, Y. & Ishikawa, T. 2000. Benzo[a]pyrene carcinogenicity is lost in mice lacking the aryl hydrocarbon receptor. *Proc Natl Acad Sci U S A,* 97**,** 779-82.
Shiovitz, S. & Korde, L. A. 2015. Genetics of breast cancer: a topic in evolution. *Ann Oncol*.
Smalley, M. & Ashworth, A. 2003. Stem cells and breast cancer: A field in transit. *Nat Rev Cancer,* 3**,** 832-44.
Smilowitz, J. T., O'Sullivan, A., Barile, D., German, J. B., Lonnerdal, B. & Slupsky, C. M. 2013. The human milk metabolome reveals diverse oligosaccharide profiles. *J Nutr,* 143**,** 1709-18.
Snell, C., Krypuy, M., Wong, E. M., Loughrey, M. B., Dobrovic, A. & investigators, k. 2008. BRCA1 promoter methylation in peripheral blood DNA of mutation negative familial breast cancer patients with a BRCA1 tumour phenotype. *Breast Cancer Res,* 10**,** R12.
Sogon, T., Masamura, S., Hayashi, S., Santen, R. J., Nakachi, K. & Eguchi, H. 2007. Demethylation of promoter C region of estrogen receptor alpha gene is correlated with its enhanced

expression in estrogen-ablation resistant MCF-7 cells. *J Steroid Biochem Mol Biol,* 105**,** 106-14.

Soo-Quee Koh, D. & Choon-Huat Koh, G. 2007. The use of salivary biomarkers in occupational and environmental medicine. *Occup Environ Med,* 64**,** 202-10.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lonning, P. & Borresen-Dale, A. L. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A,* 98**,** 10869-74.

Sorlie, T., Wang, Y., Xiao, C., Johnsen, H., Naume, B., Samaha, R. R. & Borresen-Dale, A. L. 2006. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics,* 7**,** 127.

Sproul, D., Nestor, C., Culley, J., Dickson, J. H., Dixon, J. M., Harrison, D. J., Meehan, R. R., Sims, A. H. & Ramsahoye, B. H. 2011. Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. *Proc Natl Acad Sci U S A.*

Sreekumar, A., Roarty, K. P. & Rosen, J. M. 2015. The mammary stem cell hierarchy: a looking glass into heterogeneous breast cancer landscapes. *Endocr Relat Cancer*.

Statistics, O. f. N. 2011. Cancer Survival in England: Patients Diagnosed 2005-2009 and Followed up to 2010.

Stefansson, O. A., Jonasson, J. G., Olafsdottir, K., Hilmarsdottir, H., Olafsdottir, G., Esteller, M., Johannsson, O. T. & Eyfjord, J. E. 2011. CpG island hypermethylation of BRCA1 and loss of pRb as co-occurring events in basal/triple-negative breast cancer. *Epigenetics.* United States.

Stewart, T. H., Sage, R. D., Stewart, A. F. & Cameron, D. W. 2000. Breast cancer incidence highest in the range of one species of house mouse, Mus domesticus. *Br J Cancer,* 82**,** 446-51.

Stingl, J., Raouf, A., Emerman, J. T. & Eaves, C. J. 2005. Epithelial progenitors in the normal human mammary gland. *J Mammary Gland Biol Neoplasia,* 10**,** 49-59.

Sturgeon, S. R., Arcaro, K. F., Johnson, M. A., Balasubramanian, R., Zorn, M., Jerry, D. J. & Schneider, S. S. 2014. DNA methylation in paired breast epithelial and white blood cells from women undergoing reduction mammoplasty. *Anticancer Res,* 34**,** 2985-90.

Suijkerbuijk, K. P., Fackler, M. J., Sukumar, S., van Gils, C. H., van Laar, T., van der Wall, E., Vooijs, M. & van Diest, P. J. 2008. Methylation is less abundant in BRCA1-associated compared with sporadic breast cancer. *Ann Oncol,* 19**,** 1870-4.

Suzuki, M. M. & Bird, A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet.* England.

Svanborg, C., Agerstam, H., Aronson, A., Bjerkvig, R., Duringer, C., Fischer, W., Gustafsson, L., Hallgren, O., Leijonhuvud, I., Linse, S., Mossberg, A. K., Nilsson, H., Pettersson, J. & Svensson, M. 2003. HAMLET kills tumor cells by an apoptosis-like mechanism--cellular, molecular, and therapeutic aspects. *Adv Cancer Res,* 88**,** 1-29.

Svensson, M., Duringer, C., Hallgren, O., Mossberg, A. K., Hakansson, A., Linse, S. & Svanborg, C. 2002. Hamlet--a complex from human milk that induces apoptosis in tumor cells but spares healthy cells. *Adv Exp Med Biol,* 503**,** 125-32.

Svensson, M., Fast, J., Mossberg, A. K., Duringer, C., Gustafsson, L., Hallgren, O., Brooks, C. L., Berliner, L., Linse, S. & Svanborg, C. 2003. Alpha-lactalbumin unfolding is not sufficient to cause apoptosis, but is required for the conversion to HAMLET (human alpha-lactalbumin made lethal to tumor cells). *Protein Sci,* 12**,** 2794-804.

Swerdlow, A. J., Jones, M. E., Schoemaker, M. J., Hemming, J., Thomas, D., Williamson, J. & Ashworth, A. 2011. The Breakthrough Generations Study: design of a long-term UK cohort study to investigate breast cancer aetiology. *Br J Cancer,* 105**,** 911-7.

.

Syddall, H. E., Simmonds, S. J., Martin, H. J., Watson, C., Dennison, E. M., Cooper, C. & Sayer, A. A. 2010. Cohort profile: The Hertfordshire Ageing Study (HAS). *Int J Epidemiol,* 39**,** 36-43.

Takai, D. & Jones, P. A. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A.* United States.

Talhout, R., Schulz, T., Florek, E., van Benthem, J., Wester, P. & Opperhuizen, A. 2011. Hazardous compounds in tobacco smoke. *Int J Environ Res Public Health,* 8**,** 613-28.

Tang, L. Y., Chen, L. J., Qi, M. L., Su, Y., Su, F. X., Lin, Y., Wang, K. P., Jia, W. H., Zhuang, Z. X. & Ren, Z. F. 2013. Effects of passive smoking on breast cancer risk in pre/post-menopausal women as modified by polymorphisms of PARP1 and ESR1. *Gene,* 524**,** 84-9.

Taylor, S. C. & Mrkusich, E. M. 2014. The state of RT-quantitative PCR: firsthand observations of implementation of minimum information for the publication of quantitative real-time PCR experiments (MIQE). *J Mol Microbiol Biotechnol,* 24**,** 46-52.

Terry, P. D. & Goodman, M. 2006. Is the association between cigarette smoking and breast cancer modified by genotype? A review of epidemiologic studies and meta-analysis. *Cancer Epidemiol Biomarkers Prev.* United States.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Gayther, S. A., Apostolidou, S., Jones, A., Lechner, M., Beck, S., Jacobs, I. J. & Widschwendter, M. 2009. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PloS one,* 4**,** e8274.

Thompson, P. A., Kadlubar, F. F., Vena, S. M., Hill, H. L., McClure, G. H., McDaniel, L. P. & Ambrosone, C. B. 1998. Exfoliated ductal epithelial cells in human breast milk: a source of target tissue DNA for molecular epidemiologic studies of breast cancer. *Cancer Epidemiol Biomarkers Prev,* 7**,** 37-42.

Timofeeva, M. N., McKay, J. D., Smith, G. D., Johansson, M., Byrnes, G. B., Chabrier, A., Relton, C., Ueland, P. M., Vollset, S. E., Midttun, O., Nygard, O., Slimani, N., Romieu, I., Clavel-Chapelon, F., Boutron-Ruault, M. C., Fagherazzi, G., Kaaks, R., Teucher, B., Boeing, H., Weikert, C., Bueno-de-Mesquita, H. B., van Gils, C., Peeters, P. H., Agudo, A., Barricarte, A., Huerta, J. M., Rodriguez, L., Sanchez, M. J., Larranaga, N., Khaw, K. T., Wareham, N., Allen, N. E., Travis, R. C., Gallo, V., Norat, T., Krogh, V., Masala, G., Panico, S., Sacerdote, C., Tumino, R., Trichopoulou, A., Lagiou, P., Trichopoulos, D., Rasmuson, T., Hallmans, G., Riboli, E., Vineis, P. & Brennan, P. 2011. Genetic polymorphisms in 15q25 and 19q13 loci, cotinine levels, and risk of lung cancer in EPIC. *Cancer Epidemiol Biomarkers Prev.* United States: 2011 Aacr.

Toft, D. & Gorski, J. 1966. A receptor molecule for estrogens: isolation from the rat uterus and preliminary characterization. *Proc Natl Acad Sci U S A,* 55**,** 1574-81.

Tost, J. 2010. DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. *Molecular biotechnology,* 44**,** 71-81.

Travis, R. C., Reeves, G. K., Green, J., Bull, D., Tipper, S. J., Baker, K., Beral, V., Peto, R., Bell, J., Zelenika, D. & Lathrop, M. 2010. Gene-environment interactions in 7610 women with breast cancer: prospective evidence from the Million Women Study. *Lancet.* England: 2010 Elsevier Ltd.

Tsaprouni, L. G., Yang, T. P., Bell, J., Dick, K. J., Kanoni, S., Nisbet, J., Vinuela, A., Grundberg, E., Nelson, C. P., Meduri, E., Buil, A., Cambien, F., Hengstenberg, C., Erdmann, J., Schunkert, H., Goodall, A. H., Ouwehand, W. H., Dermitzakis, E., Spector, T. D., Samani, N. J. & Deloukas, P. 2014. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics,* 9**,** 1382-96.

Tufarelli, C., Stanley, J. A., Garrick, D., Sharpe, J. A., Ayyub, H., Wood, W. G. & Higgs, D. R. 2003. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat Genet.* United States.

Tully, D. B., Jones, F. & Tully, M. R. 2001. Donor milk: what's in it and what's not. *J Hum Lact,* 17**,** 152-5.

Turner, N. C. & Reis-Filho, J. S. 2012. Genetic heterogeneity and cancer drug resistance. *Lancet Oncol,* 13**,** e178-85.

UK, S. I. T. a. C. R. 2012. *Lifetime risk as calculated using 2010 data for females and 2008-2010 data for males* [Online]. Available: http://www.cancerresearchuk.org/cancer-info/cancerstats/types/breast/incidence/uk-breast-cancer-incidence-statistics#source20.

Underwood, M. A. 2013. Human milk for the premature infant. *Pediatr Clin North Am,* 60**,** 189-207.

Updegrove, K. 2013. Nonprofit Human Milk Banking in the United States. *J Midwifery Womens Health*.

Valdenaire, O. & Schweizer, A. 2000. Endothelin-converting enzyme-like 1 (ECEL1; 'XCE'): a putative metallopeptidase crucially involved in the nervous control of respiration. *Biochem Soc Trans,* 28**,** 426-30.

Van De Voorde, L., Speeckaert, R., Van Gestel, D., Bracke, M., De Neve, W., Delanghe, J. & Speeckaert, M. 2012. DNA methylation-based biomarkers in serum of patients with breast cancer. *Mutat Res*.

Veeck, J. & Esteller, M. 2010. Breast cancer epigenetics: from DNA methylation to microRNAs. *J Mammary Gland Biol Neoplasia,* 15**,** 5-17.

Velicescu, M., Weisenberger, D. J., Gonzales, F. A., Tsai, Y. C., Nguyen, C. T. & Jones, P. A. 2002. Cell division is required for de novo methylation of CpG islands in bladder cancer cells. *Cancer Res,* 62**,** 2378-84.

Velie, E. M., Nechuta, S. & Osuch, J. R. 2005. Lifetime reproductive and anthropometric risk factors for breast cancer in postmenopausal women. *Breast Dis,* 24**,** 17-35.

Vessey, M. P. 1997. Effect of endogenous and exogenous hormones on breast cancer: epidemiology. *Verh Dtsch Ges Pathol,* 81**,** 493-501.

Villasenor, A., Garcia-Perez, I., Garcia, A., Posma, J. M., Fernandez-Lopez, M., Nicholas, A. J., Modi, N., Holmes, E. & Barbas, C. 2014. Breast milk metabolome characterization in a single-phase extraction, multiplatform analytical approach. *Anal Chem,* 86**,** 8245-52.

Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J. M., Bollen, M., Esteller, M., Di Croce, L., de Launoit, Y. & Fuks, F. 2006. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature,* 439**,** 871-4.

Visvader, J. E. 2011. Cells of origin in cancer. *Nature,* 469**,** 314-22.

Visvader, J. E. & Stingl, J. 2014. Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes Dev,* 28**,** 1143-58.

Vohr, B. R., Poindexter, B. B., Dusick, A. M., McKinley, L. T., Higgins, R. D., Langer, J. C. & Poole, W. K. 2007. Persistent beneficial effects of breast milk ingested in the neonatal intensive care unit on outcomes of extremely low birth weight infants at 30 months of age. *Pediatrics,* 120**,** e953-9.

Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H. S., Diver, W. R., Thun, M. J., Cox, D. G., Hankinson, S. E., Kraft, P., Rosner, B., Berg, C. D., Brinton, L. A., Lissowska, J., Sherman, M. E., Chlebowski, R., Kooperberg, C., Jackson, R. D., Buckman, D. W., Hui, P., Pfeiffer, R., Jacobs, K. B., Thomas, G. D., Hoover, R. N., Gail, M. H., Chanock, S. J. & Hunter, D. J. 2010. Performance of common genetic variants in breast-cancer risk models. *The New England journal of medicine,* 362**,** 986-993.

Walsh, T., Casadei, S., Coats, K. H., Swisher, E., Stray, S. M., Higgins, J., Roach, K. C., Mandell, J., Lee, M. K., Ciernikova, S., Foretova, L., Soucek, P. & King, M. C. 2006. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA,* 295**,** 1379-88.

Wang, W. & Srivastava, S. 2010. Strategic approach to validating methylated genes as biomarkers for breast cancer. *Cancer Prev Res (Phila Pa),* 3**,** 16-24.

Wang, X., Tager, I. B., Van Vunakis, H., Speizer, F. E. & Hanrahan, J. P. 1997. Maternal smoking during pregnancy, urine cotinine concentrations, and birth outcomes. A prospective cohort study. *Int J Epidemiol,* 26**,** 978-88.

Wang, Z., Zhang, X., Shen, P., Loggie, B. W., Chang, Y. & Deuel, T. F. 2005. Identification, cloning, and expression of human estrogen receptor-alpha36, a novel variant of human estrogen receptor-alpha66. *Biochem Biophys Res Commun,* 336**,** 1023-7.

Wannamethee, S. G., Lowe, G. D., Shaper, A. G., Rumley, A., Lennon, L. & Whincup, P. H. 2005. Associations between cigarette smoking, pipe/cigar smoking, and smoking cessation, and haemostatic and inflammatory markers for cardiovascular disease. *Eur Heart J.* England.

WCRF 2007. Food, Nutrition, Physical Activity and the Prevention of Cancer.

Weber, B. L. & Nathanson, K. L. 2000. Low penetrance genes associated with increased risk for breast cancer. *Eur J Cancer.* England.

Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L. & Schubeler, D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet.* United States.

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M. & Schübeler, D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet,* 39**,** 457-66.

Wei, M., Xu, J., Dignam, J., Nanda, R., Sveen, L., Fackenthal, J., Grushko, T. A. & Olopade, O. I. 2008. Estrogen receptor alpha, BRCA1, and FANCF promoter methylation occur in distinct subsets of sporadic breast cancers. *Breast Cancer Res Treat,* 111**,** 113-20.

Weigel, R. J. & deConinck, E. C. 1993. Transcriptional control of estrogen receptor in estrogen receptor-negative breast carcinoma. *Cancer Res,* 53**,** 3472-4.

Welp, E. A., Weiderpass, E., Boffetta, P., Vainio, H., Vasama-Neuvonen, K., Petralia, S. & Partanen, T. J. 1998. Environmental risk factors of breast cancer. *Scand J Work Environ Health.* Finland.

Wenbin, D., Zhuo, C., Zhibing, M., Chen, Z., Ruifan, Y., Jie, J., Cheng, Q. & Zhenming, G. 2013. The effect of smoking on the risk of gallbladder cancer: a meta-analysis of observational studies. *Eur J Gastroenterol Hepatol,* 25**,** 373-9.

West, O., Hajek, P. & McRobbie, H. 2011. Systematic review of the relationship between the 3-hydroxycotinine/cotinine ratio and cigarette dependence. *Psychopharmacology (Berl),* 218**,** 313-22.

WHO. 2009. *Cancer Factsheet* [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/index.html [Accessed 8th June 2010].

WHO 2010. Global Strategy for Infant and Young Child Feeding. *WHO*.

Widschwendter, M., Siegmund, K. D., Muller, H. M., Fiegl, H., Marth, C., Muller-Holzner, E., Jones, P. A. & Laird, P. W. 2004. Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen. *Cancer Res,* 64**,** 3807-13.

Wild, L. & Flanagan, J. M. 2010. Genome-wide hypomethylation in cancer may be a passive consequence of transformation. *Biochim Biophys Acta*.

Wilhelm-Benartzi, C. S., Koestler, D. C., Karagas, M. R., Flanagan, J. M., Christensen, B. C., Kelsey, K. T., Marsit, C. J., Houseman, E. A. & Brown, R. 2013. Review of processing and analysis methods for DNA methylation array data. *Br J Cancer,* 109**,** 1394-402.

Wilson, C. M., Tobin, S. & Young, R. C. 2004. The exploding worldwide cancer burden: the impact of cancer on women. *Int J Gynecol Cancer,* 14**,** 1-11.

Winkelstein, W. J. 2004. Vignettes of the History of Epidemiology: Three Firsts by Janet Elizabeth Lane-Claypon. *Am J Epidemiol*.

Wise, L. A., Titus-Ernstoff, L., Newcomb, P. A., Trentham-Dietz, A., Trichopoulos, D., Hampton, J. M. & Egan, K. M. 2009. Exposure to breast milk in infancy and risk of breast cancer. *Cancer Causes Control,* 20**,** 1083-90.

Wnuk, M., Lewinska, A., Gurgul, A., Zabek, T., Potocki, L., Oklejewicz, B., Bugno-Poniewierska, M., Wegrzyn, M. & Slota, E. 2013. Changes in DNA methylation patterns and repetitive sequences in blood lymphocytes of aged horses. *Age (Dordr)*.

Wong, C. M., Anderton, D. L., Smith-Schneider, S., Wing, M. A., Greven, M. C. & Arcaro, K. F. 2010. Quantitative analysis of promoter methylation in exfoliated epithelial cells isolated from breast milk of healthy women. *Epigenetics,* 5**,** 645-55.

Woodfield, G. W., Hitchler, M. J., Chen, Y., Domann, F. E. & Weigel, R. J. 2009. Interaction of TFAP2C with the estrogen receptor-alpha promoter is controlled by chromatin structure. *Clin Cancer Res,* 15**,** 3672-9.

Xiao, Z., Mak, A., Koch, K. & Moore, R. B. 2013. A molecular complex of bovine milk protein and oleic acid selectively kills cancer cells in vitro and inhibits tumour growth in an orthotopic rat bladder tumour model. *BJU Int,* 112**,** E201-10.

Xie, Y., Min, S., Harte, N. P., Kirk, H., O'Brien, J. E., Voorheis, H. P., Svanborg, C. & Hun Mok, K. 2013. Electrostatic interactions play an essential role in the binding of oleic acid with alpha-lactalbumin in the HAMLET-like complex: a study using charge-specific chemical modifications. *Proteins,* 81**,** 1-17.

Yan, L., Nass, S. J., Smith, D., Nelson, W. G., Herman, J. G. & Davidson, N. E. 2003. Specific inhibition of DNMT1 by antisense oligonucleotides induces re-expression of estrogen receptor-alpha (ER) in ER-negative human breast cancer cell lines. *Cancer biology & therapy,* 2**,** 552-556.

Yang, X., Han, H., De Carvalho, D. D., Lay, F. D., Jones, P. A. & Liang, G. 2014. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell,* 26**,** 577-90.

Yap, K. L., Li, S., Munoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M. J. & Zhou, M. M. 2010. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol Cell.* United States: 2010 Elsevier Inc.

Yee, L. D., Lester, J. L., Cole, R. M., Richardson, J. R., Hsu, J. C., Li, Y., Lehman, A., Belury, M. A. & Clinton, S. K. 2010. Omega-3 fatty acid supplements in women at high risk of breast cancer have dose-dependent effects on breast adipose tissue fatty acid composition. *Am J Clin Nutr,* 91**,** 1185-94.

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K. & Rotman, G. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol,* 21**,** 379-86.

Yilmaz, B., Sandal, S., Ayvaci, H., Tug, N. & Vitrinel, A. 2012. Genotoxicity profiles in exfoliated human mammary cells recovered from lactating mothers in Istanbul; relationship with demographic and dietary factors. *Mutat Res,* 749**,** 17-22.

Yoder, J. A., Walsh, C. P. & Bestor, T. H. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* England.

Yoon, J. Y. & Chitale, D. 2013. Adenomyoepithelioma of the breast: a brief diagnostic review. *Arch Pathol Lab Med,* 137**,** 725-9.

Yoshida, T., Eguchi, H., Nakachi, K., Tanimoto, K., Higashi, Y., Suemasu, K., Iino, Y., Morishita, Y. & Hayashi, S. 2000. Distinct mechanisms of loss of estrogen receptor alpha gene expression in human breast cancer: methylation of the gene and alteration of trans-acting factors. *Carcinogenesis,* 21**,** 2193-201.

Yu, H., Zhu, S., Zhou, B., Xue, H. & Han, J. D. 2008a. Inferring causal relationships among different histone modifications and gene expression. *Genome Res,* 18**,** 1314-24.

Yu, W., Gius, D., Onyango, P., Muldoon-Jacobs, K., Karp, J., Feinberg, A. P. & Cui, H. 2008b. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature.* England.

Yuchuan, H., Jie, Z., Dongliang, L., Ya, D., Changguo, W., Kouyan, M., Jingqiu, C., Yanrong, L., Bo, Q. & Fang, X. 2011. Circulating biomarkers of hazard effects from cigarette smoking. *Toxicol Ind Health,* 27**,** 531-5.

.

Zemach, A., Kim, M. Y., Silva, P., Rodrigues, J. A., Dotson, B., Brooks, M. D. & Zilberman, D. 2010a. Local DNA hypomethylation activates genes in rice endosperm. *Proc Natl Acad Sci U S A.* United States.

Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. 2010b. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science,* 328**,** 916-9.

Zemach, A. & Zilberman, D. 2010. Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Curr Biol.* England: 2010 Elsevier Ltd.

Zeng, J., Konopka, G., Hunt, B. G., Preuss, T. M., Geschwind, D. & Yi, S. V. 2012. Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet,* 91**,** 455-65.

Zhang, F. F., Cardarelli, R., Carroll, J., Fulda, K. G., Kaur, M., Gonzalez, K., Vishwanatha, J. K., Santella, R. M. & Morabia, A. 2011. Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. *Epigenetics,* 6**,** 623-9.

Zhao, L., Wang, L., Jin, F., Ma, W., Ren, J., Wen, X., He, M., Sun, M., Tang, H. & Wei, M. 2009. Silencing of estrogen receptor alpha (ERalpha) gene by promoter hypermethylation is a frequent event in Chinese women with sporadic breast cancer. *Breast Cancer Res Treat,* 117**,** 253-9.

Zhu, J., Adli, M., Zou, J. Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P. L., Bennett, D. A., Houmard, J. A., Muoio, D. M., Onder, T. T., Camahort, R., Cowan, C. A., Meissner, A., Epstein, C. B., Shoresh, N. & Bernstein, B. E. 2013. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell,* 152**,** 642-54.

Zhu, K., Davidson, N. E., Hunter, S., Yang, X., Payne-Wilks, K., Roland, C. L., Phillips, D., Bentley, C., Dai, M. & Williams, S. M. 2003. Methyl-group dietary intake and risk of breast cancer among African-American women: a case-control study by methylation status of the estrogen receptor alpha genes. *Cancer Causes Control,* 14**,** 827-36.

Zhu, L., Shi, G., Schmidt, C. M., Hruban, R. H. & Konieczny, S. F. 2007. Acinar Cells Contribute to the Molecular Heterogeneity of Pancreatic Intraepithelial Neoplasia. *Am J Pathol,* 171**,** 263-73.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S. 2007. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics,* 39**,** 61-69.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A. & Meissner, A. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature,* 500**,** 477-81.

Zoli, M. & Picciotto, M. R. 2012. Nicotinic regulation of energy homeostasis. *Nicotine Tob Res,* 14**,** 1270-90.

Zudaire, E., Cuesta, N., Murty, V., Woodson, K., Adams, L., Gonzalez, N., Martinez, A., Narayan, G., Kirsch, I., Franklin, W., Hirsch, F., Birrer, M. & Cuttitta, F. 2008. The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers. *J Clin Invest,* 118**,** 640-50.

Zurita, M., Lara, P. C., del Moral, R., Torres, B., Linares-Fernandez, J. L., Arrabal, S. R., Martinez-Galan, J., Oliver, F. J. & Ruiz de Almodovar, J. M. 2010. Hypermethylated 14-3-3-sigma and ESR1 gene promoters in serum as candidate biomarkers for the diagnosis and treatment efficacy of breast cancer metastasis. *BMC Cancer,* 10**,** 217.

7. APPENDIXES

Appendix 1: Knowledge regarding *ESR1* methylation analyses in human breast cancer before the start of this thesis. All primer locations are normalized to the transcription start site for *ESR1* on the genome 2006 (NCBI36/hg18), UCSC genome browser.

Reference (year)	Pathology	Cases (n)	Case details	Cont (n)	Control details	No. CpG	Distance from TSS	Assay	ESR1 expression correlated?	Main findings
(Lapidus et al., 1998)	Carcinoma	2	Primary and metastatic tumours, normal tissue	2	Cell line data	20 18 7 18 16 20	ER1 +383-579; ER2 +388-554; ER3 +560-680; ER4 +656-807; ER5 +717-835; ER6 +746-874	MSP/ SB	No	*ESR1* CGI was hypermethylated in ESR1-neg tumours and cell lines, but unmethylated in normal breast tissue and ESR1-pos cell lines.
(Iwase et al., 1999)	Carcinoma	56	Breast tumour biopsies – sporadic cases	0	n/a	 9 24	-1771-1480; +740-833	MSP/ SB	IHC	P0 – U in 15/56 (26.8%); P1 – U in 25/56 (44.6%. Hypermethylation at both promoter regions correlated with ESR1-neg status (P1, p=0.0002; P0, p=0.015). No correlation with other clinicopathological features. Heterogenous tumour types.
(Archey et al., 2002)	Carcinoma	40	All ESR1-neg tumours; 20 BRCA1-linked, 20 sporadic.	0	n/a	 14 11	+411-503; +758-802	MSP, MSRA and sequencing	Yes	Overall methylation higher in BRCA-linked tumours (p=0.031). Five CpG sites shown to be more significant, each >two-fold increase in methylation in BRCA1-

										linked tumours. Used sequencing as could not trust MSP results.
(Muller et al., 2003)	Carcinoma	65	Blood from 39 women with metastases; 26 primary tumours	10	Healthy control women	8	+403-451	Methylight	IHC	Age did not affect *ESR1* methylation, and no association was found with this and clinicopathological features.
(Widschwendter et al., 2004)	Carcinoma	148		0	n/a	13	+403-451	MethylLight	RT-PCR	ESR1 methylation was better than hormone receptor status (ER/PR) at predicting clinical response to tamoxifen
(Li et al., 2006)	Carcinoma	193	Breast tumour biopsies	0	n/a	16	+764-857	MSP	No	Higher methylation levels (89%) in pts <60 yr than >60 yr (79%, p=0.04). Higher in ductal vs. non-ductal tumours (86% vs. 60%, p=0.003). Lower methylation in ER/PR-pos than ER/PR-neg tumours. Higher methylation levels in mutant TP53 tumours than normal TP53 (p=0.027).

(Lee et al., 2007)	Carcinoma	100	Ductal lavage from 48 Korean DCIS and 52 North American DCIS pts.	31	Ductal lavage samples from 15 Korean and 16 Nth American normal breasts.	17	+713-833	QM-MSP	IHC	*ESR1* promoter methylated in 10% of women with DCIS. Statistically higher methylation levels of *ESR1* promoter in Korean than North American women with DCIS. Higher methylation in women with ER-neg DCIS.
(Mirza et al., 2007)	IDC	50	Breast tumour samples	5	Serum from 5 healthy age-matched women	18	ER4 Lapidus et al.	MSP	IHC	Significant association between DNA methylation in tumour and serum. Significant correlation between *ESR1* promoter methylation and loss of protein expression.
(Feng et al., 2007)	Carcinoma	90	Malignant breast samples	90	Normal adjacent tissue.	6	+569-693	Pyrose quencing	IHC	*ESR1* was not highly methylated in tumour samples or normal tissue, but higher methylation was seen in normal tissues.
(Lee et al., 2008)	DCIS	100	DCIS ductal lavage	100	Healthy matched women	17	+713-833	QM-MSP	IHC	*ESR1* promoter hypermethylated in 10%.
(Baker et al., 2008)	Mammary atypia	18	RFNPA; high-risk women a, prophylactic tamoxifen for 18/12	16	High-risk women with mammary atypia with no chemo-prevention	11	+768-802	MSP	IHC	*ESR1* promoter hypermethylation did not predict women who developed persistent mammary atypia on tamoxifen chemoprophylaxis.

(Suijkerbuijk et al., 2008)	Carcinoma	78	32 BRCA1 and 46 sporadic breast Ca	20	7 BRCA1 mutant carriers and 13 non-carriers	14	+376-474	QM-MSP	No	Less methylation was found in BRCA1 tumours than sporadic, but more methylation overall of *ESR1* promoter in tumours compared to controls.
(Martinez-Galan et al., 2008)	Carcinoma	166	106 pre-op sera of non-metastatic breast Ca; 60 postoperative sera.	108	34 benign breast disease; 74 healthy women	5	+564-656	QMS-PCR	No	*ESR1* methylation differed between breast cancer pts vs. controls (p=0.0112), but not pre- vs. post-op. *ESR1/14-3-3-σ* hyper-methylation predicted breast cancer (sens, 81%; spec, 88%; p<0.0001); *ESR1* methylation alone predicted ESR1-neg phenotype.
(Wei et al., 2008)	Carcinoma	120	Primary breast cancer biopsies; 59 ESR1-neg, 61 ESR1-pos.	0	n/a	20; 7; 18; 16	ER1, ER3, ER4, ER5 Lapidus et al.	MSP	RT-PCR	*ESR1*promoter methylation correlated with *BRCA1* promoter methylation. Methylated tumours tended to be higher grade (p=0.04) and ESR1-neg (p<0.00001).
(Buyru et al., 2009)	Carcinoma	77	Breast tumour biopsies from Turkish women	77	Normal breast tissue from the same cohort	9	+740-791	MS-MLPA	No	The *ESR1* promoter was only methylated in 2 patients (2.6%).

(Sharma et al., 2009)	Carcinoma	101	Archived breast tissue (FFPE); 57 ER-, 44 ER+.	0	None	18	ER4 Lapidus et al.	MSP	No	*ESR1* promoter was hypermethylated in 64%. Patients with high levels of methylation had a poorer prognosis ($p<0.001$), more likely to have lymph node metastasis ($p=0.049$). ESR1 hypermethylation correlated with loss of PR expn ($p<0.01$).
(Gaudet et al., 2009)	Carcinoma	200	Breast tumour samples	200	Normal tissue from same cohort	5 8 11	PA: -34-+96; PB: +376-464; PC: -1760-1643	MethyLight	IHC	Mild to moderate methylation at *ESR1* promoter A and B, wide ranges of methylation for promoter C, which were weakly correlated with protein expression.
(Zhao et al., 2009)	Carcinoma	138	Sporadic cancer in Chinese women	14	Benign breast biopsies	20; 7; 18; 16	ER1, ER3, ER4, ER5 Lapidus et al.	MSP	IHC	*ESR1* was methylated in 60.1% tumours, and 82.6% ER-neg tumours ($p<0.00001$). Methylation was higher at 4 specific CpG sites in cancer samples vs. controls ($p<0.00001$). ESR1 expression was negatively correlated with increasing methylation.
(Zurita et al., 2010)	Carcinoma	34	Sera from 34 pts with metastatic disease	77	Sera from disease-free women in remission		+564-656	MSP	No	*ESR1* was unmethylated in this cohort, but the data is not given to show specifically what the results were.

					5				
(Ramos et al., 2010)	Carcinoma	89	Breast tumour samples	0	n/a	16	+764-856	MSP	IHC	*ESR1* was hypermethylated in 41% of tumours, and ESR1 expression decreased with increasing methylation levels (p<0.0001). *ESR1* and CXCL2 CGI4 promoter methylation coexisted frequently (p=0.022).
(Muggerud et al., 2010)	DCIS	89	27 DCIS; 28 IDC; 34 IDCs with a DCIS componen t	5	Normal breast tissue samples	26	+702-913	Pyrose quencin g	No	No methylation of *ESR1* was observed in either the controls or tumours samples in this site.
(Matusche k et al., 2010)	Carcinoma	85	Free tumour-related DNA and circulating tumour cells	22	Healthy volunteers		FULL TEXT OF PAPER UNAVAILA BLE	EUR J MED RES		ESR1 was hypermethylated in 38% tumours vs. 23% controls. No correlation between ESR1 hypermethylation and Her2 status or elevated CA15.3.

CpG, cytosine-guanine dinucleotide; IHC, immunohistochemistry; MSP, methylation specific PCR; MSRA, methylation sensitive restriction analysis; PCR, polymerase chain reaction; PBL, peripheral blood leucocyte; QM-MSP, quantitative multiplex MSP; QMS-PCR, quantitative methylation specific PCR; RPFNA: random periareolar fine needle aspiration; RT-PCR, real-time PCR; SB, Southern blot; TSS, transcription start site.

Appendix 2: Primers for semi-nested pyrosequencing assays across ESR1.

Appendix 3: RT-PCR primers designed to amplify pairs of sense and antisense regions in MVR regions of ESR1.

Site	Direction	Primer sequence
P2 (intron 1) sense	Forward	CCAGGACACAGGAGACCACT
P2 (intron 1) sense	Reverse	TCTCCTCTCCCTCCCATTTT
P2 (intron 1) antisense	Forward	AAAATGGGAGGGAGAGGAGA
P2 (intron 1) antisense	Reverse	AGTGGTCTCCTGTGTCCTGG
P3 (intron 1) sense	Forward	TTTGTTTTCCCAGCAGTTCA
P3 (intron 1) sense	Reverse	ACCAGAATCTCCTGCATGTG
P3 (intron 1) antisense	Forward	CACATGCAGGAGATTCTGGT
P3 (intron 1) antisense	Reverse	TGAACTGCTGGGAAAACAAA
P4 (intron 2) sense	Forward	**G**CAGCTCTCCCTGTTTTCCT
P4 (intron 2) sense	Reverse	AAAGTGTTTGGCTTGAAAGCA
P4 (intron 2) antisense	Forward	TGCTTTCAAGCCAAACACTTT
P4 (intron 2) antisense	Reverse	AGGAAAACAGGGAGAGCTGC
Exon 2 sense	Forward	TTGGCCAGTACCAATGACAA
Exon 2 sense	Reverse	TGCTTCAGGCTACCATTATGG
Exon 2 sense	Forward	CCATAATGGTAGCCTGAAGCA
Exon 2 sense	Reverse	TTGTCATTGGTACTGGCCAA

Appendix 4: Additional primers designed to amplify regions of interest across *ESR1* in human biopsy material.

Site	Primer	Primer sequence
P1	F1	ATTTTGTTTGGTTGAAGTATAGGT
P1	R1	gacGGGACACCGCTGATCGTTTACAAATCATAACTTCCCCCATAAATTCA
P1	S1	TTTGGTTGAAGTATAGGTT
Prom1	F1	GGGTTGTGTTTTTTTTTTAGGTG
Prom1	R1	gacGGGACACCGCTGATCGTTTAACCATCCCAAATACTTTAATATAAAAAAT
Prom1	S1	GGGGATAAGGTTTGTATT
Prom 2	F1	AGTATTTGGGATGGTTTTATTGTATTAGA
Prom 2	R1	gacGGGACACCGCTGATCGTTTAACTTACTACTATCCAAATACACCT
Prom 2	S1	GGATGGTTTTATTGTATTAGAT
Prom 3	F1	GTTTGGGGGGTTTTTTTTTATTTAA
Prom 3	R2	gacGGGACACCGCTGATCGTTTATCTCCAAATAATAAAACACCTACTAAC
Prom 3	S1	GGGGGGTTTTTTTTTATTTAAT
P2	F1	GGTGAATATTTTTTTAGGTGGTTATAGGAT
	R1	gacGGGACACCGCTGATCGTTTAACCTAATTCAAAACCCAACCAT
	S1	GGTTATAGGATAGGTGTATT
P3	F1	AGGAATTTTTGAGTATTGGTTTAAATAGAA
	R1	gacGGGACACCGCTGATCGTTTAACTAAAACAAACCATATCCCTAATACTT
	S1	AGTTTTTAGTAAAGTGTATGGTA
P4	F1	TTTAAGGATGGTGAAGGGTTTGAT
	R1	gacGGGACACCGCTGATCGTTTAACCAAAACACTTAAAACATAAAAATAATCT
	S1	ATGGTGAAGGGTTTGATA
P5	F1	AGTTTTTATATTGAAGGGAAATTATTTGT
	R1	gacGGGACACCGCTGATCGTTTACTATTCAATATCTATTCACCACTTATACTC
	S1	ATATTGAAGGGAAATTATTTGTA
P6	F1	TTTTTAGAATGTGTTTGGTTAGAGAT
	R1	gacGGGACACCGCTGATCGTTTAACAATAACTTCCCTAAATACTCCA
	S1	AGAGATTTTGATGATTGGT
P7	F1	TTGGTTAGGAATAATAGGATATTAGTGAGA
	R1	gacGGGACACCGCTGATCGTTTAACTTAACAATTCTCCTTTCTTCTTATC
	S1	ATTTTTGTTGGAAGAGT
P8	F1	AGGGTTATAGGGTAAGTGTTAGA
	R1	gacGGGACACCGCTGATCGTTTACTACCCCAAAACACATACTTTAAAC
	S1	AGTGTTAGAGAAGGTATAGA

Appendix 5: Primer sets used in this study for bisulphite pyrosequencing and qRT-PCR of smoking-related genes.

Name	Type	Sequence
Human pyrosequencing		
Universal tag		[BIOTIN-TAG]-gacGGGACACCGCTGATCGTTTAA
AHRR_p1	Forward	AGGGGTTGTTTAGGTTATAGAT
	Reverse	gacGGGACACCGCTGATCGTTTAAAACAAAACTACCCTCTCAAAA
	Sequencing	GGGTTGTTTAGGTTATAGATT
AHRR_p3	Forward	GGTTGGTGGTGTAGGATATATT
	Reverse	gacGGGACACCGCTGATCGTTTAACCCATCCTACCCAAATCCTAA
	Sequencing	GGTTTTAGGTTTAGGGA
2q37_p1	Forward	GATGGTGTTATAGGTGAGT
	Reverse	gacGGGACACCGCTGATCGTTTAAAACTCCTACCAAAACTTCTACT
	Sequencing	GGTGTTATAGGTGAGTTTTT
2q37_p3	Forward	GGTTTTTTAGTGGGTTTTTGTATGGT
	Reverse	gacGGGACACCGCTGATCGTTTAAACTAAAACCTCTACACCCAAA
	Sequencing	GGGAAGAGTAGAGTTAG
F2RL3	Forward	TGGGTTGGGTGTTTATTAGG
	Reverse	gacGGGACACCGCTGATCGTTTACAACAACAACACTAAACCATAC
	Sequencing	GTTTTGGTGGTGGGG
Human RT-PCR		
F2RL3_HUMAN_F	Forward	CCAGCGTCTACGACGAGAG
F2RL3_HUMAN_R	Reverse	CAGGGTGTCACTGTCATTGG
AHRR_HUMAN_F	Forward	CCCTGACCTTGTCCTTGACC
AHRR_HUMAN_R	Reverse	CTTTGTGGGTCCTGGAGTCT
GAPDH_F	Forward	TCCCATCACCATCTTCCA
GAPDH_R	Reverse	CATCACGCCACAGTTTCC
Mouse RT-PCR		
AHR_F	Forward	AGCATCATGAGGAACCTTGG

AHR_R	Reverse	GGATTTCGTCCGTTATGTCG
AHRNT_F	Forward	GGTTCGGGAGCACACTGTAT
AHRNT_R	Reverse	CCTTCCTTCACAGAGCCAAG
AHRR_F	Forward	CTGCCCAGGTACTCTGAACC
AHRR_R	Reverse	AGTGTCCACAAAGCCTGACC
CYP1A1_F	Forward	GGCCACTTTGACCCTTACAA
CYP1A1_R	Reverse	CAGGTAACGGAGGACAGGAA
Rpl_F	Forward	
Rpl_R	Reverse	

APPENDIX 6: Top hits in both case-control studies from the Illumina 450K array. Probes that were overlapping between the two data sets are highlighted in bold. * P-value (and FDR) for multivariate logistic regression for methylation with smoking status adjusting for age and batch. ** adjusted for age, batch and case-control status. FDR, false discovery rate.

Breast cancer EWAS

TargetID	P*	FDR*	P**	Never	Former	Current	Effect size	Chr	MAPINFO	Strand	Gene	Relation to CGI
cg05575921	2.31E-15	1.12E-09	4.31E-15	0.84	0.79	0.68	0.17	5	373378	F	AHRR	N_Shore
cg05951221	1.80E-13	4.36E-08	3.04E-13	0.39	0.33	0.28	0.11	2	233284402	F	2q37.1	Island
cg21566642	3.73E-13	6.02E-08	6.41E-13	0.44	0.37	0.32	0.12	2	233284661	R	2q37.1	Island
cg06126421	4.96E-11	6.02E-06	7.09E-11	0.65	0.61	0.57	0.08	6	30720080	R	6p21.33	
cg03636183	8.38E-11	8.12E-06	1.62E-10	0.64	0.61	0.56	0.08	19	17000585	R	F2RL3	N_Shore
cg01940273	1.47E-10	1.19E-05	2.38E-10	0.58	0.56	0.49	0.09	2	233284934	R	2q37.1	Island
cg21161138	5.44E-08	3.77E-03	8.60E-08	0.66	0.65	0.60	0.05	5	399360	R	AHRR	
cg06644428	6.17E-07	3.74E-02	8.45E-07	0.07	0.05	0.05	0.02	2	233284112	R	2q37.1	Island
cg25949550	8.50E-07	4.58E-02	1.14E-06	0.11	0.10	0.09	0.02	7	145814306	F	CNTNAP2	S_Shore
cg23436918	2.27E-06	1.10E-01	2.39E-06	0.06	0.07	0.07	-0.01	7	16685595	F	BZW2	Island
cg25189904	3.07E-06	1.29E-01	2.84E-06	0.37	0.36	0.33	0.05	1	68299493	F	GNG12	S_Shore
cg24620673	3.36E-06	1.29E-01	3.13E-06	0.13	0.14	0.15	-0.01	4	66535655	R	EPHA5	N_Shore
cg23576855	3.46E-06	1.29E-01	3.99E-06	0.66	0.64	0.50	0.16	5	373299	R	AHRR	N_Shore
cg26718213	5.55E-06	1.92E-01	8.07E-06	0.24	0.23	0.31	-0.08	2	241976080	R	SNED1	Island

cg09935388	6.40E-06	2.07E-01	5.55E-06	0.64	0.63	0.59	0.05	1	92947588	F	GFI1	Island
cg19717773	6.98E-06	2.12E-01	1.16E-05	0.55	0.51	0.48	0.07	7	2847554	F	GNA12	
cg05145534	8.18E-06	2.33E-01	1.14E-05	0.08	0.08	0.10	-0.02	13	113097975	F	13q34	

Colon cancer EWAS

TargetID	P*	FDR	P**	Never	Former	Current	Effect size	Chr	MAPINFO	Strand	Gene	Relation to CGI
cg05575921	1.73E-09	8.40E-04	1.86E-09	0.84	0.81	0.70	0.14	5	373378	F	AHRR	N_Shore
cg01940273	5.96E-09	1.44E-03	5.87E-09	0.62	0.59	0.54	0.08	2	233284934	R	2q37.1	Island
cg21566642	4.41E-08	7.12E-03	4.81E-08	0.51	0.47	0.39	0.12	2	233284661	R	2q37.1	Island
cg05951221	1.83E-07	2.22E-02	1.81E-07	0.41	0.37	0.34	0.07	2	233284402	F	2q37.1	Island
cg03636183	3.84E-07	3.73E-02	4.08E-07	0.68	0.65	0.61	0.07	19	17000585	R	F2RL3	N_Shore
cg15342087	1.24E-06	1.00E-01	1.34E-06	0.84	0.84	0.81	0.03	6	30720209	R	6p21.33	
cg11554391	1.66E-06	1.15E-01	1.80E-06	0.18	0.16	0.15	0.03	5	321320	F	AHRR	Island
cg23574443	1.96E-06	1.19E-01	2.13E-06	0.86	0.85	0.84	0.02	17	8263941	R	LOC100128288	
cg06126421	2.46E-06	1.33E-01	2.66E-06	0.71	0.70	0.64	0.07	6	30720080	R	6p21.33	
cg25648203	3.96E-06	1.92E-01	4.19E-06	0.79	0.79	0.76	0.03	5	395444	R	AHRR	
cg07558006	4.66E-06	2.02E-01	4.58E-06	0.92	0.91	0.91	0.01	6	132874263	R	TAAR8	
cg21913886	5.00E-06	2.02E-01	5.35E-06	0.84	0.83	0.80	0.04	1	15485346	F	TMEM51	S_Shelf
cg14385325	6.60E-06	2.21E-01	7.06E-06	0.31	0.32	0.32	-0.02	17	74564821	F	ST6GALNAC2	

cg09796170	6.72E-06	2.21E-01	7.15E-06	0.71	0.68	0.67	0.03	1	156548174	R	NA	TTC24	
cg25292882	6.82E-06	2.21E-01	7.24E-06	0.83	0.82	0.81	0.02	15	39431467	F	NA	15q14	
cg21161138	8.29E-06	2.29E-01	8.87E-06	0.72	0.71	0.68	0.04	5	399360	R	NA	AHRR	
cg18791779	8.42E-06	2.29E-01	8.57E-06	0.05	0.06	0.08	-0.03	6	166689411	F	NA	6q27	
cg21611682	8.51E-06	2.29E-01	8.44E-06	0.55	0.54	0.52	0.03	11	68138269	F	NA	LRP5	
cg23576855	9.66E-06	2.47E-01	9.82E-06	0.73	0.68	0.53	0.20	5	373299	R	NA	AHRR	N_Shore

Appendix 7: R code use to assess the influence of batch effect between the three chips for Illumina 450K HumanBeadChip methylation array for the 36 DNA samples from donor EBM.

```
###Batch 1 = n=12 EBM samples done on one chip in February 2014
batch1<-read.table("Batch1_beta2.txt", sep="\t", header=T)
batch1r<-read.table("Batch1_beta2_w_rownames.txt", sep="\t", header=T)
head(batch1r)
###Batch 2 = n=24 EBM samples done on two chips in July 2014
batch2<-read.table("EBM_Batch2_beta2.txt", sep="\t", header=T)
head(batch2)
EBM_beta_annot1<-read.table("Batch1_annot2.txt", sep="\t", header=T)
EBM_beta_annot2<-read.table("EBM_Batch2_annot2.txt", sep="\t", header=T)
rownames(batch1r)=EBM_beta_annot1$TargetID
rownames(batch2)=EBM_beta_annot2$TargetID
### Purpose = to merge data from all three chips for analysis.
###Check the correct data is loaded
batch1r[1:2,]
batch2[1:2,]
head(batch2)
head(rownames)
###Rownames need to be the cg IDs.
EBM_beta_annot1<-read.table("Batch1_annot2.txt", sep="\t", header=T)
EBM_beta_annot2<-read.table("EBM_Batch2_annot2.txt", sep="\t", header=T)
rownames(batch1r)=EBM_beta_annot1$IlmnID
rownames(batch2)=EBM_beta_annot2$IlmnID
intersectbetas=intersect(rownames(batch1r),rownames(batch2))
####How many rownames intersect between the two batches?
length(intersectbetas)                ####477201
batch1new=batch1r[intersectbetas,]
batch2new=batch2[intersectbetas,]
#check same number of rows in the two datasets
dim(batch1new)
dim(batch2new)
rownames(batch1r)=EBM_beta_annot1$TargetID
```

```
mergedata=cbind(batch1new, batch2new)
head(mergedata)
###annotates which sample in table is from which chip. adjust if any samples were removed
chip=c(rep(1,12), rep(2,12), rep(3,12))
###Run FUNCTION CODE for COMBAT (given below code for creation of pdf)
mergedata_COMBAT=CorrectBatchEffect(mergedata, chip, use.M=TRUE)
pdf("COMBAT_EBM.pdf", height=10, width=20)
par(mfrow=c(1,2))
PlotDensityByLevel(mergedata,2, as.factor(chip))
PlotDensityByLevel(mergedata_COMBAT,2, as.factor(chip))
dev.off()
all_batches<-cbind(intersectbetas,mergedata_COMBAT)
write.table(all_batches,"all_batches2.txt", sep="\t", quote=F)
write.table(mergedata,"all_batches.txt", sep="\t")

############code for COMBAT and plot density functions
ToBetaValues <- function(m.values) {
   1 / (1 + 2**-m.values)
}
ToMValues <- function(beta.values, eps=1E-12) {
   log2((beta.values + eps) / (1 - beta.values + eps))
}
#####plot functions
PlotDensityByLevel <- function(X, margin, factor) {
   if (margin == 2) {
      X <- t(X)
   }
   factor <- as.factor(factor)
   if (length(factor) != nrow(X)) {
      stop("Invalid factor")
   }
   densities <- by(X, factor, function(x) {
      density(data.matrix(x), na.rm=TRUE)
   })
```

```
    x.range <- range(lapply(densities, function(d) range(d$x)))
    y.range <- range(lapply(densities, function(d) range(d$y)))
    palette <- rainbow(nlevels(factor) + 1)
    for (i in 1:nlevels(factor)) {
        if (i == 1) {
            plot(densities[[i]], col=palette[i], xlim=x.range, ylim=y.range,
                main="", xlab="", ylab="Density")
            legend("topleft", levels(factor), col=palette, lty=1)
        } else {
            lines(densities[[i]], col=palette[i])
        }
    }
}
###Code for combat
FindParametricAdjustment <- function(X, gamma.hat, delta.hat, gamma.bar, t2, a,
                                     b, tol=1E-4) {
    n <- apply(!is.na(X), 1, sum)
    gamma <- gamma.hat
    delta <- delta.hat
    change <- 1
    while (change > tol) {
        gamma.new <- (t2*n*gamma.hat + delta*gamma.bar) / (t2*n + delta)
        sqsum <- apply(sweep(X, 1, gamma.new)**2, 1, sum, na.rm=TRUE)
        delta.new <- (sqsum/2 + b) / (n/2 + a - 1)
        change <- max(
            abs(gamma.new - gamma) / gamma,
            abs(delta.new - delta) / delta
        )
        gamma <- gamma.new
        delta <- delta.new
    }
    list(gamma.star=gamma, delta.star=delta)
}
```

```
FindNonparametricAdjustment <- function(X, gamma.hat, delta.hat) {
  n <- apply(!is.na(X), 1, sum)
  result <- sapply(1:nrow(X), function(i) {
    gamma <- gamma.hat[-i]
    delta <- delta.hat[-i]
    sqsum <- sapply(gamma, function(g) sum((X[i,] - g)**2, na.rm=TRUE))
    l <- 1 / (2*pi*delta)**(n[i]/2) * exp(-sqsum / (2*delta))
    c(sum(gamma*l), sum(delta*l)) / sum(l)
  })
  list(gamma.star=result[1,], delta.star=result[2,])
}
CorrectBatchEffect <- function(X, batch,
                               method=c("parametric", "nonparametric"),
                               use.M=FALSE) {
  batch <- as.factor(batch)
  method <- match.arg(method, c("parametric", "nonparametric"))
  if (length(batch) != ncol(X)) {
    stop("Invalid batch")
  }

  if (use.M) {
    X <- ToMValues(X)
  }
  # Check for and remove unusable probes, i.e., probes for which at least one
  # batch has less than 2 finite values
  usable.values <- do.call(cbind, by(t(X), batch, apply, 2, function(x) {
    sum(is.finite(x))
  }))
  unusable.probes <- unique(which(usable.values < 2, arr.ind=TRUE)[,"row"])
  if (length(unusable.probes) > 0) {
    X <- X[-unusable.probes,]
    warning(sprintf("Removed %d unusable probe(s)",
                    length(unusable.probes)))
  }
```

```
# Compute means
probe.means <- apply(X, 1, mean, na.rm=TRUE)
# Compute pooled variances
X.centered <- X
for (b in levels(batch)) {
   X.centered[,batch==b] <- sweep(X[,batch==b], 1,
                              apply(X[,batch==b], 1, mean, na.rm=TRUE))
}
probe.pvars <- apply(X.centered, 1, var, na.rm=TRUE)
rm(X.centered)
# Normalize data
X <- sweep(sweep(X, 1, probe.means), 1, sqrt(probe.pvars), "/")
# Compute regression batch effect parameters
gamma.hat <- do.call(rbind, by(t(X), batch, apply, 2, mean, na.rm=TRUE))
delta.hat <- do.call(rbind, by(t(X), batch, apply, 2, var, na.rm=TRUE))
if (method == "parametric") {
   # Find priors
   gamma.bar <- apply(gamma.hat, 1, mean)
   t2 <- apply(gamma.hat, 1, var)
   a.prior <- apply(delta.hat, 1, function(x) {
      mu <- mean(x)
      sigma2 <- var(x)
      (2*sigma2 + mu**2) / sigma2
   })
   b.prior <- apply(delta.hat, 1, function(x) {
      mu <- mean(x)
      sigma2 <- var(x)
      (mu*sigma2 + mu**3) / sigma2
   })

   # Find parametric adjustment for each batch and apply it
   for (i in 1:nlevels(batch)) {
      b <- levels(batch)[i]
```

```
        adj <- FindParametricAdjustment(X[,batch==b], gamma.hat[i,],
                                        delta.hat[i,], gamma.bar[i], t2[i],
                                        a.prior[i], b.prior[i])
        X[,batch==b] <- sweep(sweep(X[,batch==b], 1, adj$gamma.star), 1,
                              sqrt(adj$delta.star), "/")       }
    } else {
       # Find nonparametric adjustment for each batch and apply it
       for (i in 1:nlevels(batch)) {
          b <- levels(batch)[i]
          adj <- FindNonparametricAdjustment(X[,batch==b], gamma.hat[i,],
                                             delta.hat[i,])
          X[,batch==b] <- sweep(sweep(X[,batch==b], 1, adj$gamma.star), 1,
                              sqrt(adj$delta.star), "/")     }    }
# Reverse initial normalization
    X <- sweep(sweep(X, 1, sqrt(probe.pvars), "*"), 1, probe.means, "+")
    if (use.M) {
       X <- ToBetaValues(X)
    }
    X}
#### ESR1 analysis.
allbatches<-mergedata_COMBAT
####To find the ESR1 probes of interest in the data
###Relevant CpG sites with SNPs removed
no_snps_esr1<-
c('cg08907436','cg07619683','cg07189962','cg07584093','cg10441070','cg17706972','cg013
21962','cg20893956','cg24764793','cg07746998','cg21157690','cg17264271','cg15543523','c
g26089753','cg08884395','cg01715172','cg21608605','cg20627916','cg07671949','cg231649
38','cg23165623','cg21614759','cg19411146','cg21950534','cg11813455','cg24900983','cg05
171584','cg23467008','cg22839866','cg23009221','cg13612689','cg27316393','cg00655307','
cg01777019','cg15980539','cg11251858','cg00920970','cg20253551','cg02285263','cg02720
618','cg04063345','cg15626350','cg00601836','cg06877423','cg03732055','cg21265702','cg1
0939667','cg09414638','cg19369424','cg07455133','cg12209876','cg25490334','cg02404255
','cg03037684','cg07059469')
esr1<-mergedata_COMBAT[no_snps_esr1,]
write.table(esr1,"My_did_it.txt", sep="\t", quote=F).
```

Appendix 8: Raw data from the OncoArray for the 80 EBM samples.

ns.snps<-
c("rs10759243","rs2236007","rs2981579","rs614367","rs3757318","chr14_37132769_A_G")

Donor ID	PAX9 (T/C)	RAD23B (A/C)	FGFR2 (A/G)	ESR1 (A/G)	CCND1 (C/T)
	rs2236007	rs10759243	rs2981579	rs3757318	rs614367
1	TT	CC	GG	GG	CC
18	TT	CC	GG	GG	CC
21	TC	CC	AG	GG	0
27	TC	CC	AG	GG	CC
29	CC	CC	AG	GG	TC
35	TC	CC	GG	GG	TT
36	CC	CC	AG	GG	CC
39	CC	AC	AA	GG	CC
40	CC	CC	AG	GG	CC
42	CC	AC	GG	GG	CC
46	0	CC	AG	GG	CC
47	0	AA	GG	GG	CC
50	CC	AC	AA	GG	CC
51	TC	CC	AG	GG	CC
53	CC	CC	GG	GG	CC
54	TC	AA	AG	GG	CC
55	TC	AC	AG	GG	CC
59	TC	AC	GG	GG	CC
60	TC	AC	GG	GG	CC
63	TC	CC	AG	GG	CC
64	CC	AC	GG	AG	TC
65	CC	CC	GG	GG	CC
67	0	AC	AG	GG	CC
69	CC	CC	AG	GG	CC
70	CC	CC	AG	AG	TT
71	CC	AA	GG	AG	CC
72	TC	AC	AG	GG	CC
73	CC	CC	GG	GG	CC
74	CC	CC	AA	GG	CC
76	CC	AA	AG	GG	CC
77	TC	AA	GG	GG	CC
79	TC	AC	AG	GG	CC
80	TC	AC	AA	GG	TC
81	CC	AC	AG	GG	TC
82	TC	AA	GG	GG	CC
83	CC	CC	AG	GG	CC
85	TC	AC	GG	GG	CC
86	CC	AC	AG	GG	TC
89	TC	AC	AG	GG	CC
90	TC	AC	AG	GG	CC
91	CC	AA	GG	AG	TC
92	CC	CC	AG	GG	CC
94	CC	AC	AA	GG	TT

95	TC	CC	GG	GG	TC
96	0	CC	AG	GG	CC
97	TC	AC	GG	GG	CC
98	CC	AA	AG	GG	CC
100	CC	CC	GG	GG	CC
101	TC	CC	AA	GG	TC
102	CC	AC	AG	GG	CC
106	CC	CC	GG	GG	CC
107	CC	AC	AG	GG	CC
108	TC	CC	AG	GG	CC
109	0	CC	AG	GG	TC
111	CC	CC	AG	GG	CC
112	CC	AC	AG	GG	CC
113	0	CC	AG	GG	CC
114	TT	CC	GG	GG	CC
115	TC	CC	GG	AG	CC
118	CC	AC	AG	GG	CC
120	CC	CC	AG	GG	CC
121	TC	AC	AA	GG	CC
122	CC	CC	AG	GG	CC
124	CC	CC	AG	GG	CC
127	CC	CC	GG	GG	CC
128	TC	AC	GG	GG	CC
129	TC	CC	GG	GG	TC
131	TC	AC	AG	GG	CC
133	CC	CC	AA	GG	CC
134	TC	AC	GG	GG	CC
135	TC	AC	AG	AG	TC
136	0	AC	GG	GG	TC
138	CC	AC	AG	AG	TC
139	CC	CC	GG	GG	CC
141	CC	AA	AG	GG	TC
144	0	AC	GG	GG	TC
145	TC	CC	GG	GG	CC
146	0	AC	GG	AG	CC
147	TC	AA	AG	GG	CC
153	0	CC	AG	GG	CC

Appendix 9: Full texts of publications arising from this thesis.