

Epigenetic Mechanisms of Breast Cancer Risk Exposures

Annelie Ewa Johansson

The Epigenetics Unit, Division of Cancer, Department of Surgery and Cancer,
Imperial College London

A thesis submitted for the degree of Doctor of Philosophy
March 2019

DECLARATION OF ORIGINALITY

I, Annelie Johansson, declare that all work presented within this thesis is my own, if not otherwise stated.

COPYRIGHT DECLARATION

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

PUBLICATIONS ARISING FROM THIS PROJECT

Annelie Johansson, Domenico Palli, Giovanna Masala, Sara Grioni, Claudia Agnoli, Rosario Tumino, Maria Concetta Giurdanella, Francesca Fasanelli, Carlotta Sacerdote, Salvatore Panico, Amalia Mattiello, Silvia Polidoro, Michael Jones, Minouk Schoemaker, Nick Orr, Katarzyna Tomczyk, Nichola Johnson, Olivia Fletcher, Vittorio Perduca, Laura Baglietto, Pierre-Antoine Dugué, Melissa C. Southey, Graham G. Giles, Dallas R. English, Roger L. Milne, Gianluca Severi, Srikant Ambatipudi, Cyrille Cuenin, Veronique Chajès, Isabelle Romieu, Zdenko Herceg, Anthony Swerdlow, Paolo Vineis and James M. Flanagan. *Epigenome-Wide Association Study for Lifetime Estrogen Exposure Identifies an Epigenetic Signature Associated with Breast Cancer Risk* (manuscript under review, submitted January 2019)

Clara Bodelon*, Srikant Ambatipudi*, Pierre-Antoine Dugué*, **Annelie Johansson***, Joshua N. Sampson, Belynda Hicks, Eric Karlins, Amy Hutchinson, Melissa C. Southey, Graham G. Giles, Silvia Polidoro, Zdenko Herceg**, James M. Flanagan**, Roger L. Milne** and Montserrat Garcia-Closas**. *Blood DNA methylation and breast cancer risk: a meta-analysis of four prospective cohort studies* (manuscript under review, submitted January 2019) *Co-first authors; **Co-last authors

Annelie Johansson, Domenico Palli, Giovanna Masala, Sara Grioni, Claudia Agnoli, Rosario Tumino, Maria Concetta Giurdanella, Francesca Fasanelli, Carlotta Sacerdote, Salvatore Panico, Amalia Mattiello, Anthony Swerdlow, Minouk Schoemaker, Michael Jones, Nick Orr, Olivia Fletcher, Nichola Johnson, Katarzyna Tomczyk, Paolo Vineis and James M. Flanagan. *DNA Methylation Index of Lifetime Estrogen Exposure in Breast Cancer*. Poster presented at: American Association in Cancer Research (AACR) Annual Meeting, 14-20 April 2018, Chicago, USA.

Annelie Johansson and James M. Flanagan. *Epigenome-wide association studies for breast cancer risk and risk factors*. Trends Cancer Res, 2017. **12**: p. 19-28

ACKNOWLEDGEMENTS

This PhD project would never have been completed without help from several people and collaborators. Firstly, many thanks to my supervisors Dr James Flanagan and Professor Paolo Vineis for giving me the opportunity to do this PhD project, and all the support to complete the PhD thesis in time. Thanks James for all the weekly meetings and discussions, for pushing my limits now and then, and for always having the door open for any stupid questions. I would never have completed this PhD without your supervision. Also, thank you to Breast Cancer Now for sponsoring this PhD project.

Many thanks to all the people who have joined and left the Epigenetics Unit over the years. Especially to the three PhD students who started with me and who I have shared all the PhD struggles with: Francesca Galea, Caitriona Tyndall, and John Gallon. Thanks to Fran for all the discussions about regression methods, P-values, and other EWAS-related issues, and for all the support, encouragement, and enthusiasm when it has felt as most hopeless. Thanks to Cat for all the help with various lab issues (I would never have got these western blots together without your help), for your organisation skills that has kept the group together, and for letting us stay in your house during the trip to Dublin. Thanks to John for your support with different bioinformatic and lab-related issues, for always making us laugh, but also for your smart inputs at the group meetings. I am sure all three of you will successfully complete your PhDs and that your future will be even more successful.

I owe several other people a huge thank you for helping me in the laboratory and/or providing support over the years. Thanks to Nahal Masrour for helping me get around in the lab and for your pyrosequencing skills, and thanks to Dr Nair Bonito for teaching me the cell culture techniques. Thanks to Dr Emma Bell and Dr Ed Curry for organising the bioinformatics meetings, and also, thanks to Marina, Ian, Chail, and all the other members of the Epigenetic's Unit and 4th floor at the IRDB.

This project has included several external collaborators. Thanks to the Generations Study investigators; Dr Olivia Fletcher for providing DNA samples and helpful input to the manuscript, Dr Nichola Johnson and Dr Katarzyna Tomczyk for managing the logistics with the DNA samples, and Dr Minouk Schoemaker, Dr Michael Jones, and Professor Anthony Swerdlow for all your helpful inputs to the manuscript and analyses. Thanks to all EPIC-Italy, EPIC-IARC and MCCS investigators for providing data for this project, and especially thanks to Dr Srikant Ambatipudi, Dr Vittoria Perduca and Dr Laura Baglietto for participating in the meta-analysis. Also thanks to Dr Clara Bodelon for the

inclusion in your manuscript. Thanks to Heather Thorne and Lynda Williams from kConFab to providing the LCL samples. Thanks to Professor Simak Ali and Dr Manikandan Periyasamy for providing antibodies and the cell lines for the controls in the western blot. Thanks to Fluidigm and the MRC Genomics Facility for all the help with the targeted sequencing.

Lastly, but absolutely not least, I would like to thank all my backing from home. Special thanks to my parents Eva and Björne for a lifetime of love and support. Thanks to my biggest emotional support Lisa, for being the best of friends and for always being there when I needed to talk. Thanks to all my other cheerleaders at home, especially Simone and Sandra. Thanks to Rosemary, Adrian, Beau, and Baz for welcoming me in your family, making me feel like home, and taking care of me in London. And the biggest thank you to my Alastair, for always believing in me, being there for me, for all the support and encouragements, especially when completing this thesis, and for being my safety in London.

With these words, I have now finalised this PhD. I am looking forward to my future work as a post doctoral researcher at Karolinska Institute in Stockholm and to continue my contribution in reducing the number of women affected from breast cancer.

ABSTRACT

The breast cancer incidence rate is increasing and risk prediction needs to be improved. Higher lifetime oestrogen exposure is associated with increased risk, however, the mechanisms between oestrogen exposure and breast cancer development are not clear. It was hypothesised that a woman's lifetime oestrogen exposure accumulates DNA methylation changes, detectable in blood, which can be used for risk assessment for breast cancer.

An epigenome-wide association study (EWAS) of an estimated lifetime oestrogen exposure (ELEE)-model was conducted in Illumina HumanMethylation450K (HM450K) data in EPIC-Italy (n=216). A breast cancer-associated Methylation Index (MI) of ELEE, including 31 CpG-sites, was developed using HM450K data from EPIC-Italy (n=237) and the Generations Study (n=65). The MI was validated using targeted bisulphite sequencing in the Generations Study (n=440 case-control pairs), and replicated in a meta-analysis including four study cohorts.

In the EWAS, 694 CpG sites were associated ($Q < 0.05$) with ELEE. The association between the MI and breast cancer risk was validated in the Generations Study targeted sequencing data ($OR_{Q4_vs_Q1} = 1.77$, 95%CI 1.07-2.93, $P = 0.027$) and replicated in the meta-analysis ($OR_{Q4_vs_Q1} = 1.45$, 95%CI: 1.05-2.00, $P = 0.024$). However, the correlation between the MI and ELEE was not replicated across study cohorts.

Furthermore, two mechanisms for oestrogen-induced blood DNA methylation changes were explored in lymphocyte cell-lines (LCLs): transcriptional response via the oestrogen receptors (ERs), and DNA repair response following DNA damage. No expression of the ERs was identified in the LCLs, but indications of increasing DNA damage following oestrogen-treatment.

In this project, a blood DNA methylation signature for breast cancer risk was identified, however, the signature for lifetime oestrogen exposure was not validated. Further replication of the EWAS and improvements to the MI are needed to assess if an epigenetic signature in blood can be incorporated in population risk screening. Targeted prevention for high-risk women has the potential to reduce the breast cancer incidences.

TABLE OF CONTENTS

ABBREVIATIONS.....	11
LIST OF TABLES.....	12
LIST OF FIGURES.....	13
CHAPTER 1: General Introduction.....	18
1.1. BREAST CANCER.....	18
1.1.1 Breast cancer statistics	18
1.1.2 Breast cancer subtypes	18
1.1.3 Breast cancer treatment.....	20
1.2. RISK AND PREVENTION.....	22
1.2.1 Genetic risk	22
1.2.2 Lifestyle risk factors.....	22
1.2.3 Reproductive risk factors	28
1.2.4 Breast cancer risk models	33
1.2.5 Prevention	35
1.3. EPIGENETICS.....	38
1.3.1 DNA methylation and breast cancer	39
1.3.2 Blood DNA methylation and breast cancer	40
1.3.3 Breast cancer risk factors and DNA methylation in blood	41
1.4. OESTROGEN EXPOSURE AND ITS ROLE IN BREAST CANCER	43
1.4.1 Oestrogen exposure and breast cancer.....	43
1.4.2 Mechanisms of oestrogen action.....	44
1.4.3 Oestrogens and DNA methylation	47
1.5. AIMS AND HYPOTHESES.....	51
CHAPTER 2: Epigenome-Wide Association Study of Estimated Lifetime	
Oestrogen Exposure	53
2.1. INTRODUCTION.....	53
2.1.1 EWAS for breast cancer risk using blood DNA	53
2.1.2 Previous breast cancer case-control EWAS.....	53
2.1.3 EWAS of risk factors using blood DNA.....	55
2.1.4 Estimation of lifetime exposure to oestrogen	59
2.1.5 Statistical methods in EWAS.....	60
2.1.6 Aims of this chapter	62
2.2. METHODS.....	63

2.2.1	Study cohort	63
2.2.2	Hormonal factors and breast cancer risk in EPIC-Italy	63
2.2.3	The ELEE model.....	65
2.2.4	EWAS of ELEE	67
2.2.5	EWAS of hormonal risk factors.....	69
2.2.6	Genome-wide DNA methylation and ELEE.....	69
2.3.	RESULTS.....	71
2.3.1	Hormonal factors are associated with breast cancer risk in EPIC-Italy..	71
2.3.2	The ELEE model is significantly associated with breast cancer risk	75
2.3.3	EWAS identifies 694 probes associated with the ELEE.....	75
2.3.4	There are few overlaps between the EWAS of hormonal risk factors....	81
2.3.5	Genome-wide DNA methylation and ELEE.....	85
2.4.	DISCUSSION.....	87
2.4.1	Hormonal risk factors and breast cancer.....	87
2.4.2	The ELEE model.....	88
2.4.3	EWAS of ELEE	89
2.4.4	Transcriptional mechanisms hypothesis	90
2.4.5	EWAS of hormonal risk factors.....	91
2.4.6	Genome-wide DNA methylation and ELEE.....	91
2.4.7	Conclusions.....	92
CHAPTER 3: Methylation Index of ELEE and Breast Cancer Risk.....		94
3.1.	INTRODUCTION	94
3.1.1	Replication and validation in methylation studies	94
3.1.2	Model development in DNA methylation array data.....	95
3.1.3	Aims of this chapter	96
3.2.	STUDY COHORTS	97
3.2.1	EPIC-Italy	97
3.2.2	The Generations Study.....	97
3.3.	METHODS.....	102
3.3.1	Targeted bisulphite sequencing using the Fluidigm 48.48 Access Array	102
3.3.2	MI of ELEE.....	104
3.3.3	Individual CpG sites and breast cancer risk.....	106
3.4.	RESULTS.....	108
3.4.1	Test phase.....	108
3.4.2	Quality control of target sequencing data in the Generations Study.....	112

3.4.3	No observed differences in WBC composition between cases and controls.....	115
3.4.4	Exclusion of target CpG sites and samples in the Generations Study targeted sequencing data prior to MI development.....	115
3.4.5	The MI is correlated with ELEE in EPIC-Italy but not in the Generations Study	115
3.4.6	The MI is associated with breast cancer risk	118
3.4.7	The MI in pre- and postmenopausal women.....	121
3.4.8	The association between the MI and breast cancer risk is stronger in cases with longer time to diagnosis.....	121
3.4.9	Individual CpG sites and association with breast cancer	121
3.5.	DISCUSSION.....	125
3.5.1	Conclusions.....	127
CHAPTER 4: Meta-analysis of the MI and ELEE		129
4.1.	INTRODUCTION.....	129
4.1.1	Meta-analysis	129
4.1.2	Aims for this chapter.....	130
4.2.	STUDY COHORTS	131
4.2.1	The Generations Study.....	131
4.2.2	EPIC-Italy	131
4.2.3	EPIC-IARC	131
4.2.4	MCCS.....	131
4.3.	METHODS.....	134
4.3.1	Pre-processing of the HM450K data.....	134
4.3.2	Statistical analysis	134
4.4.	RESULTS.....	136
4.4.1	Baseline characteristics	136
4.4.2	The MI is associated with breast cancer risk	136
4.4.3	The MI is not associated with ELEE	139
4.5.	DISCUSSION.....	141
CHAPTER 5: Molecular Mechanisms Between Oestrogen Exposure and DNA Methylation		144
5.1.	BACKGROUND.....	144
5.1.1	Molecular mechanisms for oestrogen-induced DNA methylation	144
5.1.2	ER- α and ER- β expression in blood.....	146
5.1.3	Aims of this chapter	146

5.2. MATERIALS & METHODS	147
5.2.1 Cell culture	147
5.2.2 γ H2AX immunofluorescence imaging.....	150
5.2.3 Bisulphite pyrosequencing.....	151
5.2.4 Western blot for ER- α and ER- β	152
5.3. RESULTS.....	156
5.3.1 DNA damage following oestrogen treatment in LCLs	156
5.3.2 DNA methylation following oestrogen treatment in LCLs	156
5.3.3 The ERs are not detectable in LCLs	160
5.4. DISCUSSION.....	162
5.4.1 Conclusions.....	164
CHAPTER 6: Discussion, Future Research Directions and Conclusions	166
6.1. GENERAL DISCUSSION.....	166
6.1.1 Lifetime oestrogen exposure and DNA methylation	166
6.1.2 Blood DNA methylation and breast cancer risk	170
6.1.3 Lifetime oestrogen exposure, blood DNA methylation, and breast cancer risk.....	171
6.2. STRENGTHS AND LIMITATIONS	174
6.3. FUTURE DIRECTIONS	177
6.4. BREAST CANCER RISK AND PREVENTION.....	179
6.5. FINAL CONCLUSIONS	180
REFERENCES	181
APPENDICES.....	209

ABBREVIATIONS

2-OHE2	2-hydroxy-oestradiol
4-OHE2	4-hydroxy-oestradiol
AI	Aromatase inhibitor
AUC	Area under the curve
AR	Androgen receptor
BMI	Body mass index
BOADICEA	The Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm
ChIP-seq	Chromatin immunoprecipitation sequencing
CI	Confidence interval
CpG	Cytosine-guanine dinucleotide
CRT	Combined hormone replacement therapy
CVD	Cardiovascular disease
DCIS	Ductal carcinoma <i>in situ</i>
DNMT	DNA methyltransferases
DSB	Double-strand breaks
E2	17- β -oestradiol
EBV	Epstein-Barr virus
ELEE	Estimated lifetime oestrogen exposure
EPIC	The European prospective investigation into cancer and nutrition study
ERE	Oestrogen responsive elements
ERT	Oestrogen-only hormone replacement therapy
EWAS	Epigenome-wide association study
ER	Oestrogen receptor
FDR	False discovery rate
FFTP	First full-term pregnancy
FSH	Follicle-stimulating hormone
γ H2AX	Phosphorylation of the histone H2AX at serine 139
GP1R	G-protein coupled ER

GWAS	Genome-wide association study
HER2	Human epidermal growth receptor 2
HM27K	Illumina HumanMethylation27K array
HM450K	Illumina HumanMethylation450K array
HM850K	Illumina HumanMethylation850K (EPIC) array
HR	Hazard ratio
HRT	Hormone replacement therapy
IARC	International Agency for Research on Cancer
ICR	International Cancer Institute
IGF2	Insulin-like growth factor 2
ILC	Invasive lobular carcinoma
kConFab	Kathleen Cuninghame Foundation Consortium for research into Familial Breast Cancer
LCL	Lymphocyte cell line
LH	Luteinising hormone
LINE-1	Long interspersed nuclear elements
LUMA	Luminometric methylation assay
MCCS	Melbourne Collaborative Cohort study
MI	Methylation Index
OC	Oral contraceptives
OR	Odds ratio
ORperSD	Odds ratio per one standard deviation
PABC	Pregnancy-associated breast cancer
PAH	Polycyclic aromatic hydrocarbons
PARP	Poly(ADP-Ribose) polymerases
PBMNCs	Peripheral blood mononuclear cells
PLCO	The Prostate, Lung, Colorectal, and Ovarian cancer screening trial
PR	Progesterone receptor
PRS	Polygenic risk scores
QQ-plots	Quantile-Quantile plots

ROS	Reactive oxygen species
RR	Risk ratio
RRBS	Reduced representative bisulphite sequencing
SAM	S-adenosyl methionine
SDS-PAGE	Sodium dodecyl sulphate-polyacrylamide gel electrophoresis
SERM	Selective oestrogen receptor modulators
SHBG	Sex hormone binding globulin
SNP	Single nucleotide polymorphism
TDLU	Terminal duct lobular units
TNBC	Triple-negative breast cancer
TSS	Transcription start site
UTR	Untranslated region
WBC	White blood cell
WGBS	Whole-genome bisulphite sequencing
WHI	Women's Health Initiative

LIST OF TABLES

Table 1.1	Hormonal risk factors in breast cancer
Table 2.1	Table of characteristics for EPIC-Italy questionnaire data
Table 2.2	Table of characteristics for EPIC-Italy HM450K methylation data
Table 2.3	ELEE models and breast cancer risk in EPIC-Italy
Table 2.4	Top hits from the EWAS of ELEE
Table 3.1	Table of characteristics for EPIC-Italy and the Generations Study HM450K data
Table 3.2	Table of characteristics for case-control pairs in EPIC-Italy HM450K data
Table 3.3	Table of characteristics for the Generations Study subjects for targeted bisulphite sequencing
Table 3.4	The 42 CpG sites included in the targeted bisulphite sequencing
Table 3.5	Summary of the targeted sequencing data in the Generations Study
Table 3.6	The MI by quartile and breast cancer risk
Table 3.7	The MI in pooled breast and ovarian cancer cases and controls
Table 3.8	The MI, ELEE and breast cancer risk in pre- and postmenopausal women
Table 3.9	Target CpG sites associated with breast cancer risk
Table 4.1	Table of characteristics for the meta-analysis
Table 4.2	Meta-analysis of the MI and breast cancer risk
Table 4.3	Meta-analysis of the MI by the quartile and breast cancer risk
Table 4.4	Meta-analysis of the MI, breast cancer risk and time to diagnosis
Table 4.5	Correlation between the MI and ELEE across study cohorts
Table 5.1	The expression of oestrogen receptors in blood
Table 5.2	Summary of LCL samples from kConFab

LIST OF FIGURES

Figure 1.1	Breast cancer statistics
Figure 1.2	Mechanisms of oestrogen action
Figure 2.1	Beta-values and M-values in HM450K methylation array data
Figure 2.2	Hormonal breast cancer risk factors in EPIC-Italy
Figure 2.3	EWAS identifies CpG sites significantly associated with ELEE in EPIC-Italy
Figure 2.4	Inflation and bias in the EWAS of ELEE
Figure 2.5	Sensitivity analyses for the EWAS of ELEE
Figure 2.6	Functional genomic regions and histone modifications among EWAS of ELEE hits
Figure 2.7	EWAS of hormonal risk factors
Figure 2.8	Genome-wide DNA methylation and ELEE
Figure 3.1	Workflow for targeted bisulphite sequencing and data analysis
Figure 3.2	CpG probes for estimation of WBC composition
Figure 3.3	Test phase: quality control for targeted bisulphite sequencing using the Fluidigm 48.48 Access Array
Figure 3.4	Estimation of WBC composition in the Generations Study targeted sequencing data
Figure 3.5	Excluded target CpG sites in the Generations Study targeted sequencing data prior to development of the MI
Figure 3.6	The Methylation Index of ELEE
Figure 3.7	Receiver operating characteristic curves for the MI and breast cancer risk
Figure 3.8	Development of the MI in controls-only
Figure 3.9	The MI, breast cancer risk and time to diagnosis
Figure 4.1	Forest plot of the MI and breast cancer risk
Figure 4.2	Forest plot of the MI by the quartile and breast cancer risk
Figure 4.3	Forest plot of the MI, breast cancer risk and time to diagnosis
Figure 5.1	EBV-transformed Lymphocyte Cell Lines (LCLs)
Figure 5.2	Analysis of γ H2AX immunofluorescence using CellProfiler
Figure 5.3	DNA damage following oestrogen treatment in LCLs
Figure 5.4	DNA methylation in LCLs following oestrogen treatment
Figure 5.5	DNA methylation in LCLs following oestrogen treatment, phenotype specific

- Figure 5.6 DNA methylation in LCLs following oestrogen treatment, per individual sample
- Figure 5.7 Western blot of ER- α and ER- β in LCLs
- Figure 6.1 Estimating a woman's total lifetime oestrogen exposure

CHAPTER 1:

General Introduction

1 CHAPTER 1: General Introduction

1.1. BREAST CANCER

1.1.1 Breast cancer statistics

Breast cancer is the most common female cancer with over 55,000 diagnosed cases in the UK in 2015 [1]. It is estimated that over 2 million women will be diagnosed with breast cancer in 2018 worldwide [2]. The mortality rate has decreased by 38% in the UK since the 1970s after the introduction of population mammographic screening in the late 1980s, which allows detection of the disease at an early stage (**Figure 1.1**). This in combination with the advancement in treatment, such as targeted therapy with Tamoxifen, means that more women than ever survive their disease [1, 3, 4]. However, breast cancer is still the second most common cause of death from cancer after lung cancer, and with increasing incidence rate it is expected that the number of deaths due to this disease may yet increase [1]. Besides an aging population, differences in lifestyle and reproductive factors over the years are reasons for the increase in the incidence, and it is estimated that 23% of cases could be preventable [1, 5].

1.1.2 Breast cancer subtypes

Breast cancer is a heterogeneous disease and the biology, prognosis and treatment vary between patients [6]. Ductal carcinoma *in situ* (DCIS) is a non-invasive form of a ‘pre-cancer’ state of breast cancer developed inside the milk ducts. The incidence rate for DCIS has dramatically increased after the start of mammography routine screening, which detects DCIS that might never have been detected otherwise, or alternatively detects the disease before it has developed to invasive breast cancer [7]. Invasive breast cancers have the ability to spread to the surrounding tissues and are classified into ‘special type’ or ‘no special type’. No special type (previously known as invasive ductal carcinoma) is the most common type of breast cancer, makes up 60-70% of all cases, and as the name anticipates, has no distinct morphological features [8, 9]. Invasive lobular carcinoma (ILC) is the most common breast cancer of special type and comprises around 5-15% of all breast cancers. However, ILC is still a heterogeneous breast cancer type, and better understanding of breast cancer heterogeneity is needed [9, 10].

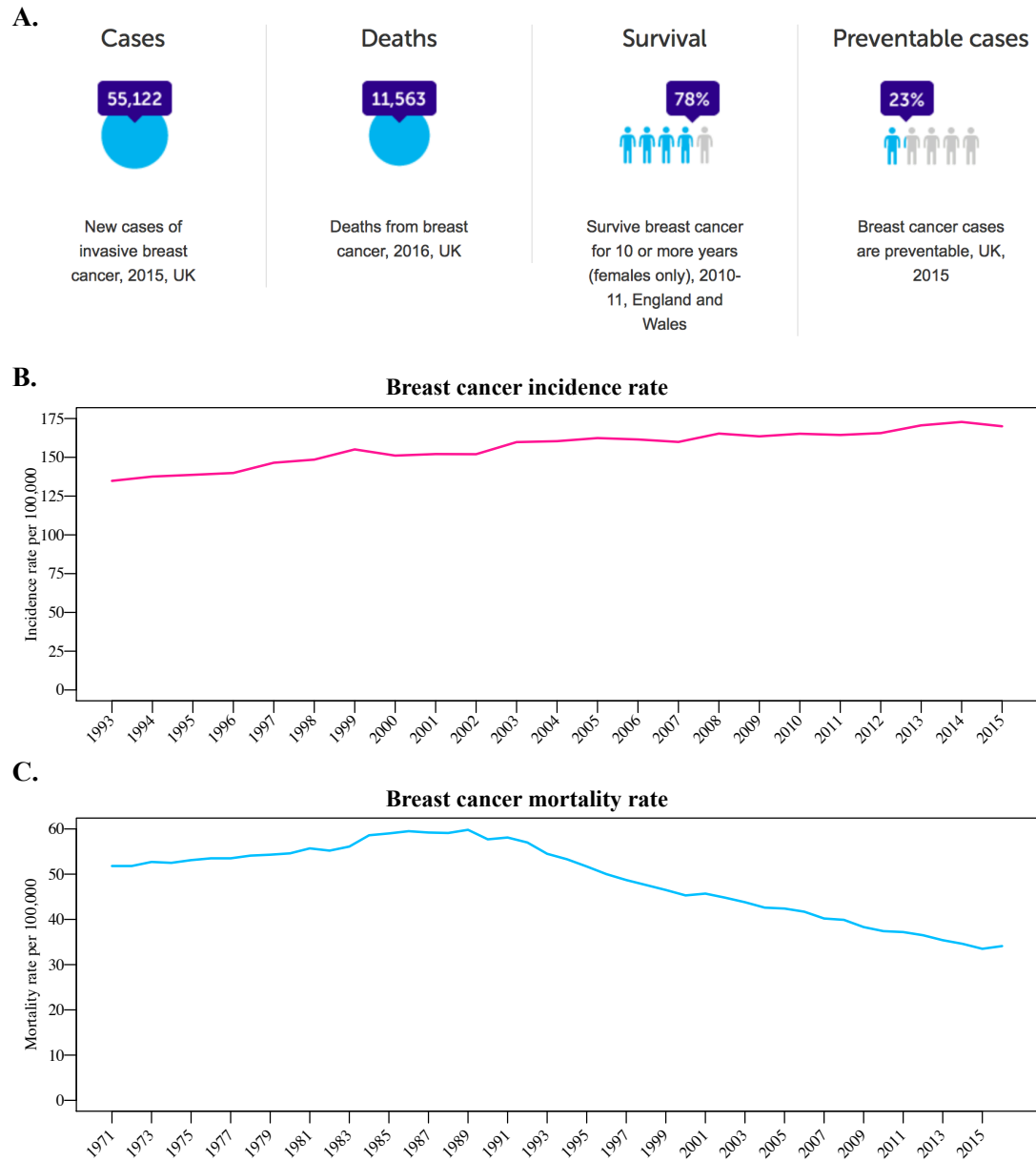


Figure 1.1: Breast cancer statistics

A. Latest statistics on number of breast cancer cases, deaths, survival, and preventable cases from Cancer Research UK [1]. **B.** The incidence rate has increased over the years, and is expected to keep increasing. **C.** The mortality rate has decreased after the introduction of population screening in the late 1980s. Data from Cancer Research UK [1].

Another way to classify breast cancers is into molecular subtypes based on their gene expression, where the status of the oestrogen receptor alpha (ER- α) in the tumour is the main defining feature. The majority, around 75% of all breast cancers, are ER-positive and thereby responsive to oestrogens [11]. The progesterone receptor (PR) is also regulated by oestrogens, and tumours that express both the ER and the PR (ER+/PR+) have a better prognosis than ER-/PR- due to the opportunity for oestrogen-targeted therapy [12]. The molecular subtypes luminal A and B both express the ER and the PR. Luminal A has a higher hormonal expression and better prognosis than luminal B, and the human epidermal growth receptor 2 (HER2) is positively expressed in 30% of all luminal B breast cancer. The HER2-enriched subtype has high expression of the HER2 and low expression of the ER and the PR. The basal-like subtype is the most heterogeneous subtype. A majority (80%) are triple negative breast cancer (TNBC, ER-/PR-/HER2-) and do not respond to endocrine treatment. TNBC cases are more frequent in women with mutations in the breast cancer susceptibility gene *BRCA1* in contrast to *BRCA2* mutations carriers that are more frequently ER-positive breast cancers. Other molecular subtypes described are Claudin-low and normal-like breast cancer [13, 14].

1.1.3 Breast cancer treatment

The histological and molecular characteristics of the breast tumours will determine the appropriate treatment. The current therapy for tumours that express the ER is different anti-oestrogens. Selective oestrogen receptor modulators (SERMs), including Tamoxifen and Raloxifene, bind to the ERs and inhibit their action on gene transcription by blocking co-activators and promoting co-repressors in the breast [15]. Tamoxifen has been widely used over the years and proven to significantly increase the survival of breast cancer patients. However, there are concerns of the effects on other tissues since it acts as an antagonist in some tissues but has oestrogenic activity in others, including the endometrium [15, 16]. Raloxifene, which was originally developed for osteoporosis, has significantly less endometrial oestrogenic activity, but is not as effective as Tamoxifen [15-17]. Another category of anti-oestrogens is selective oestrogen receptor-down-regulators (SERDs) including Fulvestrant, which bind to the ER and structurally block their action, and are described as ‘true’ anti-oestrogens [16].

Aromatase inhibitors (AIs) block the activity of the enzyme aromatase that is involved in the conversion of androgens to oestrogens, and thereby decrease the oestrogen levels. It has proven effective for postmenopausal women, however it is less effective in premenopausal women. This is due

to a reduction in the oestrogen negative feedback system following aromatase inhibition, which increases the synthesis of gonadotropins that stimulate both oestrogen and aromatase production in the ovaries, which increase the oestrogen levels [18, 19]. There are two types of AI, steroidal analogous of the androgen androstenedione that bind to aromatase irreversibly and thereby inactivates the enzyme (for example Anastrozole), and non-steroidal AI that bind reversibly to aromatase and inhibit its action (for example Exemestane) [19].

Unfortunately, resistance occurs in a large proportion of the women treated with anti-oestrogens. Either the tumours are not responsive to anti-oestrogens (intrinsic resistance) or the resistance is acquired after an initial response. The mechanisms for this are not fully understood, but include alterations and mutations to the *ESR1* gene that codes for ER- α , deregulations of co-activators of the ERs, overexpression of HER2, and ER activation in the absence of oestrogens. For example epigenetic alterations that lead to changes in the cholesterol synthesis have been observed, which can activate the ER [20-24].

ER-negative breast cancer will not respond to anti-oestrogens and are typically treated by surgery, radiation therapy, and chemotherapy, because of the lack of a more specific therapeutic target. The response to different chemotherapies differs between individuals; for example might carboplatin-treatment for TNBC be more effective in *BRCA1* or *BRCA2*-mutation carriers, but not in non-mutation-carriers [25]. Other treatment alternatives include Trastuzumab (Herceptin) for HER2 breast cancer, a recombinant monoclonal antibody that affects the HER2 receptors and has proven very effective for these tumours [26], and poly(ADP-ribose) polymerases (PARP)-inhibitors for ER-negative breast cancer in *BRCA1* mutation carriers, which are involved in DNA repair [27].

1.2. RISK AND PREVENTION

1.2.1 Genetic risk

Family history is an important contribution to a woman's breast cancer risk. Having one first-degree relative (mother, sister, daughter) with breast cancer is associated with an almost doubling of the risk (odds ratio, OR=1.80, 95% confidence interval, CI: 1.69 – 1.91) compared to a woman with no affected relative [28]. However, eight out of nine women that are diagnosed with breast cancer have no affected first-degree relative. Heritable genetic variants might explain up to a quarter of the breast cancer risk [29]. The two most common genes with hereditary genetic risk variants are *BRCA1* and *BRCA2*. Women carrying mutations in *BRCA1* or *BRCA2* have at the age of 70 a cumulative breast cancer risk of 65% and 45%, respectively [30]. The genes are involved in genomic stability and DNA repair mechanisms of double-strand breaks (DSB) [31]. However, mutations in the population are rare with prevalence of 1-3% in the general population [32, 33]. Other genes that moderately contribute to the breast cancer risk are different tumour suppressor genes and oncogenes controlling cell growth (*TP53*, *PTEN*, *STK11*, *CHD1*) and genes involved in DNA repair mechanisms (*CHEK2*, *ATM*, *PALB2*) [34, 35].

Several other genes with breast cancer risk single nucleotide polymorphisms (SNPs) have been identified in large genome-wide association studies (GWAS) [36, 37]. In fact, close to two hundred of risk-SNPs have been associated with breast cancer susceptibility [38]. The individual contribution from each SNP to breast cancer risk is small and not useful in risk prediction. However, when combined into polygenic risk scores (PRS) they have a much larger effect and can be used to stratify women into different risk groups [38-41].

1.2.2 Lifestyle risk factors

As already mentioned, it is estimated that 23% of all cases are potentially preventable by lifestyle changes. The most well studied lifestyle risk factors are increased body mass index (BMI), alcohol consumption, smoking, and physical inactivity (**Table 1.1**).

1.2.2.1 BMI in postmenopausal women

It is well established that increased BMI is associated with breast cancer risk in postmenopausal women [42-47]. For each 5 kg/m² increase in BMI the breast cancer risk is elevated by 12% [43].

However, the increase in risk with increasing BMI is stopped around a BMI of 28 kg/m² in postmenopausal women [46]. The reason for this plateau might be that obese postmenopausal women tend to also be obese at younger ages, and obese premenopausal women have a decreased breast cancer risk. When an obese woman enters the menopause, the breast cancer risk is only slightly increased [42, 45, 48]. Categorising postmenopausal women into normal-weight (< 25 kg/m²), overweight (25-30 kg/m²) and obese (> 30 kg/m²) groups, the risk was increased by 30% and 31% in overweight and obese women compared to normal-weight women, respectively [45]. The largest increase in risk might be for women putting on weight after the menopause [48, 49].

For postmenopausal women the main source of oestrogen is from the skin or adipose tissue, and the increased body fat will therefore lead to an increase in the oestrogen production. Higher BMI has been associated with increased aromatase activity and lower levels of the sex hormone binding globulin (SHBG), leading to a higher concentration of free bioavailable oestrogens [50, 51]. Higher BMI is mainly associated with ER-positive disease [52, 53], suggesting an effect in breast cancer risk from the increasing circulating oestrogen levels. The increased aromatase activity in obese women leads to higher levels of insulin, decreased production of SHBG in the liver, and increased circulating free bioavailable oestrogens [48, 51]. Furthermore, the increase in risk with increasing BMI seems to be limited to women who never have used hormone replacement therapy (HRT) [45, 47, 49, 52]. Therefore, it has been suggested that it is only in women whose oestrogen levels are low, that increased adiposity leads to an increase in breast cancer risk [48]. In fact, one study observed an inverse association between BMI and breast cancer risk among postmenopausal HRT users; obese women had a 34% reduction in the risk compared to normal-weight women [45]. Additionally, the breast cancer risk might also be mediated by inflammation in obese postmenopausal women [54].

1.2.2.2 Premenopausal BMI

An opposite, inverse association between BMI and breast cancer risk has been observed in premenopausal women [43, 44, 46, 55, 56]. Risk ratio (RR) of 0.92 per 5 kg/m² increase in BMI was observed for premenopausal women [43]. Obesity in premenopausal women (BMI > 31 kg/m²) was associated with almost half of the risk (RR=0.54) compared to women with BMI < 21 kg/m² [46]. A recent large study pooled data from 19 prospective cohorts and observed a stronger decrease in breast cancer risk among younger women, with significant hazard ratios (HRs) of 0.77 for women aged 18 to 24, and HR of 0.88 for women aged 45 to 54 [56].

The mechanisms between BMI and breast cancer risk in premenopausal women are not fully understood. Similar to postmenopausal women, the levels of SHBG are lower in obese premenopausal women. This leads to higher levels of free bioavailable oestrogens, but they are not as dramatically increased as in postmenopausal women [55]. The increased circulating oestrogen levels in young women might lead to early breast differentiation and irregular cycles with increased frequency of anovulation, which may reduce the breast cancer risk [45, 48, 56]. However, risk reductions has been associated with both ER-positive and ER-negative subtypes in premenopausal women [53, 56], as well as for TNBC [57], which implies that other mechanisms than oestrogens might be involved, such as insulin resistance and chronic inflammation.

1.2.2.3 Alcohol consumption

A positive linear relationship between alcohol consumption and breast cancer risk has been observed in both pre- and postmenopausal women [50, 58-61]. For each additional 10 g/day alcohol intake there is around 10% increase in breast cancer risk [62]. Even for light drinkers (defined as < 12.5 g/day or < 1 drink/day) a modest but significant increase in risk (RR=1.04) compared to non-drinkers was observed [60]. Largest effects were observed for heavy drinkers (> 3 drinks per day) with 40-50% increase in risk [59, 60, 63, 64].

Many different mechanisms have been proposed to explain the relationship between alcohol consumption and breast cancer [65]. Ethanol is metabolised to acetaldehyde, a weak carcinogen and mutagen, which can interact with DNA and proteins and form adducts. The ethanol metabolism can also produce reactive oxygen species (ROS) that are associated with DNA damage [65, 66]. Alcohol consumption also affects the oestrogen metabolism, and higher alcohol consumption is associated with increased circulating oestrogen levels in both pre- and postmenopausal women [50, 58, 60, 67]. The half-life of oestradiol is approximately three hours and thereby blood samples collected for oestrogen level measurements may miss the increased effect of alcohol consumption. However, it is hypothesised that the total cumulative lifetime oestrogen exposure will be influenced by alcohol consumption and lead to increased oestrogen levels and breast cancer risk in a more long-term perspective [58]. Ethanol stimulates the ER- α , cell proliferation, and aromatase expression in breast cells [68], down regulates *BRCA1* [69], and increases premenopausal mammographic density [58]. Furthermore, alcohol consumption decreases folate levels, which has an effect on the one-carbon metabolism and DNA methylation [65, 70]. However, it is not clear if folate levels affect breast cancer risk or if this is only

limited to postmenopausal women [66, 70, 71]. Alcohol consumption has been linked to both ER-positive and ER-negative subtypes, including TNBC [59, 71].

1.2.2.4 Smoking

Smoking is a well-known risk factor for many cancers, but it does not have a strong effect on breast cancer risk. The International Agency for Research on Cancer (IARC) has reported limited evidence for smoking and breast cancer [72]. However, several large studies have found significant associations between smoking and breast cancer risk. One study showed that the risk is increased by 16% in current smokers, 14% in former smokers and 10% in current passive smokers, compared to women who never smoked and were not exposed to passive smoking [73]. Similar results have been observed in other studies [74-77], and the risk might be elevated up to 20 years after smoking cessation [76]. Longer duration or intensity smoking may increase the risk further [76, 78]. Also, the timing of starting to smoke seems to be important with particularly increased risk in women who smoked between menarche and first full-term pregnancy (FFTP) [73, 78]. In contrast to this, there are also studies that have found no association with breast cancer risk for neither smoking status, duration or intensity [79].

The mechanisms between smoking and breast cancer risk, if there is a true association, might be from different carcinogens that cause DNA damage or alterations in the oestrogen metabolism. In premenopausal women, lower oestrogen levels in the luteal phase have been observed in current smokers [80], but also no effect on oestrogen levels has been observed [55]. Changes in the oestrogen metabolism have been observed in smokers, with increased levels of DNA damaging oestrogen-metabolites [80]. Postmenopausal heavy smokers may have increased oestrogen levels [50], but controversially an anti-oestrogenic effect has been observed in smoking postmenopausal women [81]. The mechanisms are not fully understood, but smoking women reach the menopause earlier, which is associated with a decrease in the breast cancer risk [81, 82]. It might also be that the hepatic degradation of oestrogens is enhanced in postmenopausal smokers [81]. Some studies have even observed a reduced breast cancer risk in postmenopausal smoking women [73].

Table 1.1. Hormonal risk factors in breast cancer

Summary of the most recent and largest publications on different lifestyle and reproductive factors and the risk for breast cancer

Risk factor	Risk increase/decrease
Postmenopausal BMI	RR=1.12 per 5 kg/m ² increase [43]
Postmenopausal obesity	RR=1.31 (> 30 kg/m ² vs < 25 kg/m ²) [46]
Premenopausal BMI	RR=0.92 per 5 kg/m ² increase [43]
Premenopausal obesity	RR=0.54 (> 31 kg/m ² vs < 21 kg/m ²) [46]
Alcohol consumption	OR=1.10 for each additional 10g/day intake [62]
Current smoking	HR=1.16 compared with never smokers not exposed to passive smoking [73]
Former smokers	HR=1.14 compared with never smokers not exposed to passive smoking [73]
Passive smoking	HR=1.10 compared with never smokers not exposed to passive smoking [73]
Physical inactivity	RR=0.88 for highest compared with lowest levels [85]
Age at menarche	RR=1.05 for each year younger [82]
Age at menopause	RR=1.03 for each year older [82]
Nulliparity	RR=1.30 compared to parous women [98]
Number of pregnancies	RR=0.93 for each additional birth [100]
Age at FFTP	RR=1.03 for each year older [100]
Ever breastfeeding	RR=0.61 compared to parous women who never breastfed [117]
Breastfeeding duration	RR=0.96 for each additional 12 months breastfeeding [100]
HRT duration	RR=1.02 for each additional year* [95]
ERT	RR=1.40 compared to never users* [121]
CRT	RR=2.00 compared to never users* [121]
OC	OR=1.08 compared to never users* [130]

* Increase in risk was only observed for current use and up to 5 - 10 years after cessation

The association between smoking and breast cancer risk might be confounded by alcohol consumption. Smoking and alcohol consumption are highly correlated, and one study found no effect on risk when restricting the analysis to non-drinkers [61]. However, in another study the association was restricted to non-drinkers [73], and no potential confounding effect from alcohol consumption has been found in other studies [50, 75, 77]. Furthermore, smokers tend to be leaner which is associated with a decreased risk in postmenopausal women [50, 73, 81]. Further studies are needed to conclude whether smoking has a causal effect on breast cancer risk.

1.2.2.5 Physical inactivity

Physically active women have a significantly reduced risk for breast cancer compared to less active women [83-87]. Studies indicate that at least 150 minutes of vigorous physical activity per week is needed for a decrease of 9% in lifetime risk for breast cancer [85]. Both household, recreational, and total physical activity are associated with risk reductions [84, 87, 88]. Some studies have reported less strong associations for physical activity and breast cancer in premenopausal women [84, 86], but a recent large meta-analysis observed no differences in menopausal status [85]. Furthermore, use of HRT might be confounding the associations in postmenopausal women and was observed to almost cancel out the effects on risk-reduction of physical activity in HRT users [85]. A somewhat stronger association has been observed for leaner women, but this could be confounded by BMI, with smaller risk reductions in postmenopausal heavier women who already are at elevated risk [84]. Independent effects for BMI and physical activity have also been observed [86].

The biological mechanisms between physical activity and breast cancer risk are not fully understood, but might be due to changes in hormone levels [88]. Risk reductions have been observed for both ER+/PR+ and ER-/PR- breast cancer, but it has been suggested that the PR might play a more important role in the breast cancer development [85, 87]. Lower circulating oestrogen levels have been observed in the luteal phase in highly physically active women [88], but also no effect on oestrogen levels has been observed [86]. Very high levels of physical activity can affect the timing of menarche in young women, and affect the menstrual cycle in premenopausal women by making it longer and increasing the number of anovulatory cycles [87, 88]. This will result in a reduction in the total cumulative lifetime exposure to high levels of oestrogen, and thereby decrease the breast cancer risk. There are many other factors that might be influenced by physical activity, such as weight reduction,

changes in the metabolism of insulin and growth factors, changes in immune function, and decreased systemic inflammation [84, 86].

1.2.3 Reproductive risk factors

Other reproductive factors have been associated with breast cancer risk (**Table 1.1**), which is thought to be mainly through alterations of oestrogen exposure. Recent trends in reproductive factors are thought to be a reason to explain the increase in breast cancer incidences over the last years, including a younger age at menarche, fewer pregnancies, older age at first pregnancy, and shorter breastfeeding duration.

1.2.3.1 Age at menarche

Breast cancer risk is modified in young ages through timing of different pubertal stages [89]. The menarche marks the start of menstrual cycles, confirms ovarian and endocrine reproductive activity, and is an approximate start of the exposure to high oestrogen levels [82, 89, 90]. With changes in the hormonal environment, the number of undifferentiated breast cells increases, which makes the breast tissue more susceptible to carcinogens [89]. A younger age at menarche is associated with a higher breast cancer risk compared to later age at menarche, and will increase the time when a woman is exposed to the high oestrogen levels through the menstrual cycles [82, 89-91]. Each year younger age at menarche is associated with 5% increase in breast cancer risk [82]. In recent years the age at menarche has decreased in European countries, which is thought to be a result from lifestyle changes in food intake, BMI and physical activity [91, 92].

1.2.3.2 Age at menopause

A woman is exposed to high levels of circulating oestrogens produced in the ovaries between menarche and menopause, which mark the reproductive time for a woman. The number of menstrual cycles is an approximation of the reproductive time and has been positively associated with risk of breast cancer [93, 94]. The total oestrogen exposure is further affected by time spent pregnant, duration of breastfeeding, OC and HRT use, and length and regularity of menstrual cycles [90]. The menopause marks the stop of exposure to high circulating levels of oestrogens: the oestrogen production in the ovaries is stopped and the oestrogen levels decrease rapidly [82]. An older age at menopause is associated with increased breast cancer risk [82, 95, 96] and each additional year before menopause is

associated with around 3% increase in breast cancer risk [82]. The same risk reduction is seen for women who have undergone bilateral oophorectomy [97]. After menopause, the breast cancer risk decreases dramatically. Premenopausal women aged 45 to 54 years have a 43% increased breast cancer risk compared to postmenopausal women of the same age [82], which probably has to do with the decrease in oestrogen levels [82, 95]. On the other hand, the risk for breast cancer in postmenopausal women can be increased by HRT use and higher BMI [95]. Whether early menarche and late menopause are independent breast cancer risk factors is not clear. It is hypothesised that a longer reproductive time increases the breast cancer risk but this has been poorly studied. The two life events are likely to be correlated and it has been shown that oestrogen levels in postmenopausal women are greater for women with an early age at menarche [82].

1.2.3.3 Parity and age at first full-term pregnancy (FFTP)

Nulliparous women have around 30% increased lifetime risk of developing breast cancer compared to parous women [91, 96, 98, 99]. Each additional birth is associated with 7% decrease in breast cancer risk, independent from breastfeeding [100]. However, each year older age at FFTP is associated with 3% increase in the risk [100], and women with an age at FFTP over 35 compared with below 20 have a 40% increased risk for breast cancer [91, 96, 98]. This increase in risk might be greater for premenopausal breast cancer [91]. Women with an age at FFTP of 35 or older might have a greater risk compared with nulliparous women, however this was observed in old data and might need to be updated [101]. A more recent study found that the risk is increased shortly after FFTP, and that the risk increase is greater and stay elevated for a longer time for women with an older age at FFTP. The risk can be elevated up to 30 years after giving birth for women with an age at FFTP of 30 or over [102].

The protective mechanisms from pregnancy are thought to be from the hormonal changes during and after pregnancy as well as a result of lower number of undifferentiated breast cells. During pregnancy, the breast cells undergo a maturation process, including developing of lobules, branching, and milk production [103]. The morphological changes of the breast and the exposure to high levels of hormones make the breast cells less susceptible to carcinogens [103, 104]. It has been observed that the hormonal changes might lead to an increased expression of the tumour suppressor gene *p53* in the breast and increased apoptotic activity, which might be involved in the protective effect [104]. The biological protective effects are only seen for pregnancies lasting longer than 33 weeks, suggesting that important changes are introduced late in the pregnancy [105]. Parous women have a decreased risk for

developing ER+/PR+ tumours of luminal subtype [106-108], but the risk might also be decreased for ER-negative and TNBC [109]. Parous premenopausal women have decreased levels of oestrogens throughout the menstrual cycle compared to nulliparous women [110], and parous postmenopausal women have decreased oestrogen levels and increased levels of SHBG [111]. This may be a result of better oestrogen metabolism and more regular cycles after being exposed to high levels of oestrogen during pregnancy [110, 111].

On the other hand, the short-term effect from being pregnant is an increased risk for breast cancer. The cancers developed during or shortly after giving birth are referred to as pregnancy-associated breast cancer (PABC) [112]. This is thought to be due to the exposure of high oestrogen and progesterone levels during pregnancy [113], and is especially seen in women with an older age at first pregnancy [91, 112]. The younger age at the FFTP, the earlier the breast cells undergo the maturation process, and the longer a woman can benefit from the long-term protective effects [91]. However, the number of terminal duct lobular units (TDLU, the structures in the breast that produce milk) is increased in parous women compared to nulliparous women, especially after giving birth and the 10 following years, which has been associated with breast cancer risk [114]. The PABC cases have poor survival and a high risk of relapse, which is probably due to a delay in diagnosis since the breast density increases during pregnancy and breastfeeding, which makes the cancer hard to detect [112]. In summary, pregnancy is therefore associated with a short-term increase in risk, but a long-term decrease in risk for breast cancer [105].

1.2.3.4 Breastfeeding

The association between breastfeeding and breast cancer risk have been inconsistent in earlier studies, but recent large studies have been able to establish a moderate but significant decrease in risk with longer duration breastfeeding. Parous women who have been breastfeeding compared with parous women who never breastfed have a significant decrease in breast cancer risk, which has been reported in several large meta-analyses [100, 115-117] with risk estimates ranging from OR=0.86 (95% CI: 0.82 – 0.90 [115]) to RR=0.61 (95% CI: 0.44 – 0.85 [117]). Every additional 12 months of breastfeeding decreases the risk by 4.3% [100]. A few studies report only a protective effect in premenopausal breast cancer and it is thought that the protective effect in postmenopausal women is masked by the already decreased risk for breast cancer [115, 118].

Several potential biological protective mechanisms of breastfeeding have been proposed. The protective effect may be mediated by the hormonal changes following both from having been pregnant and breastfeeding [90, 116]. Breastfeeding is closely related to parity and women who have more children spend more time breastfeeding [100, 116]. Women breastfeeding for a longer time have reduced lifetime numbers of menstrual cycles, and the protective effect might be mediated through shorter total lifetime oestrogen exposure [90, 115-117, 119]. However, breastfeeding has been associated with a decreased risk in both ER+/PR+ and ER-/PR- breast cancer, suggesting a potentially hormone-independent protective effect [107]. The differentiation process of breast cells during and after pregnancy for milk production makes them less susceptible to carcinogens, as described earlier, and could lead to decreased breast cancer risk [116, 117].

1.2.3.5 *Hormone Replacement Therapy (HRT)*

When a woman enters the menopause there is a decrease in oestrogen and progesterone levels. HRT is often prescribed to replace the hormonal production in the ovaries and to relieve the symptoms that can occur, such as hot flashes [120]. However, HRT use is associated with an increase in breast cancer risk and it was estimated that 20,000 extra breast cancer cases were related to HRT use in the UK in the twentieth century [121]. For each additional year of HRT use, the breast cancer risk is increased by 2%, which is similar to one year delay in the menopause [95]. The risk increase is further affected by the hormonal formulations in the HRT. Fifty years ago oestrogen-only HRT (ERT) was mostly prescribed, but has been replaced with combined oestrogen and progestins (CRT) because of the observed increased risk of endometrial cancer from ERT [120]. However, in several studies a greater effect on breast cancer risk has been observed for CRT compared to ERT [97, 121-124]. In the Million Women Study in the UK, current use of ERT was associated with 40% increase in risk, but a 100% increase in risk for CRT [121]. Less than one year of CRT increased the risk with 45% compared to never users [121]. After the Women's Health Initiative (WHI) clinical trial of HRT that showed an increased risk for breast cancer in CRT users [125], many women stopped taking HRT in the US, which was suggested to be a reason for the reduction in breast cancer incidence between 2002 and 2003 [120]. However, the study was criticised: almost half of the women did not follow the hormone therapy through, the number of cases was small, and some women had been taking ERT before the start of the study.

The risk increase seems to follow a complex pattern: in the WHI, the risk was increased during the intervention phase and decreased quickly after cessation [126]. This is similar to other studies that have found a significant increase in breast cancer risk in current users but not 2-5 years after cessation [95, 121]. The risk is further increased with longer duration of HRT [95, 121, 124]. The increased risk in HRT users seems to be limited to ER+/PR+ subtype [124, 127]. This, in combination with the greater risk increase in CRT compared to ERT, suggests that the PR is the driving mediator of the breast cancer caused by HRT [124].

Different confounders can influence the association between HRT and breast cancer risk. Importantly, a larger number of ERT users are thought to be a reason for non-significant results between HRT and breast cancer risk in older studies [95, 120]. Different doses as well as different oestrogen and progestin formulations may also have different effect on breast cancer risk, but the results are inconsistent [95, 121, 123]. Also the timing of menopause and start of HRT may have an effect on the breast cancer risk, where women starting to use HRT a short time after the menopause and for a long duration seem to have a particularly increased breast cancer risk [128]. A lower risk increase has been observed in women with higher BMI, suggesting that BMI can overtake the effect HRT has on breast cancer risk [95, 127]. Furthermore, breast cancer diagnosed in HRT users seems to be less clinically advanced with often more low-grade breast tumours that are less likely to spread [95, 120]. Therefore, while HRT increases the risk of cancer diagnosis, it may not increase the risk of breast cancer related mortality.

1.2.3.6 Oral Contraceptives (OC)

Hormonal contraceptives have been widely used since 1957 and contribute to the primary prevention for ovarian cancer with a risk reduction of 50% or more in long-term users [129]. The concern for its effect on breast cancer risk was highlighted early, but the long-term effects have not been possible to study until recently. Ever use of OC increases the risk of breast cancer by 8% compared to never users [130], with potential greater effect for breast cancer diagnosed before menopause [131]. However, similar to HRT, the risk seems to be increased for current users by around 20%, and then decrease back to normal risk-levels within 10 years after last use [132, 133]. Comparably, more than 5 years use of OC has been associated with breast cancer mortality up to 10 years since last use [134]. OC use in young ages does not seem to affect the breast cancer risk later in life [135], and another study only found a significant association between recent (< 1 year) OC use and breast cancer risk [136].

Furthermore, the duration of OC use might increase the risk further [133, 137, 138], with 14% increase in risk per 10 year use of OC [138]. However, the risk increase might be a consequence of increase in age, which will itself increase the risk for developing breast cancer [139].

The breast cancers associated with OC use seem to be limited to the ER-positive subtype [136]. This is in concordance with the increase in risk among current users only, which suggests no genotoxic effect from OCs [132]. Furthermore, OC use does not seem to affect the risk in high-risk women with family history of *BRCA1* or *BRCA2* mutations [140]. Different formulations of OCs in oestrogen-only, progestin-only, or combined oestrogens and progestins might have different effects on the risk. The breast cancer risk might not be increased in users of low-dose oestrogen-only OCs [136]. Other studies did not find any particular variation between different hormonal combinations or formulations [132, 133]. The effect from OCs on oestrogen exposure is complex: OC use can cause anovulation, which might be protective for breast cancer, but also the exogenous exposure to oestrogens and progestins can promote breast cancer development [137]. The breast cancer diagnosed from OC use may be less clinically advanced, which might be a combination from early diagnosis or a biological effect from OC use [132]. Nevertheless, several studies highlight that the benefits from OC use overtakes the small increase in risk that is seen in mainly premenopausal women where the breast cancer incidence rate is low to start with [139].

1.2.4 Breast cancer risk models

Different breast cancer risk models have been developed that combine different risk factors and family history to estimate a woman's lifetime risk of developing breast cancer or in a 5 or 10-year period [141]. The Gail model was developed in 1989 and estimates the risk of *in situ* breast cancer for a woman given her age, age at menarche, age at first live birth, information of previous breast biopsies, and number of first-degree relatives with breast cancer [142]. The Gail model was later modified to estimate the risk for invasive breast cancer only and is often referred to as Gail model 2 [143]. The Gail model 2 might predict the overall breast cancer risk better than model 1, but both models seem to underestimate the risk for women in the lowest risk groups, and overestimate the risk for women in the highest risk groups [143]. The Tyrer-Cuzick model (IBIS Breast Cancer Risk Evaluation Tool [144]) also incorporates second-degree family history, age of onset of cancer in the family, use of HRT, and information of *BRCA1* and *BRCA2* mutations. The risk model estimates both the risk for developing

breast cancer and a woman's risk of carrying any high-risk mutations [145]. Other models, such as the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA, [146]), predicts risk for women with family history of breast cancer and computes *BRCA1* and *BRCA2* mutation carrier probabilities, but was recently updated to include more information [147].

Even though the additional information incorporated in the Tyrer-Cuzick model improved the risk prediction in comparison to other risk models [148], it is not well calibrated across different populations. Attempts to improve the risk models have been implemented by adding information of breast density and polygenic risk score (PRS) [141]. Breast density is positively associated with breast cancer risk; women with a breast density of 75% or more have almost five times higher risk compared to women with less than 5% [149]. In one UK study, the breast density itself was a stronger risk factor than the Tyrer-Cuzick or Gail model alone. Adding information of breast density to the risk models improved the risk prediction slightly [150]. Similar results have been observed in other studies where the additional information of breast density or PRS improved the models significantly but modest [141]. A recent study including PRS of 88 risk-SNPs observed similar prediction accuracy for the PRS as for the Tyrer-Cuzick model, but it was substantially improvement in the combined model [38]. The recently updated BOADICEA model now includes *BRCA1*, *BRCA2*, *PALB2*, *CHEK2*, and *ATM* in a PRS (with a total of 313 risk-SNPs), as well as mammographic density and different reproductive and lifestyle risk factors (age at menarche, age at menopause, parity, age at FFTP, OC use, HRT use, height, BMI and alcohol intake). It has the potential to improve risk stratification and is well calibrated in UK and European populations, but needs further validation [147].

The clinical aim of risk models is to be able to categorise women into high- and low-risk groups, with few women in the middle-risk groups. The addition of PRS to the Tyrer-Cuzick model led to more women being categorised into high or low risk groups [151]. Improved risk stratifications have been observed when combining risk factors, family history, breast density, and PRS [40]. The continuous discovery of more risk-SNPs to include in the PRS will probably improve risk assessment further. Other potential improvements to the risk models are to include information on hormonal exposures and lifestyle risk factors. However, including hormonal measurements is tricky due to the variation in hormone-levels measurements [152, 153]. A model estimating a woman's total lifetime oestrogen exposure, similar to the Pike model, could be relevant. The Pike model is a model for 'breast

tissue aging' including information of a woman's age at menarche, age at FFTP, and age at menopause [154].

Also, gene expression profiles, proteomics, and epigenetic risk markers in blood have the potential to be included in risk models and improve risk stratification [5, 141]. For now, the risk models have good prediction accuracy for a population, but not at the individual level [5, 141, 155]. Further understanding of the biological mechanisms leading to breast cancer development is needed to improve risk assessment.

1.2.5 Prevention

As already mentioned, it has been estimated that 23% of all breast cancer cases in the UK are preventable by lifestyle changes. This includes changes in modifiable risk factors, including the above-mentioned lifestyle factors. However, it is believed that incorporating other preventable approaches for women at high risk, such as chemoprevention, will reduce the number of cases further.

1.2.5.1 Chemoprevention

For prevention of ER-positive breast cancer, the blocking or inhibition of oestrogen activity by SERMs, SERDs or AIs has been successful. At least five years of Tamoxifen can reduce the breast cancer risk by 33% overall (including DCIS), with the biggest effect for ER-positive disease with a risk reduction of 44% [156]. The risk is reduced for a long period after treatment for both ER-positive disease and DCIS [157]. The risk could be further reduced for women at high risk: one study with women taking Tamoxifen for 5 years with a 5-year risk of 1.66% or more (estimated by the Gail model 2) the RR was 0.57 compared to the placebo group [158]. Tamoxifen shows better risk-reductions compared to Raloxifene, but it comes with more side effects [156]. Lasofoxifene, another SERM, is a promising candidate since it not only reduces the risk for breast cancer but also decreases the risk for stroke and cardiac events, however larger studies with longer follow-up time are needed [156]. Even greater risk reductions have been observed for different AIs: Exemestane has shown to reduce the risk by 65% and Anastrozole by 53% [141, 159, 160]. Compared to SERMs, they are not associated with increased risk in thromboembolic events or endometrial cancer, but are associated with moderate bone and muscle pain, reduced bone density, and hot flashes. However, they can only be used for postmenopausal women [141, 159, 160].

When should chemoprevention be recommended? For which women will the benefits in breast cancer risk reduction overtake the risks of the side effects? To answer these questions better risk assessment models are needed to clearly identify women at high risk of developing breast cancer. In the USA, women with a 5-year projected absolute risk of 1.66% or more, estimated by the Gail model 2, should discuss the use of chemoprevention [141]. In the UK, chemoprevention should be considered for women at moderate risk (lifetime risk of 17-30%) and offered to women at high risk ($\geq 30\%$) [40, 161]. Further improvements are needed to understand which women are at a high risk and what the best therapeutic or other preventative approaches would be.

Importantly, chemoprevention will only target hormone-dependent cancers. It might be desirable to decrease the risk for hormone receptor negative cancers as well, especially since this cancer is more aggressive with poor prognosis, and often affects younger women. Hormone-independent approaches to reduce breast cancer risk might include statins to block cholesterol synthesis, the type-2 diabetes drug Metformin that has showed to decrease the breast cancer risk, PARP-inhibitors involved in DNA repair (especially for women with BRCA mutations), nonsteroidal anti-inflammatory drugs including aspirin, and more. However, the mechanisms are yet not fully understood for these alternatives [162, 163]. Other approaches are changes in lifestyle factors to decrease risk for breast cancer overall.

1.2.5.2 *Lifestyle changes*

At least three lifestyle changes are widely agreed to be able to prevent a large number of breast cancer cases. These are: maintaining a healthy weight, reduction in alcohol drinking, and increase in physical activity [164]. There is substantial evidence that increased alcohol consumption and postmenopausal weight gain is associated with increased breast cancer risk, as outlined above. It has been estimated that 8% of all breast cancer cases are caused by being overweight or obese, and another 8% from alcohol drinking [1]. Also physical activity reduces the risk, however, it could be confounded by other factors such as weight loss. Of the already mentioned reproductive factors, breastfeeding should be encouraged, but other recommendations might not be appropriate due to different social, economical, and health reasons [141].

The lifestyle factors are highly correlated: individuals that tend to exercise regularly more often maintain a healthy weight, consume a healthy diet with moderate alcohol intake, and tend to not

smoke [165]. All of these lifestyle factors will not only decrease the risk for breast cancer, but many other cancers and diseases including cardiovascular disease (CVD) and diabetes. It is therefore generally advised to follow the national guidelines. Several studies have reported breast cancer risk reduction for both ER-positive and ER-negative disease in women who adhere to the health guidelines, including women with family history [165]. In the UK, the guidelines for alcohol consumption are to not drink more than 14 units of alcohol per week (one unit corresponds to 8 gram pure alcohol), to spread it out over at least three days, and to keep several days alcohol-free [166]. On the contrary, studies have found increased cancer mortality in non-drinkers compared to light-drinkers, suggesting a J-shaped relationship between alcohol consumption and cancer mortality [167]. However, this might be confounded by several factors: individuals with poorer health tend to drink less, and high socio-economical status is associated with light drinking [167]. Furthermore, for weight maintenance it is recommended to keep a BMI between the normal-weight range (18.5 – 25.0), waist circumference less than 80 cm, and to be physically active for at least 150 minutes per week [166]. There is clear evidence that weight maintenance or weight loss is associated with a reduction in postmenopausal breast cancer risk [168, 169].

There is some evidence suggesting that dietary patterns can have an effect on breast cancer risk. For example, increased fibre intake [170] and reduced intake of red and processed meat might decrease the risk for breast cancer [171]. A Mediterranean diet has been associated with decreased risk for invasive breast cancer [172], which might be more relevant to ER-negative disease [173]. Overall, dietary patterns with high grain, fruit, vegetables, and fish intake might reduce the breast cancer risk, but further studies are needed to conclude this [174]. In summary, there is a complex interaction between lifestyle, genetic, reproductive, and environmental exposures and breast cancer risk. To get more individuals to be adherent to the public health guidelines, and to reduce the breast cancer incidences, more evidence is needed that sustainable lifestyle changes reduce breast cancer risk on the individual level.

1.3. EPIGENETICS

Epigenetics is heritable changes in gene function that cannot be explained by genetics, and is responsible for the stable inactivation of the X chromosome in females and cell differentiation, giving rise to different cell types with the same genetic sequences [175-178]. The term ‘epigenetics’ is not completely defined [176], but includes changes in chromatin structure, histone modifications, and methylation of the DNA [175]. Inside the cell nucleus the DNA is wrapped around nucleosomes, which are made up of eight histones, and is further organised by other high-dimensional structures. The accessibility for different transcription factors and other proteins to bind to the DNA will affect the gene expression in the cell [177, 178]. Condensed, highly packed DNA (heterochromatin) will silence gene expression, and open chromatin will be accessible to the transcriptional machinery. The chromatin is further altered by different histone modifications, including acetylation, methylation, and phosphorylation [178, 179]. For example, trimethylation of histone H3 at lysine 4 (H3K4me3) is associated with active promoters, and trimethylation of histone H3 at lysine 9 (H3K9me3) is associated with repressed promoters [180].

In this project the focus has mainly been on DNA methylation. The role of DNA methylation was proposed already in 1975 and is the most well characterised epigenetic mechanism [175, 177]. DNA methylation is the addition of a methyl-group to cytosine often positioned adjacent to a guanine (CpG site). Long DNA sequences with high-density of CpG sites are termed CpG islands. Earlier, the main research focus was on DNA methylation in promoter CpG islands that lead to silencing of that gene. However, approximately half of the promoters do not contain CpG islands and the focus has now shifted to an important role of DNA methylation outside CpG islands and gene bodies [178]. DNA methylation is a reversible mechanism affected by environmental exposures, and altered by different methyltransferases (DNMTs) [175, 177]. The “maintenance” methyltransferase DNMT1 is responsible for the methylation of hemi-methylated (methylation at a CpG site on only one of the DNA strands) during DNA replication, and is important for the inheritance of DNA methylation patterns. DNMT3A and DNMT3B are known as *de novo* methyltransferases, even if DNMT1 also has been associated with *de novo* methylation, and are thought to be important in early development. For example DNMT3B might be responsible for the X chromosome inactivation [181]. However, the mechanisms of methylation and demethylation are yet not fully understood.

DNA methylation can be quantified using various different assays. The Illumina human methylation array measures DNA methylation at pre-designed probes, and has been improved from covering over 27,000 (the HM27K array), over 450,000 (the HM450K array) and now over 850,000 (the HM850K or EPIC array) probes across the genome [182]. It is designed to cover a mixture of CpG islands, promoters, enhancers and other regulatory regions, and both intra- and intergenic CpG sites. Bisulphite pyrosequencing is often used as the golden standard in DNA methylation studies to investigate candidate genes or short DNA sequences [183]. The DNA is treated with bisulphite before sequencing, which converts unmethylated cytosines to uracils and later on thymines, which lets the researcher differ between methylated and unmethylated CpG sites. To cover a larger number of CpG sites, whole-genome bisulphite sequencing (WGBS) can be used, however, for now, the sequencing cost is relatively high to achieve deep sequencing coverage for a large number of samples [184]. Instead, the Illumina human methylation arrays are often used, which is more cost-effective for large studies. The reduced representation bisulphite sequencing (RRBS) is an alternative, but will have a bias towards CpG islands (and thereby promoter regions) due to the use of restriction enzymes targeting 5'CCGG3' sites, and will only cover approximately 10% of the genome [184]. The LUMInometric Methylation Assay (LUMA) does not require bisulphite treatment by the use of two restriction enzymes; one methylation-sensitive that will only cut unmethylated DNA, and one methylation-insensitive, followed by polymerase extension and a pyrosequencing procedure [185]. However, again, this technique is biased due to the choice of restriction enzymes. Various other assays have been developed, including different PCR-based methods for candidate genes-approach, and the [³H]-methyl acceptance assay, which uses radiolabelled S-adenosylmethionine (SAM), an important methyl-group donor, that methylates all unmethylated CpG sites, resulting in a signal inversely related to DNA methylation levels [186].

1.3.1 DNA methylation and breast cancer

Changes in DNA methylation have been recognised in early cancer development, including breast cancer. Generally, a decrease in the genome-wide DNA methylation and hypermethylation at promoter CpG islands have been observed [178, 179, 187, 188]. This leads to aberrant gene expression, inactivation of tumour-suppressor genes, activation of oncogenes and disruption of DNA damage mechanisms. The genome-wide reduction in DNA methylation is also thought to lead to genomic

instability. The now “unprotected” DNA can result in a larger proportion of translocations, deletions, and chromosomal rearrangements [179]. Hypomethylation occurs predominantly outside of genes, which makes it difficult to link to a transcriptional gene function, but is probably enriched in important regulatory regions such as enhancers. One of the genes that are hypomethylated in breast cancer is the insulin-like growth factor 2 (IGF2).

Several hundreds of genes have been found hypermethylated and thereby down regulated in breast cancer [188]. This includes methylation of tumour suppressor genes, such as *p16* that is involved in the normal growth of mammary epithelial cells. The silencing of *p16* by DNA methylation is thought to be an important step in immortalisation in early breast cancer development [187, 189]. Hypermethylation has also been observed in the gene *GSTP1* that is involved in DNA repair [189], the growth suppressor gene *HIN1*, *HOXA5* that is involved in the p53 synthesis, *THBS1* that inhibits angiogenesis, and *CDH1* coding for E-cadherin that suppress cell invasion and metastasis [187]. Methylation of *BRCA1* occurs to some extent [189], but is more often found in rare subtypes such as basal-like hormone receptor negative breast cancer [188, 190].

Breast cancer cases with *BRCA2* mutations have overall higher DNA methylation levels compared to *BRCA1* mutations [191]. Also, different DNA methylation profiles have been observed for the molecular subtypes, especially between Luminal A, Luminal B and basal-like breast cancer [191]. The basal-like breast cancer has overall lowest DNA methylation levels and Luminal B the highest. Differentially methylated CpG sites have been observed between ER+/PR+ and ER-/PR- breast cancer [192]. A larger number of hypermethylated loci were observed for ER-positive breast cancer, but more promoter-methylation for ER-negative [193]. Methylation of the *ESR1* gene coding for the ER- α might be a mechanism that leads to the loss of ER expression in ER-negative breast cancer [187, 194], however, results are inconsistent and there is no clear link between ER methylation and ER status [188].

1.3.2 Blood DNA methylation and breast cancer

It is clear that DNA methylation is important in breast cancer development, however, the mechanisms are not fully understood. The fact that changes in DNA methylation have been observed in early cancer development, and that DNA methylation is altered from exogenous and endogenous exposures associated with cancer risk, has led to the hypothesis that it can be used as risk biomarker for cancer

[195]. Even though methylation profiles are tissue-specific there is evidence that methylation changes associated with cancer can be detectable in blood DNA [186-188, 195-197]. Both loci-specific and genome-wide DNA methylation profiles in blood have been associated with breast cancer (this is summarised more in detail in the introduction to **Chapter 2**). Different methylation profiles have been observed in ER-positive and ER-negative breast cancer, for example unmethylated promoter and increase in intragenic methylation in *ESR1* in ER-positive disease [193, 198]. The *BRCAl* promoter is often hypermethylated in *BRCAl*-mutated breast cancer cases [199, 200]. In addition to this, hypermethylation of *ATM* [201, 202] and intergenic CpG island hypomethylation of 2q37.1 have been associated with breast cancer [203]. Genome-wide hypomethylation in blood DNA has been observed in breast cancer cases; however, results are inconsistent [195, 197, 204-206]. Furthermore, an increase in DNA methylation at promoters and CpG islands, and decrease in gene bodies and intragenic regions have been associated with breast cancer risk [197, 204, 206].

Specific DNA methylation profiles for breast cancer could be promising as early biomarkers for breast cancer and might improve risk assessment. DNA methylation is a relatively stable marker and affected by different risk exposures [195, 197, 203, 207]. However, so far no strong methylation signatures have been identified for breast cancer risk prediction [208]. It might seem questionable to use DNA methylation in blood to predict breast cancer risk as it is not the target tissue, however, DNA methylation might represent a signature of long-term exposures [207], and blood sampling is an accessible and less invasive method for risk screening purposes. Since DNA methylation changes can last longer than the lifetime of most of the mature blood cells, an “epigenetic memory” of past exposures that are embedded in stem cells has been hypothesised [209]. Exposures will independently act on more than one cell type, and similar DNA methylation patterns have been identified across multiple tissues. For example, DNA methylation alterations in blood found to be associated with smoking and BMI have replicated in lung [203, 210, 211] and adipose tissue [212, 213]. In this study the use of blood DNA methylation as risk biomarker for breast cancer will be further explored.

1.3.3 Breast cancer risk factors and DNA methylation in blood

Numerous endogenous and exogenous exposures have been found to affect DNA methylation in blood. This includes different breast cancer risk factors, such as aging, smoking, alcohol consumption, and increased BMI. Blood DNA methylation signatures for aging, or ‘epigenetic clocks’, can be used to

estimate the biological aging, which can be affected by different exposures [214-216]. Biological accelerating aging (having an epigenetic predicted age higher than the actual age) has been associated with various diseases, including cancer [217, 218],

Smoking is the most well characterised exposure of the risk factors for DNA methylation [203, 210, 211, 219-223]. Two of the genes found associated with smoking status are the aryl hydrocarbon receptor repressor gene, *AHRR*, and the coagulation factor II (thrombin) receptor like gene, *F2RL3*. The gene *AHRR* is a tumour suppressor that is involved in the metabolism of toxins from tobacco smoking and interacts with polycyclic aromatic hydrocarbons (PAHs), which affect gene expression and DNA methylation. It is not fully understood how the gene *F2RL3* might mediate smoking-related risk; however it codes for a protein expressed in various tissues including blood and lung tissue, and has been associated with CVD, lung cancer, and mortality [210, 211]. Interestingly, some DNA methylation changes observed in smokers are reversible and can go back to normal levels after cessation, however, some DNA methylation changes can remain differentially methylated up to 35 years after smoking cessation [221].

Alcohol consumption is associated with DNA methylation alterations in blood [224-227]. There are several mechanisms linking together alcohol consumption and DNA methylation. For example, the oxidation of ethanol produces the carcinogen acetaldehyde that has an impact on DNMTs [66]. Alcohol consumption also decreases folate levels that are involved in the one-carbon metabolism, and elevates homocysteine levels that is metabolised to S-adenosyl methionine (SAM), an important methyl group donor [66, 228].

Obesity related traits, such as BMI, waist circumference and waist-to-hip ratio, and their impact on DNA methylation has been reported in numerous studies [212, 213, 229-232]. No consistent genes or CpG sites have been found between studies, however, a majority of the reported genes are involved in lipoprotein metabolism and inflammatory pathways. Furthermore, similar DNA methylation patterns have been observed in blood and replicated in adipose tissue, which supports the use of blood DNA as a surrogate tissue [212].

1.4. OESTROGEN EXPOSURE AND ITS ROLE IN BREAST CANCER

It is well established that oestrogens are involved in the aetiology of breast cancer, especially in postmenopausal women [233]. However, the biological mechanisms behind oestrogen exposure and breast cancer are not fully understood [137]. A woman is exposed to oestrogens during her whole lifetime, but the oestrogen levels will vary by age, menopausal status and different hormonal and lifestyle risk factors. Oestrogens have many important functions in the human body. This includes development of sexual characteristics, regulation of the menstrual cycle, maintenance of bone mass density, response to insulin, and cognitive function [51, 234]. Oestrogens are mainly produced in the granulosa cells in the ovaries of premenopausal women, the placenta of pregnant women, and the muscle, skin and adipose tissue of postmenopausal women. The most important oestrogens are oestradiol, oestriol and oestrone [51, 137, 234]. Oestradiol is dominant during a woman's reproductive time, oestriol during pregnancy and oestrone after menopause [235]. The enzyme aromatase, expressed by the gene *CYP19A*, converts androgens to oestrogens in a tissue-specific manner and is expressed in the ovaries, placenta, skin, adipose tissue, brain, and bone [51, 137, 234]. Oestrone is a weaker form of oestrogen and is produced in the skin and adipose tissue. The oestrogen production in subcutaneous tissues is increased with age and becomes the dominant source of oestrogens for postmenopausal women [51, 234]. In premenopausal women only 5% of the total oestradiol synthesis is in the subcutaneous tissues, and the total circulating oestrogen levels are controlled via a negative feedback system. Oestrogens inhibit the release of gonadotropins (luteinising hormone, LH, and follicle-stimulating hormone, FSH) in the brain, which controls the oestrogen production in the ovaries [236]. Therefore, in extremely obese women, the increased oestrogen synthesis in the adipose tissue leads to activation of the negative feedback system and can result in absence of menstruation (amenorrhea) [237]. The negative feedback system is reduced in postmenopausal women [236].

1.4.1 Oestrogen exposure and breast cancer

It is well established that high circulating oestrogen levels increase the risk for breast cancer in postmenopausal women [233]. For premenopausal women findings are more inconsistent [55, 238-241]. This might be due to issues with statistical power in premenopausal studies since breast cancer is less common before the menopause, difficulties to control for menstrual cycle day, or other confounding factors. Studies have showed that higher oestrogen levels in the luteal or follicular phase

of the menstrual cycle might be associated with increased breast cancer [240, 241], but non-significant findings for premenopausal oestrogen levels have also been observed [239]. Further studies are needed to understand the role of oestrogen and breast cancer in premenopausal women. However, from large epidemiological studies of different hormonal risk factors, known to affect both lifetime oestrogen exposure and breast cancer risk, it is clear that lifetime oestrogen exposure has an important role even in premenopausal women [55]. For example the number of menstrual cycles, as an approximation to lifetime oestrogen exposure, is positively associated with breast cancer risk [93, 94].

1.4.2 Mechanisms of oestrogen action

Oestrogen can affect gene expression through different mechanisms (**Figure 1.2**). This includes a transcriptional response via activity of the ERs and non-genomic response by binding to the membrane-bound G-protein coupled ER, GPER1, which triggers different fast non-genomic responses and is involved in several signalling pathways [242]. Also, different oestrogen-metabolites can have genotoxic effects by binding to the DNA and cause DNA damage. Also, oestrogen exposure can lead to changes in DNA methylation, which will be further explored in this project.

1.4.2.1 Transcriptional mechanisms

A transcriptional response is triggered when oestrogens (predominantly 17- β -oestradiol, E2) bind to and activate the intracellular ERs, ER- α and ER- β . The ERs have three domains: the NH₂ terminal, the DNA binding domain, and the ligand-binding domain. The NH₂ terminal has an oestrogen-independent transcriptional activation function involved in protein-protein interactions, and is the major functional difference between the ER- α and ER- β with more activity in the ER- α . The other domains are very similar between the two receptors [243, 244]. When oestrogen binds to the ERs they often undergo dimerization and translocate to the nucleus, where they interact with the DNA at oestrogen responsive elements (ERE) and recruit co-activators and other transcription factors. The outcome is an increased gene expression of ER-target genes often involved in cell proliferation, particularly for the ER- α , which is associated with cancer development [11, 137, 198, 243-245]. The role for ER- β is not clear but it might have antiproliferative effects [244, 246]. The majority of the ER-binding sites are distant to the target genes and often function as enhancers [11]. The ER-DNA binding is mediated by pioneer factors such as FOXA1 and GATA3, which form chromatin loops that contact with the promoters of specific oestrogen-responsive gene targets [11]. Tissue-specific ER-activity and oestrogen-

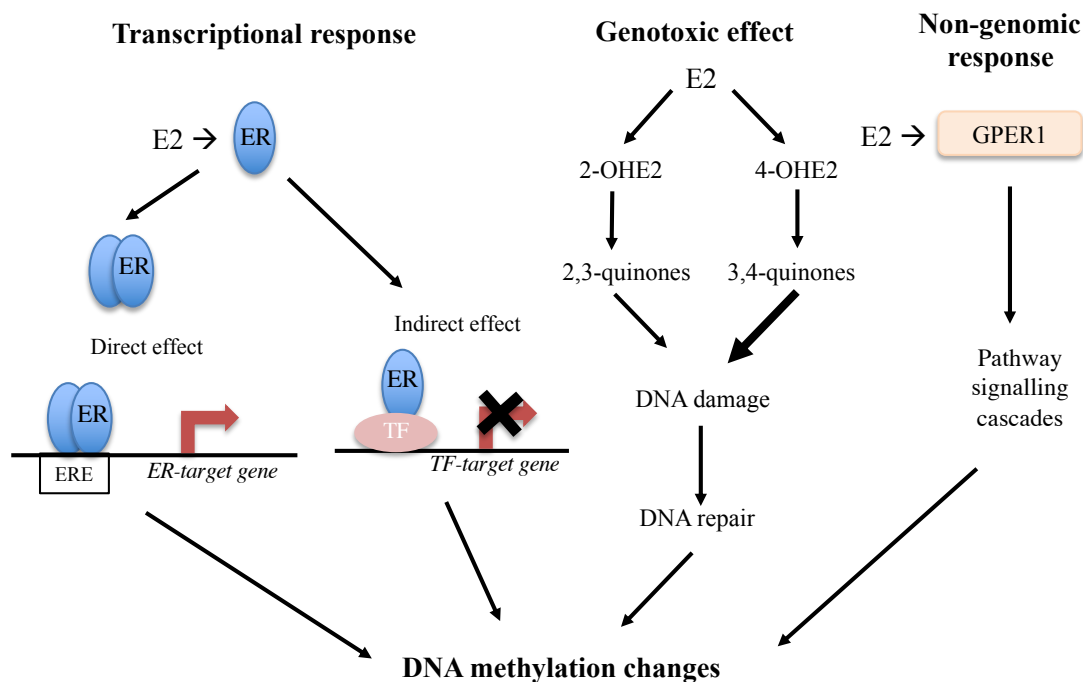


Figure 1.2. Mechanisms of oestrogen action

Oestrogen exposure has several biological outcomes, here categorised into transcriptional response, genotoxic effect and non-genomic response. Oestrogen (mainly 17- β -oestradiol, E2) can bind to the intracellular oestrogen receptors ER- α or ER- β , that undergo dimerization and bind to the DNA at oestrogen responsive elements (ERE), and have a direct transcriptional effect on ER-target genes, often involved in increasing cell proliferation. The ERs can also bind to other transcription factors (TF) and thereby indirectly affect transcription on other genes, both by increasing and inhibiting gene expression. The genotoxic effect involves the conversion to different oestrogen metabolites that can cause DNA damage, mainly from the formation of depurinating DNA adducts of 3,4-quinones. Oestrogen can also cause non-genomic response via the membrane-bound GPER1, that is involved in several pathway signalling cascades. It is hypothesised that these mechanism can lead to changes in DNA methylation.

responsiveness is dependent on the expression levels of FOXA1 and other factors [247], as well as the expression levels of different isoforms and subtypes of ER- α and ER- β [244]. Oestrogen and different SERMs, such as Tamoxifen, can both have agonist and antagonist effects on the ER, which is not completely understood. The ERs do not always undergo dimerization, and can indirectly interact with the DNA by interactions with other transcription factors, and are involved in both the inhibition and activation of their respective target genes. Oestrogens, via the ERs, can therefore both stimulate and inhibit gene expression. For example, it can inhibit the interleukin-6 (IL-6) by binding to the NF κ B complex, and activate the activation protein 1 (AP1) that is involved in the gene regulation of several important cellular processes [243].

1.4.2.1.1 ER-expression in breast cancer

The ER- β is less well characterised compared to the ER- α and has lower affinity for oestrogen, but oestrogen activity is mediated both through ER- α and ER- β [137, 234]. ER- β is expressed in 80% of the cells in normal breast cells, and loss of ER- β expression is observed in invasive oestrogen-dependent breast cancer [246]. This is suggested to be an early indicator of breast cancer progression and may be a result of a growth advantage for ER- α positive cells, where the breast cancer initiation from oestrogens starts. Surprisingly, the ER- α is expressed in only 10-15% of normal epithelial breast cells, and it is thought that these ER-positive cells influence surrounding ER-negative cells to grow [248]. In early breast cancer there are significantly more (60-80%) ER- α -positive cells, but these numbers are reduced in invasive breast cancer, indicating a role in the progression but not in invasion of breast cancer [246, 248]. The treatment regimes are complicated by the fact that the status of the ER is not homogenous in a tumour. When ER-positive breast cells are exposed to oestrogen, they produce signals that can cause neighbouring ER-negative cells to proliferate. The result is a mix of ER-positive and ER-negative breast epithelial cells in the tumour [107]. Also, oestrogen is most important in ER-positive breast cancer, but the different hormonal nuclear receptors ER, PR and androgen receptor (AR), and their respective activated hormones work in a complex relationship. These can affect each other's outcome and binding position. For example, the AR can work as a substitute to the ER and be recruited by the transcription factor FOXA1 instead of ER [11].

1.4.2.2 Genotoxic effects

Another ER-independent outcome of oestrogen exposure is DNA damage of different genotoxic oestrogen metabolites. The catechols 2- and 4-hydroxy-oestradiol (2-OHE2 and 4-OHE2) can oxidise to 2,3-quinones and 3,4-quinones, where the latter ones are far more carcinogenic than the former [249, 250]. The 2,3-quinones can interact with the DNA and form stable adducts, but the 3,4-quinones interact with the DNA more frequently and form unstable, depurinating adducts at adenine and guanine sites that lead to mutations and cancer development [251-253]. Indeed, higher levels of 4-OHE2 have been observed in breast cancer cases [251, 254]. Furthermore, the quinones can form reactive oxygen species (ROS) that can lead to oxidative damage of the DNA [253, 255]. The DNA damage is repaired by a number of different repair-pathways that either try to repair the damage or initiate apoptosis if extensive damage has occurred [256]. DNA damage, including DSBs and oxidative damage, lead to the recruitment of histone modifications, the maintenance methyltransferase DNMT1, and the *de novo* methyltransferase DNMT3B, which increase the DNA methylation. This leads to gene silencing and protection against replication of the damaged sites [257-260]. It is therefore hypothesised that a DNA repair mechanism following DNA damage can lead to mainly hypermethylation, and our group has previously shown an increase in DNA methylation in ovarian cancer following treatment with the DNA damaging agent platinum [261].

1.4.3 Oestrogens and DNA methylation

Oestrogen-treatment of breast cells leads to epigenetic alterations. For example, oestrogen-induced repression of the imprinted gene *CDKN1C* via ER- α signalling in ER-positive breast cancer cells. The repression was mediated via chromatin interactions and DNA methylation of the imprinted control region, which regulates the expression of the gene cluster including *CDKN1C* [262]. Also, oestrogen treatment resulted in promoter methylation and repression of the tumour suppressor gene *PTPRO*, facilitated via the ER- β and not the ER- α . Following Tamoxifen treatment, the expression of *PTPRO* was increased [263]. A number of other genes and large gene clusters have been found to be epigenetic silenced following oestrogen treatment [264-266]. Repression of large gene clusters might be via oestrogen-induced long-range DNA looping, partly mediated via *ESR1* [266]. This results in DNA hypermethylation and alterations in histone modifications, such as increasing H3K27me3, which is a marker for repressed promoters [265, 266]. Furthermore, genome-wide hypomethylation following

oestrogen treatment has been observed in a rat-model for human sporadic breast cancer [267, 268]. The mechanisms for oestrogen-induced DNA methylation are not fully understood. However, oestrogen-treatment also leads to up regulation of DNMT1, DNMT3A and DNMT3B [264, 268], which can explain promoter hypermethylation, but not the genome-wide hypomethylation. Interestingly, the morphological precancerous changes following oestrogen treatment have been observed later than the DNA methylation changes, indicating that epigenetic alterations can be early indicators for cancer development [267].

1.4.3.1 Hormonal risk factors and DNA methylation

Associations between reproductive risk factors and DNA methylation further support the hypothesis that lifetime oestrogen exposure alters DNA methylation. Parity-related epigenetic alterations in the breast are thought to be through changes in the oestrogenic activity following pregnancy. The FFTP initiates specific chromatin remodelling that works as a safeguard and protects the breast cells from cancer development [269]. More histone modifications for condensed chromatin (H3K9me2 and H3K27me3) have been observed in breast cells from parous women compared to nulliparous women, which protects the DNA from DNA damaging agents [269]. Nulliparity and later age at FFTP have been positively associated with DNA methylation of *ESR1* in breast tumours (n=99 women, no information of ER-status) [270]. Also, hypermethylation of *FOXAI* in breast cells of parous women has been observed, which will reduce the ability for DNA looping between enhancers and ER-target genes, and might be an important factor for a decrease in breast cancer risk [271]. Furthermore, methylation in the promoter of the tumour suppressor *p16* was observed in women who have not been breastfeeding [272].

Reproductive risk factors have also been associated with DNA methylation in blood. Inconsistent findings have been observed for DNA methylation and age at menarche; in a prospective population-based study cohort, later age at menarche was associated with decreased genome-wide DNA methylation measured with LUMA (n=332, OR=1.32, 95% CI: 1.14 – 1.53, $P < 1.0 \times 10^{-4}$), but this was not replicated using the HM450K array [273]. Another multi-ethnic study found that a later age at menarche was associated with higher DNA methylation levels using the [³H]-methyl acceptance assay in n=92 subjects [186]. These inconsistent findings might have been reflected from the different techniques used; for example, LUMA is more biased towards CpG-rich regions due to the use of restriction enzymes [185]. No difference in the LINE-1 assay throughout the menstrual cycle was

observed in a study of young, healthy, premenopausal women (n=53-56 samples from blood-donators at three different time points) [275]. It was suggested this was due to low ER-expression in blood, since a decrease in LINE-1 methylation was observed in oestrogen-treated ER-positive breast cancer cell line [275]. Age at menopause has been inversely correlated with DNA methylation in the repetitive elements Alu and LINE-1 in a study including n=161 hospital-recruited women, suggesting that DNA methylation can be used as a biomarker for age at menopause [274]. In a meta-analysis including three different blood datasets, accelerating epigenetic aging, estimated by using an age-associated DNA methylation signature, has been found inversely associated with age at menopause [276]. Also number of children has been inversely associated with epigenetic aging in a prospective population-based cohort study (n=105) [216].

Differentially methylated CpG sites, mainly hypomethylation, between HRT users and non-users have been observed in twin pairs (n=20 pairs from a twin cohort) [277], and long-term use of OC has been associated with genome-wide hypomethylation (n=67 OC users compared to n=72 non-OC users, hospital-recruited women) [278]. In an asthma study, DNA methylation levels linked to a SNP in *GATA3* were associated with OC use (n=34 young women from a prospective study cohort) [279].

In addition to this, DNA methylation changes in relation to breast cancer have been observed from different lifestyle risk factors. This supports oestrogen-dependent mechanisms that lead to alterations in DNA methylation and breast cancer development. In ER-positive breast cancer tumours from obese women, hypermethylation at several loci with functions such as immune response, cell growth, DNA repair and DNMT coding genes, has been observed (n=90 obese compared to n=118 normal-weight women in a prospective case-control study) [280]. The oestrogen production in the adipose tissue of obese women is regulated through tissue-specific promoter methylation of genes and affects aromatase expression, which might be an important factor in postmenopausal breast cancer [281].

Alcohol consumption has been associated with increased methylation of *CDH1*, which is often found in breast cancer, and decreased methylation of *p16* in postmenopausal breast tumours in a case-case study (n=803) [282]. Another study observed increased DNA methylation and lower expression levels of *p16* in normal breast cells, as a result of lower folate levels following increased alcohol consumption (cross-sectional study of women undergoing reduction mammoplasty, n=138)

[283]. Besides affecting the one-carbon metabolism, ethanol might also alter activity of DNMTs [66, 282, 283].

Smoking is well known to alter DNA methylation, but methylation changes associated with breast cancer are not conclusively detected. However, DNA methylation at the 2q37.1 locus in blood associated with smoking has been suggested to be associated with breast cancer risk in a prospective population study cohort (n=374) [203]. Other potential blood DNA methylation changes from smoking might be the decreased methylation of *BRCA1* and *ESR1*, as well as tumour suppressor candidates *HIN1* and *DAPK*, observed in cases from a population-based study (n=765-851) [284].

Furthermore, physical inactivity has been positively associated with increased global methylation in two different population-based studies (n=647 women with family history of breast cancer and n=161 cancer-free women) and this could be mediated by changes in oestrogen levels [285, 286]. Promoter methylation of the tumour suppressor gene *APC* has been associated with breast cancer risk and was observed to be inversely associated with physical activity in a cross-sectional study in breast samples from women without breast cancer (n=45) [287]. The *APC* promoter hypermethylation is associated with high oestrogen levels, so it was hypothesised that the changes in oestrogen levels from physical activity lead to this change [287].

All together, this suggests that oestrogen exposure is important in early breast cancer development, and that oestrogens can cause DNA methylation changes detectable in blood. However, the mechanisms between lifetime oestrogen exposure, DNA methylation, and breast cancer risk are still not clear.

1.5. AIMS AND HYPOTHESES

A woman's breast cancer risk is influenced throughout her lifetime by several reproductive and lifestyle risk factors and her total exposure to oestrogen. However, the mechanisms between oestrogen exposure and breast cancer development are not fully understood. Breast cancer risk models need further improvements to be able to stratify women into low and high-risk groups. If women with high risk for breast cancer can be identified, and prevention approaches can be targeted for these women, the breast cancer incidence rate could be reduced.

In this PhD project the association between lifetime oestrogen exposure, DNA methylation measured in blood, and breast cancer risk was further assessed. It was hypothesised that the lifetime oestrogen exposure accumulates changes in DNA methylation detectable in blood, which are associated with breast cancer risk. A DNA methylation signature in blood following oestrogen exposure might add valuable information and improve breast cancer risk assessment.

It was aimed to first perform an epigenome-wide association study (EWAS) using HM450K array data to identify potential CpG sites altered by the lifetime oestrogen exposure (**Chapter 2**). For this, an Estimate Lifetime (o)Estrogen Exposure (ELEE) model was identified. A Methylation Index (MI) was developed, including CpG sites found to be associated with ELEE in the EWAS, and assessed for association with breast cancer risk. Validation of the association between the MI and breast cancer risk was conducted in a large independent study cohort using targeted bisulphite sequencing (**Chapter 3**), and a meta-analysis for replication of the MI was conducted including a total of four independent case-control study cohorts (**Chapter 4**). Lastly, the mechanisms of oestrogen-induced changes in DNA methylation in blood were explored (**Chapter 5**).

CHAPTER 2:
Epigenome-Wide Association Study
of Estimated Lifetime Oestrogen
Exposure

2 CHAPTER 2: Epigenome-Wide Association Study of Estimated Lifetime Oestrogen Exposure

2.1. INTRODUCTION

2.1.1 EWAS for breast cancer risk using blood DNA

Endogenous and exogenous exposures leave epigenetic signatures, which are hypothesised to work as prognostic disease biomarkers. Alterations in blood DNA methylation have been observed in breast cancer, but as reviewed, previous EWAS for breast cancer, using DNA methylation data measured in blood in a nested case-control design, have so far not identified any strong methylation signatures for risk prediction [208]. In this study, the hypothesis was that circulating oestrogens alter DNA methylation detectable in blood, and the association between lifetime oestrogen exposure as a risk factor for breast cancer and blood DNA methylation was explored.

2.1.2 Previous breast cancer case-control EWAS

The study reporting the most number of significant CpG sites associated with breast cancer risk identified 250 differentially methylated probes using the Illumina HumanMethylation27K (HM27K) array in 298 cases and 612 controls in the prospective US Sister Study (including women not diagnosed with breast cancer themselves at recruitment but with a sister with breast cancer) [288]. A majority of the CpG sites (>75%) were hypomethylated in cases compared to controls and mostly located in CpG islands. Unfortunately, the differences in methylation levels between cases and controls were often very small, ranging from 0.001 to 2.53%. Small changes in DNA methylation could make them difficult to be used as biomarkers and to replicate in other studies. Furthermore, they developed a methylation signature including 57 CpG sites that showed a better risk prediction than the Gail model and nine risk-SNPs with an area under the curve (AUC) of 0.66 (95% CI: 0.61 – 0.71), however, this was over-inflated due to the bias in the probe selection (the same dataset was used in the discovery phase). The most recent breast cancer EWAS identified 24 heritable CpG sites where carrier probabilities were significantly (Bonferroni P -value threshold $< 1 \times 10^{-5}$) associated with breast cancer risk in an Australian multiple-case breast cancer family study (210 individuals from 25 families). Of

these, four CpG sites, located in relation to the genes *GREB1*, *PNKD*, *C7orf50* and *TMC3*, were replicated to be associated with breast cancer risk in another study cohort [289].

2.1.2.1 Gene specific DNA methylation

Some studies have focused on gene-specific DNA methylation in relation to breast cancer risk. Mutations in *BRCA1* increase the breast cancer risk and hypermethylation in the *BRCA1* promoter in blood DNA has been observed in both retrospective and prospective cases-control studies [199, 200]. In the prospective study (n=72 *BRCA1* mutants and n=72 *BRCA1* wild-type), a *BRCA1* mutation DNA methylation signature of 1,829 CpG sites was developed that predicted breast cancer in two different UK based study cohorts (n=19 cases vs n=77 controls, AUC=0.65, 95% CI: 0.51 – 0.78, $P=0.02$; n=119 cases vs 122 controls, AUC=0.57, 95% CI: 0.50 – 0.64, $P=0.03$) as well as breast cancer mortality (n=119 cases vs 122 controls AUC=0.67, 95% CI: 0.51 – 0.83, $P=0.02$) [199].

Furthermore, hypermethylation in the *ATM* intragenic loci ATMmvp2b has been observed for bilateral breast cancer patients compared to controls in another UK based study cohort (n=190 case-control pairs, $OR_{Q4_vs_Q1}=3.20$, 95% CI: 1.78 – 5.86, $P=8.3 \times 10^{-5}$) and this was replicated in two out of three prospective study cohorts with an OR for the highest versus lowest quintiles of 1.89 (95% CI: 1.27 – 2.82, $P=1.7 \times 10^{-3}$) in a meta-analysis [201, 202].

An EWAS of smoking, including n=374 women from an Italian breast cancer-case cohort from the European prospective into cancer and nutrition (EPIC) study, observed hypomethylation at the loci 2q37.1 that was significantly associated with breast cancer risk. They also observed a strong correlation with HM450K methylation levels and case-control status. A total of 533 CpG probes were associated with breast cancer survival, which could be predicted between groups with different DNA methylation profiles [203].

2.1.2.2 Genome-wide DNA methylation

Another approach is to study DNA methylation globally by analysing the genome-wide DNA methylation levels and breast cancer risk. Genome-wide hypomethylation in 420 cases matched to 420 controls was observed in an Australian prospective study with an OR per 1 standard deviation (OR_{perSD}) of 0.69 (95% CI: 0.50 – 0.95, $P=0.02$). They observed an increase in DNA methylation at promoters and a decrease outside promoters [204]. Hypomethylation was observed for two out of three prospective study cohorts (one being the Australian study just mentioned, the other one being EPIC-

Italy that is also used in this study) in a meta-analysis with a summarised ORperSD of 0.61 (95% CI: 0.47 – 0.80, $P=4 \times 10^{-4}$) [197]. The hypomethylation was mostly located to gene bodies and intragenic regions. Another nested case-control study with samples from the prospective cohort EPIC did not observe any significant difference in genome-wide DNA methylation between 451 cases matched to 451 controls, but observed significant CpG island hypermethylation with an ORperSD of 1.20 (95% CI: 1.03 – 1.40, $P=0.02$) [206].

Studies measuring the genome-wide DNA methylation levels using pyrosequencing of repetitive elements as a surrogate measurement have frequently not found any significant association with risk in either retrospective or prospective studies [195, 205]. However, one study found a significant decrease in DNA methylation of LINE-1 elements in cases compared to controls in the Sister Study (n=294 cases and n=646 non-cases, $HR_{Q1_vs_Q4}=1.75$, 95% CI: 1.19 – 2.59, $P=0.002$) [290], and another American population-based study (n=1,055 cases and n=1,101 controls) observed significant increase in DNA methylation in cases compared to controls ($OR_{Q5_vs_Q1}=2.41$, 95% CI: 1.83 – 3.16, $P < 1.0 \times 10^{-4}$) using the LUMA, but this was not repeated for LINE-1 DNA methylation. This was thought to be due to differences in the techniques; the LUMA measures DNA methylation of a repetitive motif (C^mCGG) that is over-represented in promoters, however, the hypermethylation in promoters was also observed in the HM27K array [291].

In summary, studies of blood-based DNA methylation as a breast cancer risk marker are inconsistent. Differences in study cohorts, techniques to measure DNA methylation levels, statistical methods, and small sample sizes (resulting in low statistical power), make it difficult to compare the results between studies. Other methylation studies not mentioned above have used retrospective data [205], which cannot be used to draw conclusions on DNA methylation for risk prediction due to the risk of confounding reverse causation, i.e. that the cancer itself is causing the changes in DNA methylation. There is a need for larger, collaborative studies as well as replication and validation steps to conclude if DNA methylation signatures in blood can be used for breast cancer risk prediction.

2.1.3 EWAS of risk factors using blood DNA

Different lifestyle factors have been shown to alter DNA methylation and increase the risk for diseases such as cancer. EWAS for risk factors have successfully identified methylation signatures for aging, smoking, alcohol consumption, and BMI in blood DNA. The signatures could be a more reliable option

to self-reported data and be used in studies of underlying mechanisms and the impact between risk exposures, DNA methylation and diseases.

2.1.3.1 Age

Aging is the biggest risk factor for nearly all cancers (except paediatric cancers), including breast cancer, and hundreds of age-related DNA methylation alterations, both hypo- and hypermethylation, has been observed across multiple tissues, including blood, brain, kidney, lung, buccal cells and brain [214-216]. Different DNA methylation signatures, or ‘epigenetic clocks’ can predict age with high accuracy and might be used to reflect the biological aging affected by different exposures [214, 215]. However, the correlation between the epigenetic clocks and aging seems to be lower in breast tissue, which might be due to hormonal effects [215]. Aging-associated hypermethylated CpG sites are more commonly shared across cell types, while hypomethylated CpG sites are more often observed in blood than other tissues [215, 292, 293]. The hypermethylated CpG sites are enriched for CpG islands and gene promoters with bivalent chromatin marks (both active and inactive markers), which are important in development and often silenced in cancer [293].

Accelerated epigenetic aging in blood has been associated with cancer, CVD, all-cause mortality, and postmenopausal breast cancer risk [206, 217, 218]. In breast cancer, the epigenetic age acceleration differs between different breast cancer subtypes with luminal A and B showing the highest epigenetic age acceleration [215], which however might have been a result of older age at diagnosis for these subtypes compared to e.g. basal-type breast cancer. Furthermore, the age acceleration is influenced throughout the lifetime by different exposures and has been associated with smoking [294], alcohol consumption [216], and obesity and BMI [216, 295, 296].

2.1.3.2 Smoking

Numerous studies of DNA methylation and smoking have been conducted over the years, and it is clear that smoking alters DNA methylation in blood [203, 210, 211, 219-223]. The two top genes are *AHHR* and *F2RL3*, but several other genes have been found associated with smoking [210, 211]. For some CpG sites there seem to be a clear dose-response relationship between cigarette consumption and DNA methylation [219], and both stable and reversible CpG sites after smoking cessation have been identified [76, 221]. An EWAS on time since smoking cessation observed two groups of smoking-associated CpG sites: reversible CpG sites where the DNA methylation levels goes back to normal

levels after smoking cessation, and more stable, irreversible CpG sites where DNA methylation levels remained differentially methylated up to 35 years after smoking cessation [221].

Another EWAS of smoking using the HM450K array observed hypomethylation for *F2RL3*, *AHRR* and two intergenic loci 2q37.1 and 6p21.33. This was replicated and validated using bisulphite pyrosequencing in both blood and lung tissue, supporting the use of blood DNA as a surrogate measurement for methylation in lung tissue [203]. In the same study a methylation index was developed using DNA methylation levels at *AHRR*, 6p21 and 2q37 that could distinguish former smokers from never smokers with an AUC of 0.83 (95% CI: 0.79 – 0.96) in the validation dataset (n=180) [223]. It has also been shown that CpG sites alone can distinguish smokers from non-smokers with high accuracy up to an AUC of 0.99 [222].

Furthermore, DNA methylation at smoking-associated CpG sites has been associated with various disease outcomes. In an EWAS for lung cancer in a Norwegian prospective study cohort hypomethylation at the CpG sites in *AHRR* and *F2RL3* were significantly associated with lung cancer risk with an OR of 0.37 for *AHRR* methylation (95% CI: 0.31 – 0.54, $P=3.33 \times 10^{-11}$, n=132 case-control pairs). This was replicated in four other prospective study cohorts, and a risk prediction model was developed using smoking status and methylation of *AHRR* and *F2RL3* that could predict lung cancer with high accuracy (AUC=0.78) [220]. Another German prospective study further explored the association between methylation at *F2RL3* and lung cancer and found 33% increased risk for lung cancer per 10% decrease in methylation levels. Their risk model prediction for lung cancer showed an AUC of 0.86 (95% CI: 0.81 – 0.91) when combining DNA methylation at *F2RL3* and pack years smoking for subjects aged 65 or above [211].

2.1.3.3 Alcohol consumption

There is a clear link between alcohol consumption and DNA methylation changes, since ethanol has an affect on the one-carbon metabolism [66, 228]. Unfortunately, most of the methylation studies of alcohol consumption or alcohol dependence have had low sample sizes (n < 100), used DNA extracted from different white blood cell (WBC) types (lymphoblast, lymphocytes, peripheral blood mononuclear cells (PBMNCs), CD3+ T-cells), and have reported inconsistent results. One study reported decreased methylation levels in abstinent compared to mild, moderate and heavy drinking individuals [297], a second one a majority of hypomethylation in alcohol dependent patients [224], a

third one mostly hypermethylation in individuals with chronic alcohol consumption [226], and a fourth one a mixture of hypo- and hypermethylated changes [227]. There are no overlaps among top genes between the studies, however, they mostly report genes are associated with alcoholism [224-226].

Combining methylation and GWAS data, one larger Swedish study (n=619) identified 17 SNP-CpG regions significantly associated (false discovery rate, FDR, Q -value < 0.1) with alcohol use using methyl-CpG binding domain protein capture sequencing [298]. The recent largest prospective EWAS and meta-analysis (total n=13,317) used the HM450K array and developed DNA methylation signatures including 5 to 144 CpG sites that could discriminate heavy drinkers from non-drinkers with AUC values > 0.8 across all five study cohorts. Similar DNA methylation patterns were observed in former and never drinkers, suggesting that DNA methylation levels are reversible after alcohol abstinence, similar to what was observed in former smokers [299]. Similar, another study in CD3+ T-cells observed reversible DNA methylation levels at seven CpG loci after treatment for chronic alcohol dependence [227].

DNA methylation alterations following alcohol consumption have been linked to liver disease and various cancers, including liver, colorectal, head and neck [66]. Furthermore, in relation to breast cancer, alcohol increases oestrogen levels in blood and decreases folate levels, which is associated with increased risk for postmenopausal breast cancer [66].

2.1.3.4 BMI

BMI is associated with DNA methylation changes detectable in blood, but no consistent genes or CpG sites have replicated across studies [212, 213, 229-232]. One European consortium identified five CpG probes significantly associated with BMI, whereof three were located in the gene *HIF3A*. The gene *HIF3A* is a component of the hypoxia inducible transcription factor, which regulates a variety of cellular and physiological responses to hypoxia, and has been associated with CVD and cancer. The association between the three CpG sites in *HIF3A* and BMI was replicated in adipose tissue, which again supports the use of blood DNA as a surrogate tissue [212].

Other EWAS have observed associations between BMI and DNA methylation for genes involved in different metabolic functions: *ABCG1* associated with blood lipid levels [213, 230, 232], *SREBF1* involved in cholesterol biosynthesis [213, 230, 231], and *PHGDH* involved in serine biosynthesis [213, 229]. These have also been associated with various diseases: promoter methylation at the gene *ABCG1* has been associated with coronary heart disease, *SREBF1* has been associated with

diabetes traits [230], and overexpression of *PHGDH* is found in 70% of ER-negative breast cancer [229]. Furthermore, in one of the most recent and largest prospective EWAS of BMI (total n=10,261) similar DNA methylation patterns were observed in both blood and adipose tissue for these genes [213].

Overall, these studies confirm that BMI is associated with DNA methylation changes in blood and that these signatures can be linked to various diseases. The adipose tissue is the main source of oestrogen for postmenopausal women, and obesity increases the oestrogen levels by increased expression of the enzyme aromatase that converts androgens to oestrogen. This links together obesity, DNA methylation, and ER-positive postmenopausal breast cancer [51].

2.1.4 Estimation of lifetime exposure to oestrogen

Epidemiological studies have identified several hormonal breast cancer risk factors (**Table 1.1**). These risk factors will affect a woman's oestrogen exposure either by altering oestrogen levels directly (e.g. being overweight or obese for postmenopausal women), in the long-term perspective (parous postmenopausal women have lower oestrogen levels compared to nulliparous women), or by defining the time when a woman is exposed to high levels of oestrogens through ovulatory cycles (age at menarche and menopause). Therefore, a combination of risk factors could be used to estimate a woman's lifetime exposure to oestrogens. This has been done previously in studies of Alzheimer's disease [235], cognitive performance [300], prospective memory [301], CVD [302], ischemic stroke [303] and mortality [304], where higher oestrogen has a protective role, and for breast cancer [154, 305]. However, the models vary from simple models calculating a woman's reproductive span (age at menopause minus age at menarche) [303, 304], to calculating a woman's total number of menstrual or ovulatory cycles [235, 302], and more complex combinations of reproductive factors including postmenopausal weight or BMI, duration of HRT and time since menopause [300, 301, 305]. For now, there is no standardised, generalised model to estimate a woman's lifetime exposure to oestrogen, and in this study a new model has been proposed.

2.1.5 Statistical methods in EWAS

In this study methylation data measured using the HM450K array has been used. The array measures DNA methylation levels at 485,512 CpG sites across the genome designed to focus on CpG sites in promoters, CpG islands, and enhancer regions. The DNA is bisulphite converted and the array generates data based on the signals from protected, methylated cytosines and unprotected, unmethylated thymines in a CpG context at a query probe site. Beta-values represent the proportions of methylated cytosines at given positions in the range of 0 to 1, where 1 is equal to 100% methylation. The beta-values follow a bimodal beta distribution with peaks around 0.2 and 0.8 with non-constant variances known as heteroscedasticity. The variances are often smaller for beta-values close to 0 and 1 and larger for beta-values around 0.5.

The beta-values are challenging some of the assumptions for linear regression modelling, where the values are assumed to follow a normal distribution with a constant variance (homoscedasticity). To overcome this problem previous EWAS have often used log2-transformation of the beta-values (M-values) prior to the regression analysis. The M-values tend to be more normally distributed with less heteroscedasticity (**Figure 2.1**), however, this is not always the case and they lose their biological interpretability [306-308]. M-values tend to perform better at the low and high ranges compared to beta-values, but the performance for the middle-range remains similar [306].

Recently beta regression has been proposed as an alternative approach for modelling continuous proportions in a range, such as beta-values [309]. The beta regression is a parameterised generalised linear model with two parameters: μ for the mean and Φ for the precision (or the variance), and handles heteroscedasticity and skewed distributions better. The mean is linked to the responses through a link function and a linear predictor [310, 311]. Beta regression has been applied in a few recent methylation studies [312-317] and might be a better statistical approach in methylation studies using array data compared to linear regression.

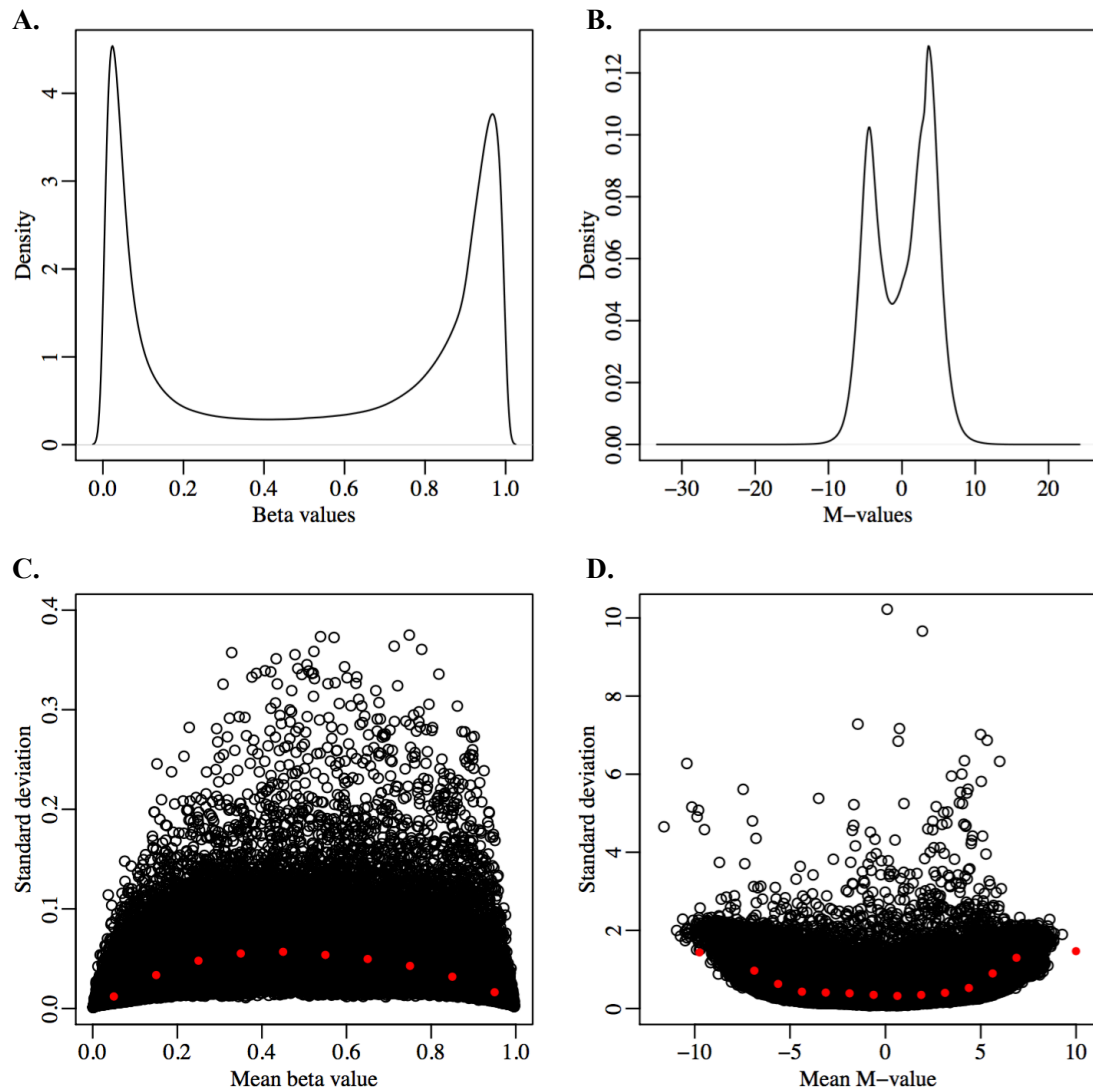


Figure 2.1. Beta-values and M-values in HM450K methylation array data

Reconstruction of figures from Du *et al.* (2010) [301] using EPIC-Italy HM450K data (n=324). **A.** Distribution of beta-values. **B.** The beta values are often $\log_2$ -transformed to M-values in EWAS, which are more normally distributed. **C.** The beta-values have non-constant variances, known as heteroscedasticity. The plots shows the mean beta-values plotted against standard deviations, red dots indicate mean standard deviations. **D.** Mean M-values against standard deviations, red dots indicate mean standard deviations. The M-values have more constant variances (homoscedasticity).

2.1.5.1 Inflation and bias in EWAS

In comparison with GWAS, EWAS are more sensitive to technical and biological confounding effects. This might lead to incorrect test statistics by inflation and bias, resulting in an overestimation in the significance, a large number of false positives and biased estimates of effect sizes. The genomic inflation factor is used in GWAS to estimate the inflation, however, this is not suitable in EWAS where there might be a larger number of expected true associations. To control for this a Bayesian method has been proposed that fits a three component normal mixture to the observed set of test statistics using a Gibbs sampling algorithm [318]. The method captures the fraction of true associations observed in the data that is assumed to be an unknown minority of the tests and estimates the inflation and bias without being affected by this.

2.1.6 Aims of this chapter

In this study the aim was to further explore hormonal risk factors, focusing on lifetime oestrogen exposure, and DNA methylation in blood. The results from breast cancer EWAS have been inconsistent, but previous studies have successfully identified DNA methylation signatures for smoking, alcohol consumption, BMI and aging, and associated these with various diseases. Therefore, it was hypothesised that an EWAS of hormonal risk factors could be a better approach to identify CpG sites associated with breast cancer compared with a traditional case-control design. In this chapter an Estimated Lifetime (o)Estrogen Exposure (ELEE) model has been defined and it was hypothesised that ELEE will affect DNA methylation detectable in blood. An EWAS of ELEE has been performed to identify CpG sites where DNA methylation levels are associated with lifetime oestrogen exposure.

2.2. METHODS

2.2.1 Study cohort

Participants in this study were from the Italian sub cohort of the European prospective investigation into cancer and nutrition (EPIC) study, The EPIC is a multi-centre prospective cohort study including epidemiological data from over 521,000 participants from 23 centres across 10 European countries. The study was designed to investigate cancer and other chronic diseases and the relationships with diet, lifestyle and environmental factors. Healthy participants were recruited between 1992 and 2000 and questionnaire and interview data has been collected on diet and non-dietary variables (such as reproductive and lifestyle factors), anthropometric measurements, and blood samples. The study is followed up for cancer, other disease occurrences and mortality [319, 320]. EPIC-Italy consists of over 57,000 participants recruited between 1993 and 1998 across five Italian centres: Varese, Turin, Florence, Ragusa and Naples. After consent and ethical review from the IARC, questionnaire data and blood samples were collected [321]. The follow-up loss has been reported to be less than 2% and incident cancer cases have been identified from cancer registries [197].

2.2.2 Hormonal factors and breast cancer risk in EPIC-Italy

2.2.2.1 *EPIC-Italy questionnaire data*

At the start of this study EPIC-Italy included epidemiological questionnaire data from 32,059 women, of whom 1,388 (4.3%) had been diagnosed with breast cancer at last follow-up (July 2012). ER-status was missing for a majority of the breast cancer cases and therefore cases were stratified on age at diagnosis ≥ 50 to enrich for ER-positive disease, which is more common in older cases. This left 1,193 prospective cases for the study with an average time to diagnosis of 8.2 years (**Table 2.1**).

2.2.2.2 *Risk factors analysed*

Hormonal reproductive and lifestyle factors and the association with breast cancer risk were explored in EPIC-Italy questionnaire data. The following reproductive risk factors (all variables were reported at the time of recruitment) were explored: age at menarche, age at menopause, parity (parous versus nulliparous), number of full-term pregnancies, age at first pregnancy, if a woman had ever breastfed, and total lifetime breastfeeding duration (in years). Explored lifestyle risk factors included BMI, ever

Table 2.1. Table of characteristics for EPIC-Italy questionnaire data

The EPIC-Italy questionnaire data was used to explore different hormonal factors and breast cancer risk. All analyses were adjusted for age and cases were stratified on age at diagnosis ≥ 50 to enrich for ER-positive disease (ER status was missing for most of the subjects)

			EPIC-Italy questionnaire data			<i>P</i> value***
			All (n=31,864)	Cases (n=1,193)	Controls (n=30,671)	
Age	mean (st.dev.), yrs		50.7 (8.1)	53.2 (6.8)	50.6 (8.1)	$< 1 \times 10^{-16}$
Time to diagnosis	mean (st.dev.), yrs		8.2 (4.2)	8.2 (4.2)	NA	
Menopausal status	n (%)	<i>Premenopausal</i>	15563 (48.8%)	477 (40.0%)	15086 (49.2%)	2×10^{-10}
		<i>Postmenopausal</i>	15866 (49.8%)	703 (58.9%)	15163 (49.4%)	
Age at menarche	mean (st.dev.), yrs		12.5 (1.5)	12.5 (1.4)	12.5 (1.5)	0.342
Age at menopause	mean (st.dev.), yrs		48.3 (4.9)	48.9 (4.9)	48.3 (4.9)	0.001
Parity	n (%)	<i>Parous</i>	28185 (88.5%)	1060 (88.9%)	27125 (88.4%)	0.843
		<i>Nulliparous</i>	3266 (10.2%)	120 (10.1%)	3146 (10.3%)	
Number of pregnancies	mean (st.dev.)		1.9 (1.1)	1.8 (1.0)	1.9 (1.1)	0.044
Ever breastfed*	n (%)		22773 (80.8%)	835 (78.8%)	21938 (80.9%)	0.096
Breastfeeding duration*	mean (st.dev.), yrs		0.6 (0.6)	0.6 (0.6)	0.6 (0.6)	0.361
OC ever	n (%)		12871 (40.4%)	437 (36.6%)	12434 (40.5%)	0.006
OC duration	mean (st.dev.), yrs		1.6 (3.1)	1.5 (3.1)	1.6 (3.2)	0.144
HRT ever	n (%)		4961 (15.6%)	222 (18.6%)	4739 (15.5%)	0.004
HRT duration	mean (st.dev.), yrs		0.3 (1.2)	0.4 (1.2)	0.3 (1.2)	0.169
BMI	mean (st.dev.), kg/m ²		25.8 (4.4)	26.3 (4.5)	25.8 (4.3)	5×10^{-5}
Drink alcohol	n (%)	<i>Yes</i>	18937 (59.4%)	737 (61.8%)	18200 (59.3%)	0.523
		<i>Previously</i>	6778 (21.3%)	282 (23.6%)	6496 (21.2%)	
		<i>No</i>	602 (1.9%)	21 (1.8%)	581 (1.9%)	
Alcohol consumption	mean (st.dev.), intensity**		2.6 (2.7)	2.4 (2.7)	2.6 (2.7)	0.105
Smoking status	n (%)	<i>Smoker</i>	8240 (25.9%)	285 (23.9%)	7955 (25.9%)	0.072
		<i>Former</i>	6323 (19.8%)	223 (18.7%)	6100 (19.9%)	
		<i>Never</i>	16888 (53.0%)	672 (56.3%)	16216 (52.9%)	
Smoking duration	mean (st.dev.), yrs		6.5 (11.4)	6.9 (12.2)	6.5 (11.4)	0.283

st.dev = standard deviation; yrs = years

* Breastfeeding variables in parous women only

**Alcohol consumption intensity was calculated as average reported intensity (1-9) from beer, wine, and spirits at ages 20, 30, 40, and 50.

*** *P*-values indicate differences between cases and controls: Student's t-test was used for continuous variables and chi-squared test for categorical variables.

and duration use of OC and HRT (in years), alcohol drinking and alcohol consumption intensity, and smoking status, duration and intensity. Alcohol drinking was summarised from the answers “Yes”, “Never” and “Previously” reported for wine, beer, cider, and spirits or liquor. The intensity for alcohol consumption was calculated as average reported intensity (self-reported numbers 1 to 9) at ages 20, 30, 40 and 50. Smoking status was reported as smoker, former smoker or never smoker. Duration of smoking was calculated as the time (in years) between age at start smoking and age at stop smoking (or age at recruitment for smokers) minus reported number of years giving up smoking. Smoking intensity was calculated as average reported smoking intensity (self-reported numbers 1 to 9) at ages 20, 30, 40 and 50 for smokers and former smokers.

2.2.2.3 Statistical analysis

The association between the risk factors and breast cancer was estimated using Cox proportional hazards regression (function ‘coxph’ in R package ‘survival’) adjusted for age at recruitment [322, 323]. The number of years between date at recruitment and date at diagnosis was used as the time-to-event variable for cases and the time between date at recruitment and date at last follow up for controls. All risk factors were explored as continuous or categorical variables when suitable in all, premenopausal and postmenopausal women.

2.2.3 The ELEE model

Different combinations of time-dependent breast cancer risk factors were explored for the ELEE model in the EPIC-Italy questionnaire data. The associations between the ELEE models and breast cancer risk were estimated using Cox proportional hazards regression (function ‘coxph’ in R package ‘survival’) adjusted for age at recruitment. The explored variables included a woman’s reproductive time (age at menopause minus age at menarche for postmenopausal women and age at recruitment minus age at menarche for premenopausal women), number of pregnancies, breastfeeding duration and duration of OC and HRT use. The selected ELEE model included as many of these risk factors as possible without affecting the association with breast cancer risk in pre- and postmenopausal women.

Table 2.2. Table of characteristics for EPIC-Italy HM450K methylation data

The data set was used in the EWAS of ELEE. Cases were stratified on age at diagnosis ≥ 50 to enrich for ER-positive disease and subjects with missing information of ELEE were excluded, leaving 216 subjects for the analysis.

			EPIC-Italy (n=216)
Prospective cases	n (%)		97 (44.9%)
Age	mean (st.dev.), yrs		54.2 (6.4)
Time to diagnosis	mean (st.dev.), yrs		5.3 (4.5)
Menopausal status	n (%)	<i>Premenopausal</i>	58 (26.9%)
		<i>Postmenopausal</i>	132 (61.1%)
Age at menarche	mean (st.dev.), yrs		12.8 (1.5)
Age at menopause	mean (st.dev.), yrs		49.7 (3.6)
Parity	n (%)	<i>Parous</i>	180 (83.3%)
		<i>Nulliparous</i>	36 (16.7%)
Number of pregnancies	mean (st.dev.)		1.7 (1.1)
Ever breastfed*	n (%)		180 (100%)
Breastfeeding duration*	mean (st.dev.), yrs		0.7 (0.7)
OC ever	n (%)		75 (34.7%)
OC duration	mean (st.dev.), yrs		5.6 (6.0)
HRT ever	n (%)		35 (16.2%)
HRT duration	mean (st.dev.), yrs		2.1 (3.0)
BMI	mean (st.dev.), kg/m ²		25.6 (4.3)
Alcohol consumption	mean (st.dev.), average g/day		6.8 (8.9)
Smoking status	n (%)	<i>Smoker</i>	41 (19.0%)
		<i>Former</i>	42 (19.4%)
		<i>Never</i>	133 (61.6%)

* Breastfeeding variables in parous women only
st.dev = standard deviation; yrs = years

2.2.4 EWAS of ELEE

2.2.4.1 *EPIC-Italy HM450K data*

The HM450K array data was measured in DNA samples extracted from blood in 332 women from the EPIC-Italy sub cohort, including 166 incident breast cancer cases and 166 healthy controls matched on age at recruitment $\pm$ 5 years and study centre [197]. Blood samples were collected at recruitment, processed the same day and stored in liquid nitrogen [321]. Matched pairs were excluded if 5% of the measured probes in either a case or a control had low detection *P*-values, leaving 324 women (162 incident breast cancer cases and 162 controls) for the analysis. Subjects with missing information of ELEE were excluded (n=87) as well as cases with age at diagnosis less than 50 (n=28), to enrich for ER-positive disease, leaving 216 women for the EWAS of ELEE (**Table 2.2**).

2.2.4.2 *Generation and processing of HM450K array data*

DNA extraction from blood samples and methylation array processing were conducted in the Italian Institute for Genomic Medicine (previously known as the Human Genetics Foundation), as described previously [197]. For the 332 samples 500ng of DNA were bisulphite converted using the EZ-96 DNA Methylation-Gold Kit (Zymo Research) and hybridised on the Illumina Infinium HumanMethylation450K BeadChip. The BeadChips were scanned using the Illumina HiScan SQ scanner and intensities were extracted from the images using GenomeStudio (v.2011.1) and its Methylation module (1.9.0). Correction for the beta-value distributions from the two different probe types (Infinium I and II) was done by peak-based correction [324] and the beta-mixture quintile normalisation method [325]. Batch-effects were reduced using the empirical Bayes method ‘ComBat’ [326] and missing data was imputed using the k-nearest neighbour averaging method (R function ‘impute.knn’ in package ‘impute’) [327]. Probes overlapping SNPs (n=65) or with low detection *P*-value in more than 20% of the samples (n=36,655) were excluded from the analysis resulting in 448,857 probes. Pre-processing to this stage was performed by Karin van Veldhoven [197]. At this stage the data was received, and furthermore cross-hybridising probes were removed using the R function ‘rmSNPandCH’ in the package ‘DMRcate’, leaving 404,596 probes for the analysis [328]. Methylation levels from the HM450K array are expressed as beta-values between 0 and 1 representing the fraction of methylated cytosine at the given probe.

2.2.4.3 Statistical analysis

An EWAS with ELEE as the exposure and DNA methylation as the outcome was conducted in 216 samples from the EPIC-Italy HM450K data measured in blood. Beta regression on beta-values was used to investigate DNA methylation levels at individual CpG sites and their association with ELEE; DNA methylation levels were modelled as a dependent variable in a generalised linear model with beta-distributed responses (R function 'vglm' in package 'VGAM' [309]). Confounders known to influence methylation and/or oestrogen levels were adjusted for including age, BMI, alcohol consumption (average gram per day) and smoking duration in years (all variables were reported at the time of recruitment), as well as technical confounders including batch, position on batch and WBC composition estimated using the Houseman method [329]. The *P*-values were corrected for multiple testing using the false discovery rate (FDR) *Q*-values in the R function 'p.adjust'. For interpretation, linear mixed-effects regression modelling on beta-values multiplied by 100 was used, to get DNA methylation changes per one year longer ELEE as estimates. Manifest and annotation data were extracted from the R package 'IlluminaHumanMethylation450kanno.ilmn12.hg19' and gene function, disease association, and interaction with oestradiol were extracted from the UCSC genome browser (<https://genome-euro.ucsc.edu>) and the GeneCards Human Gene Database (<https://www.genecards.org>).

2.2.4.4 Inflation and bias

Inflation and bias were explored using the R package 'bacon' [318] and Quantile-Quantile (QQ) plots of the *P*-values in the R package 'qqman'. The 'bacon' package uses a Bayesian method to estimate the empirical null distribution from observed test statistics, and fits a three component normal mixture to the test statistics using a Gibbs sampling algorithm, which estimates the bias and inflation factor lambda. Lambda equal to 1 indicates no inflation, and greater than 1 indicates inflated *P*-values. Inflation and bias can affect test statistics in EWAS, and to explore this in the EWAS of ELEE two more beta regression models were investigated: one without adjusting for WBC composition and another one not adjusted for any of the confounding variables. Inflation and bias were compared between the three models.

2.2.4.5 Sensitivity analyses

Sensitivity analyses were conducted for the following conditions: (i) exploring the effects of stratifying on age at diagnosis ≥ 50 to enrich for ER-positive disease by including all subjects with complete information of ELEE in the analysis ($n=237$) and (ii) effects from including prospective breast cancer cases in the analysis by repeating the analysis in controls only ($n=119$). Estimates were compared using the Pearson's correlation coefficient.

2.2.4.6 Functional genomic regions and histone modifications

Probes were classified into categories reflecting their location in relation to functional genomic regions (promoter, gene body and intergenic region) and CpG islands (CpG island, shore, shelf and open sea) using annotation data for the HM450K array. Promoter regions were defined by combining CpG probes overlapping TSS200 (located 0 to 200bp from a transcription start site), 5'UTR and 1st exon, and body and 3'UTR were combined for gene body. Furthermore, the probes were classified into categories overlapping chromatin immunoprecipitation sequencing (ChIP-seq) data of different histone modifications. The ChIP-seq data (bed-files with pre-processed narrow, stable peaks) of PBMNCs was downloaded from the ENCODE project (<https://www.encodeproject.org>). Active promoters were defined as promoter region overlapping histone modifications H3K4me3 or H3K9ac, repressed promoters were defined as promoter regions overlapping H3K9me3 or H3K27me3, and active enhancers were defined as intergenic region overlapping H3K27ac or H3K4me1 [180]. ORs and P -values from Fisher's Exact Test for count data were estimated using the R function 'oddsratios' from the package 'epitools'.

2.2.5 EWAS of hormonal risk factors

Four other EWAS of the hormonal risk factors included in the ELEE model were conducted using a beta regression model with the risk factor as the exposure and DNA methylation as the outcome, adjusting for the same confounders as in the EWAS of ELEE. The P -values were corrected for multiple testing and significant probes ($\text{FDR } Q < 0.05$) from the five different EWAS were overlapped.

2.2.6 Genome-wide DNA methylation and ELEE

Genome-wide hypomethylation in association with breast cancer risk has previously been observed in this dataset [197]. Therefore, it was explored if genome-wide DNA methylation levels were associated

with ELEE. The genome-wide DNA methylation was calculated as the average beta-values across the 404,596 CpG sites on the HM450K array for all women with complete information of ELEE in EPIC-Italy (n=216). The association between the genome-wide DNA methylation and ELEE was analysed using generalised linear regression adjusting for age, BMI, alcohol consumption (average gram per day), smoking duration in years (all variables were reported at the time of recruitment), batch, position on batch and WBC composition.

2.3. RESULTS

2.3.1 Hormonal factors are associated with breast cancer risk in EPIC-Italy

Different hormonal factors and the association with breast cancer risk were explored in the EPIC-Italy questionnaire data (**Table 2.1**). At baseline, cases were older compared to controls (53.2 and 50.6 years) and more cases had reached menopause (58.9 and 49.4%) with a slightly older age at menopause (48.9 and 48.3 years). Controls had a slightly higher number of pregnancies compared with cases (1.9 and 1.8), more OC users (40.5% and 36.6%), less HRT users (15.5 and 18.6%), and lower BMI (25.8 and 26.3 kg/m²). The results are found in **Appendix 1** and summarised in **Figure 2.2**.

2.3.1.1 Age at menarche and menopause

An older age at menarche was inversely associated with the breast cancer risk for all and postmenopausal women (HR=0.94, 95% CI: 0.91 – 0.98, $P=0.003$ and HR=0.94, 95% CI: 0.89 – 0.98, $P=0.008$), and older age at menopause was positively associated with an increase in risk (HR=1.02, 95% CI: 1.01 – 1.04, $P=0.006$).

2.3.1.2 Parity-related factors

Parity (parous compared to nulliparous women) had no significant risk effect in EPIC-Italy, however, each additional pregnancy was inversely associated with the risk for all (HR=0.92, 95% CI: 0.87 – 0.97, $P=0.003$) and premenopausal women (HR=0.86, 95% CI: 0.78 – 0.94, $P=0.002$). A trend towards a protective association between number of pregnancies and breast cancer risk was observed for postmenopausal women (HR=0.94, 95% CI: 0.88 – 1.01, $P=0.070$). An older age at first full-term pregnancy was associated with increase in breast cancer risk for all (HR=1.02, 95% CI: 1.00 – 1.03, $P=0.008$) and premenopausal women (HR=1.03, 95% CI: 1.01 – 1.05, $P=0.006$). Ever breastfeeding was not associated with breast cancer risk, however, longer total lifetime breastfeeding duration was inversely associated the risk in all women (HR=0.88, 95% CI: 0.80 – 0.97, $P=0.008$).

2.3.1.3 Lifestyle factors

Of the lifestyle factors, an increase in breast cancer risk with increasing BMI was observed for all and postmenopausal women (HR=1.01, 95% CI: 1.00 – 1.03, $P=0.035$ and HR=1.02, 95% CI: 1.01 – 1.04, $P=0.010$). However, BMI was not associated with breast cancer risk in premenopausal women. Smokers compared to never smokers had a lower risk among postmenopausal women (HR=0.81, 95%

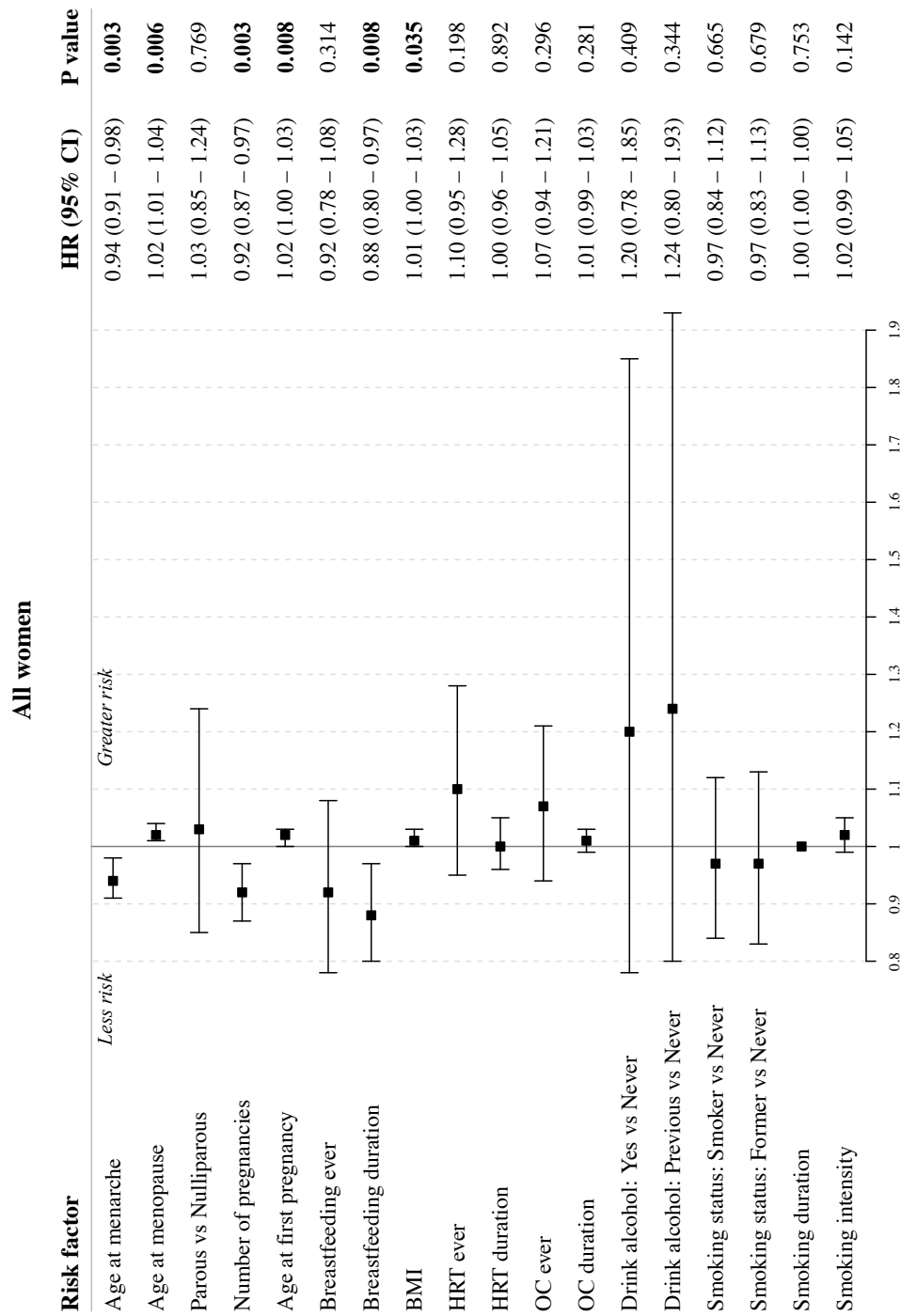


Figure 2.2A. Hormonal breast cancer risk factors in EPIC-Italy

The associations between hormonal risk factor and breast cancer risk were estimated using the Cox proportional hazards regression modelling adjusted for age at recruitment. The square box represents the Hazard Ratios (HRs) and the lines the 95% Confidence Intervals (CIs). All women includes 30,671 controls and 1,193 cases.

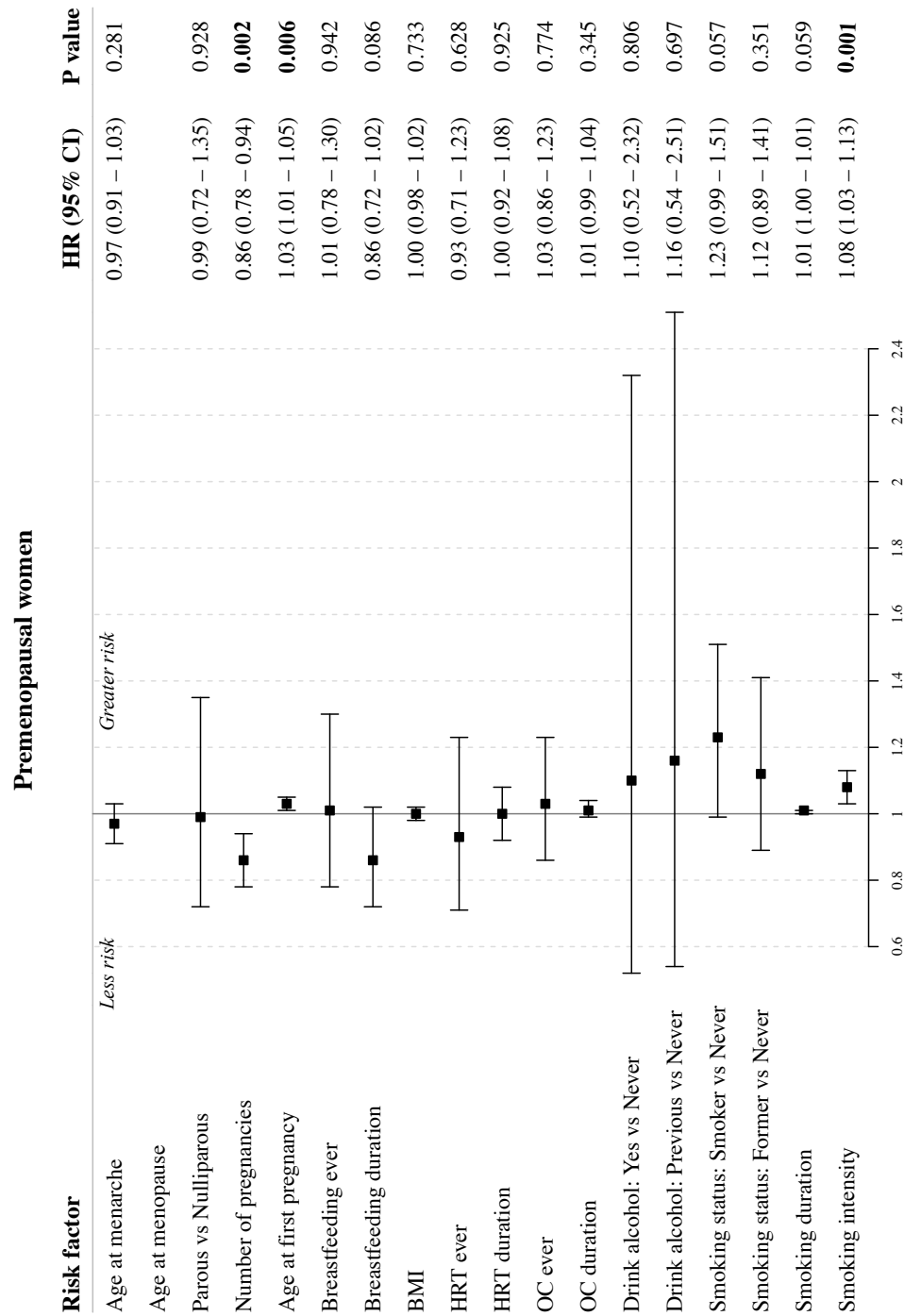


Figure 2.2B. Hormonal breast cancer risk factors in EPIC-Italy

The associations between hormonal risk factor and breast cancer risk were estimated using the Cox proportional hazards regression modelling adjusted for age at recruitment. The square box represents the Hazard Ratios (HRs) and the lines the 95% Confidence Intervals (CIs). Premenopausal women included 15,086 controls and 477 cases

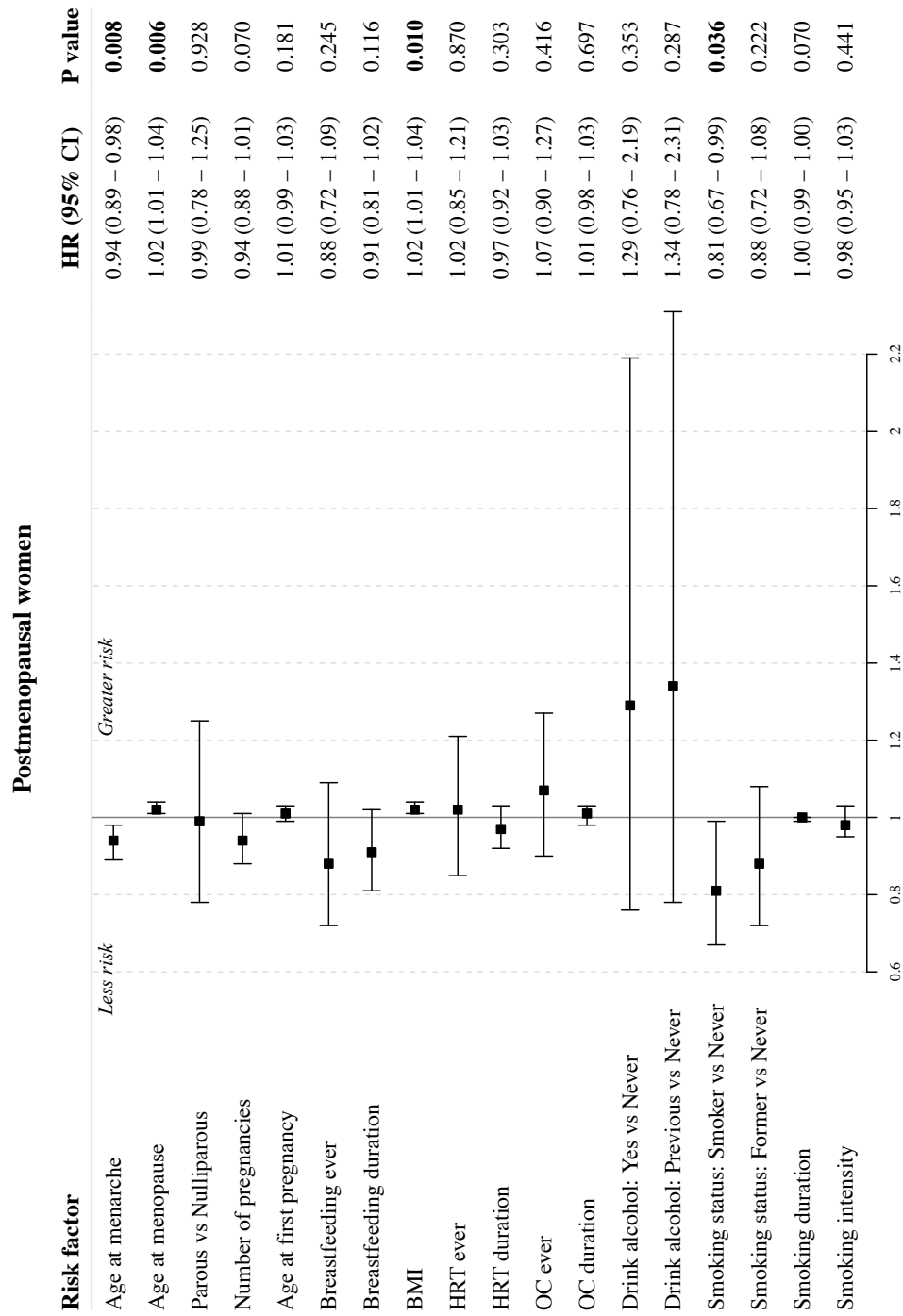


Figure 2.2C. Hormonal breast cancer risk factors in EPIC-Italy

The associations between hormonal risk factor and breast cancer risk were estimated using the cox proportional hazards regression modelling adjusted for age at recruitment. The square box represents the Hazard Ratios (HRs) and the lines the 95% Confidence Intervals (CIs). Postmenopausal women included 15,163 controls and 703 cases.

CI: 0.67 – 0.99, $P=0.036$) but a higher smoking intensity was associated with an increase in the risk for premenopausal women (HR=1.08, 95% CI: 1.03 – 1.13, $P=8.09 \times 10^{-4}$). Neither reported use of OC or HRT was associated with breast cancer risk in EPIC-Italy, nor smoking duration or alcohol drinking. To note, less than 2% of the women reported never drinking.

2.3.2 The ELEE model is significantly associated with breast cancer risk

The final ELEE model was calculated as a woman's reproductive time (age at menopause or age at recruitment for premenopausal women minus age at menarche) minus one year per pregnancy and total duration breastfeeding in years up to recruitment. This will in theory give the time in years when a woman has been exposed to high levels of endogenous oestrogens, but also takes into account the long-term effect on oestrogen levels for parous women and the duration breastfeeding, which will shorten the total lifetime endogenous oestrogen exposure. Different combinations of time-dependent risk factors were explored for the ELEE model, summarised in **Table 2.3**. There was no apparent difference in using 9 months or one year per pregnancy; therefore one year for each pregnancy was used in the model since it also takes into account a delay in ovulation after pregnancy. The duration of OC and HRT use were found to not be significantly associated with breast cancer risk in EPIC-Italy (**Figure 2.2**) and the addition of this to the ELEE calculation did not improve the model. Also, only current or recent OC and HRT use has been reported to increase the risk and oestrogen levels, and it might not be relevant to include duration use. In all women in EPIC-Italy ($n=1,193$ cases and 30,671 controls, **Table 2.1**) one year longer ELEE, ranging from 5 to 49, was associated with a 5% increase in breast cancer risk (HR=1.05, 95% CI: 1.04 – 1.07, $P=3.14 \times 10^{-12}$, **Table 2.3**), and was associated with breast cancer risk in both pre- and post-menopausal women (HR=1.09, 95% CI: 1.03 – 1.15, $P=0.002$ and HR=1.03, 95% CI: 1.01 – 1.05, $P=6.07 \times 10^{-4}$). By removing outliers in the ELEE range (women with ELEE below and above two standard deviations from mean value) very similar results were obtained (HR=1.06, 95% CI: 1.04 – 1.08, $P=7.52 \times 10^{-12}$ for all women; HR=1.09, 95% CI: 1.03 – 1.15, $P=0.003$ for premenopausal women; HR=1.04, 95% CI: 1.02 – 1.06, $P=1.63 \times 10^{-4}$ for postmenopausal women).

2.3.3 EWAS identifies 694 probes associated with the ELEE

An EWAS of the ELEE was conducted in EPIC-Italy ($n=216$, **Table 2.2**) using a beta regression model adjusting for confounding biological and technical variables. After correction for multiple testing the EWAS identified 694 CpG probes significantly associated with ELEE in EPIC-Italy (FDR $Q < 0.05$

Table 2.3. ELEE models and breast cancer risk in EPIC-Italy
Different time-dependent risk factors were included in the ELEE models and tested for their association with breast cancer risk (using Cox proportional hazards regression adjusted for age at recruitment) in all women (n=30,671 controls and 1,193 cases), premenopausal women (15,086 controls and 477 cases) and postmenopausal women (n=15,163 controls and 703 cases) in EPIC-Italy. The final ELEE-model used throughout this project is highlighted in grey.

ELEE model	All women		Premenopausal women		Postmenopausal women	
	HR (95% CI)	P value	HR (95% CI)	P value	HR (95% CI)	P value
Reproductive time	1.05 (1.04-1.07)	8x10⁻¹³	1.04 (0.97-1.1)	0.281	1.03 (1.01-1.05)	4x10⁻⁴
Reproductive time - number of pregnancies*0.75	1.05 (1.04-1.07)	2x10⁻¹³	1.08 (1.01-1.14)	0.015	1.03 (1.01-1.05)	4x10⁻⁴
Reproductive time - number of pregnancies*1	1.05 (1.04-1.07)	1x10⁻¹³	1.08 (1.02-1.14)	0.007	1.03 (1.01-1.05)	3x10⁻⁴
Reproductive time - number of pregnancies*1 - breastfeeding duration	1.05 (1.04-1.07)	3x10⁻¹²	1.09 (1.03-1.15)	0.002	1.03 (1.01-1.05)	6x10⁻⁴
Reproductive time - number of pregnancies*1 - breastfeeding duration + HRT duration	1.05 (1.03-1.06)	6x10⁻¹²	1.06 (1.02-1.10)	0.008	1.03 (1.01-1.04)	1x10⁻³
Reproductive time - number of pregnancies*1 - breastfeeding duration + OC duration	1.03 (1.02-1.05)	2x10⁻⁹	1.02 (0.99-1.04)	0.142	1.02 (1.01-1.04)	1x10⁻³

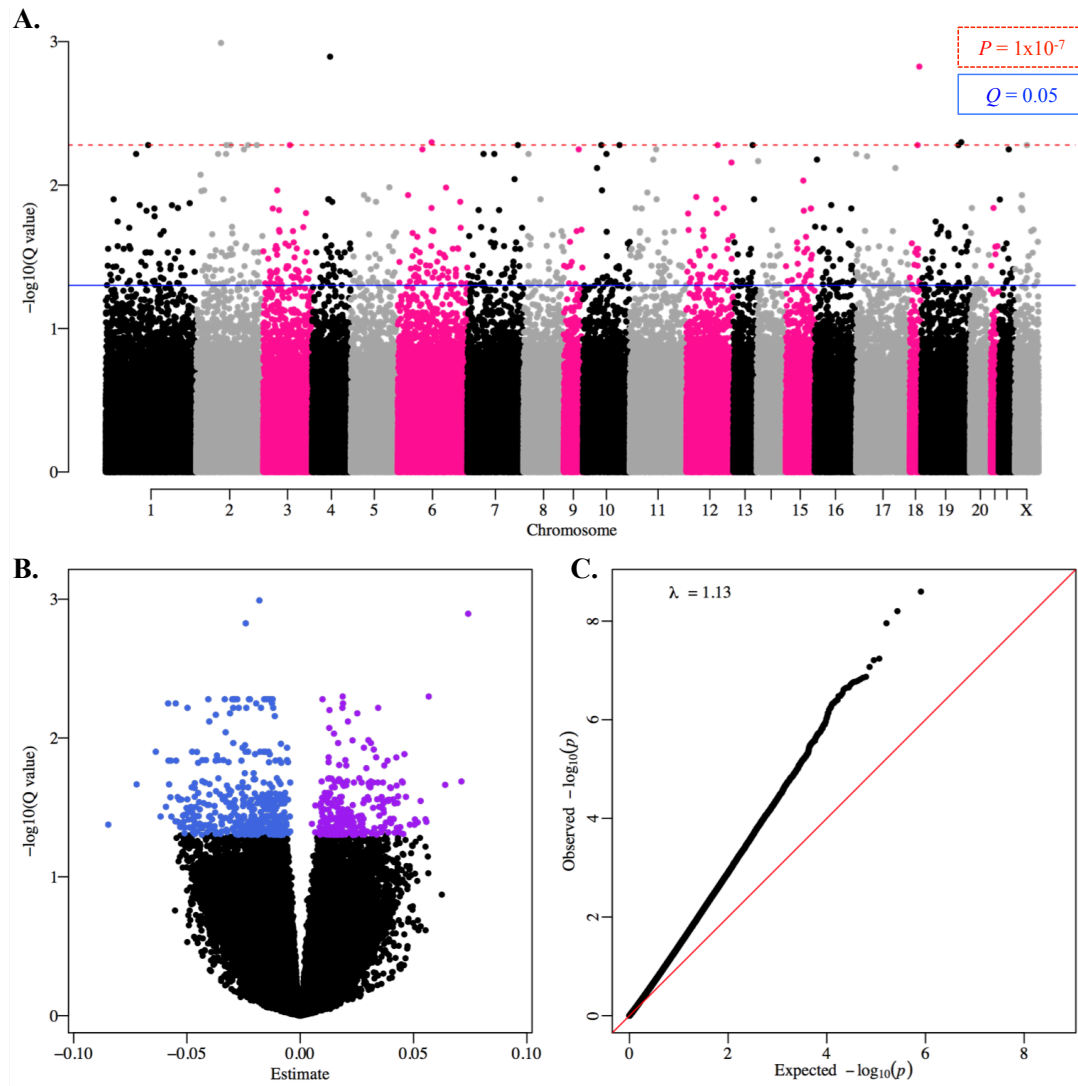


Figure 2.3. EWAS identifies CpG sites significantly associated with ELEE in EPIC-Italy

An EWAS with DNA methylation (HM450K beta-values) as outcome and ELEE as exposure was conducted in EPIC-Italy (**Table 2.2**, $n=216$) using a beta regression model adjusted for age, BMI, alcohol consumption, and smoking, batch, position on batch and WBC composition using the Houseman method. **A.** Manhattan plot of minus $\log_{10} Q$ -values (FDR-corrected P -values) from the EWAS of all 404,596 probes across the genome. Blue line indicates FDR Q -value threshold of 0.05 and red dotted line represents genome-wide threshold P -value of 1×10^{-7} . **B.** Volcano plot of the regression coefficients (estimates) from the beta regression model, showing significant hypomethylated ($n=424$, in blue) and hypermethylated probes ($n=270$, in purple). **C.** The expected and observed P -values from the EWAS of ELEE. Inflation factor lambda was calculated using the R package ‘bacon’.

Figure 2.3A) with a mixture of hypo- and hypermethylated probes with increasing ELEE, but mostly hypomethylated (61%, **Figure 2.3B**). The *P*-values were moderately inflated with a lambda of 1.13 (**Figure 2.3C**).

2.3.3.1 Inflation and bias in the EWAS of ELEE

Inflation and bias of the EWAS of ELEE were investigated by repeating the analysis without adjusting for WBC composition and with no adjustments. The test statistics (*Z* values) and *P*-values for the three models are visualised in **Figure 2.4**. The bias and inflation factor for the main analysis was -0.12 and 1.13 respectively suggesting that there were some bias and inflation presented in the fully adjusted model. The unadjusted models showed possibly larger bias (-0.30), suggesting that cellular heterogeneity and not adjusting for WBC composition might result in biased results. The QQ-plots and lambda inflation factors were smaller than in the fully adjusted model (1.11 and 1.02), suggesting a larger inflation for the fully adjusted model (**Figure 2.4D-F**). However, these statistical tools have not yet been explored for beta regression and these conclusions should be taken with caution. No significant CpG sites were observed for the unadjusted model (FDR *Q*-value < 0.05) and 242 significant CpG sites in the model without adjusting for WBC.

2.3.3.2 Sensitivity analyses

To explore the effects of (i) excluding cases with an age at diagnosis < 50 and (ii) including prospective cases in the EWAS of ELEE sensitivity analyses were conducted including all cases and controls (n=237) and in controls only (n=119). Estimates in the sensitivity analyses were highly correlated with the estimates from the main EWAS of ELEE for both analyses (*r*=0.997 and *r*=0.963, respectively, **Figure 2.5A-B**) and all 694 probes remained significant (FDR *Q* < 0.05, n=237) in the first analysis and 563 (81% of the 694 probes) in the second analysis.

2.3.3.3 CpG sites associated with ELEE are enriched in CpG islands, CpG shelves and active promoters

The 694 significant CpG probes were categorised by their relation to different functional genomic regions, relation to CpG islands, and overlap with histone modifications, and analysed for enrichment. The CpG sites associated with ELEE were significantly (*P*-value < 0.05) enriched in CpG islands, CpG shelves, and active promoters, and depleted in intergenic regions, open sea, other promoters (defined by all promoters not overlapping any of the above mentioned histone modifications) and active

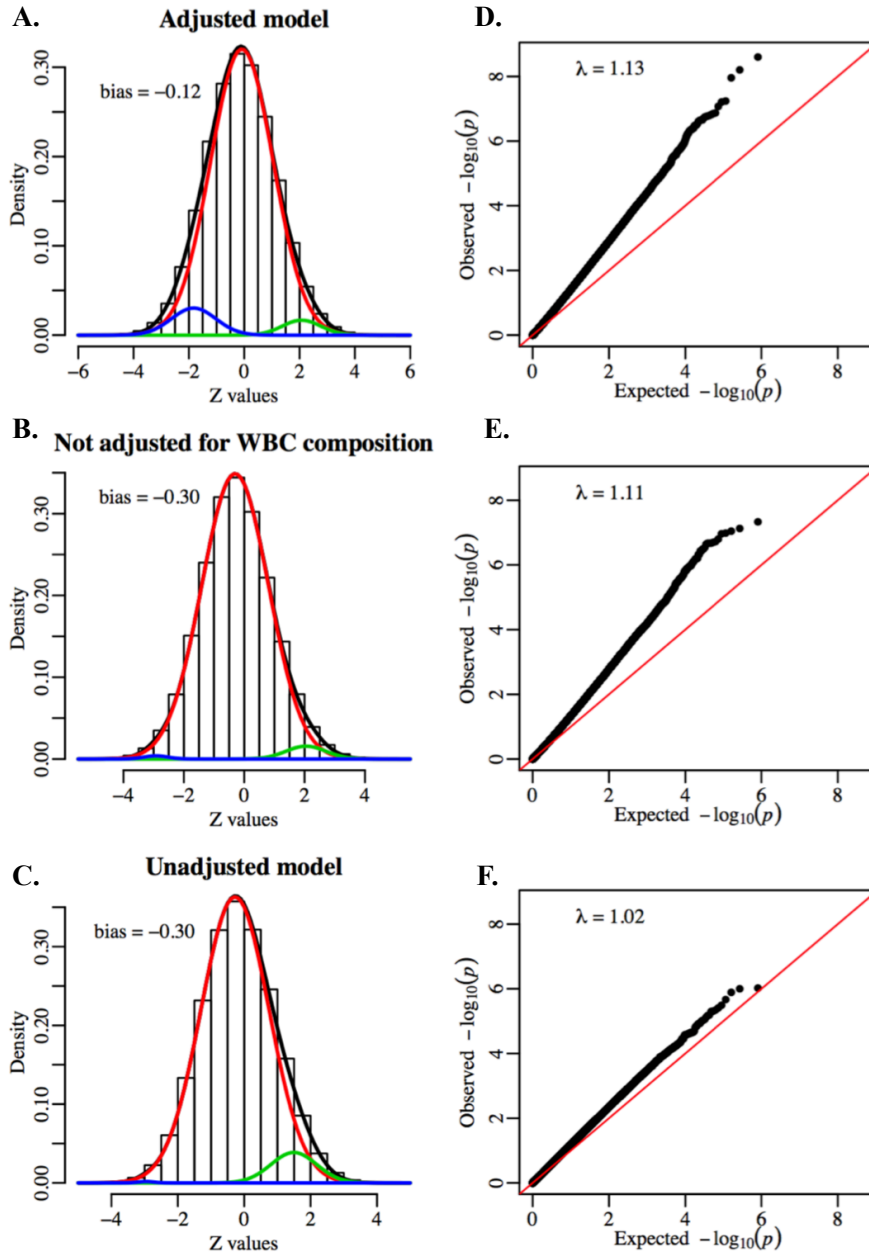


Figure 2.4. Inflation and bias in the EWAS of ELEE

A-C. The distribution of test statistics (Z values) from the EWAS of ELEE and the three components from the Bayesian method used in the R package ‘bacon’ for (A) the fully adjusted model (age at recruitment, BMI, smoking duration, alcohol consumption, batch, position on batch, WBC composition) (B) without adjusting for WBC composition and (C) without any adjustments. The black line is the fit for the data, the red line is the fit of the null component (the empirical null distribution) and the blue and green lines for the fits of the negatively and positively true associations. **D-F.** QQ-plots of P -values from the beta regression models and inflation factor lambda, calculated in R package ‘bacon’.

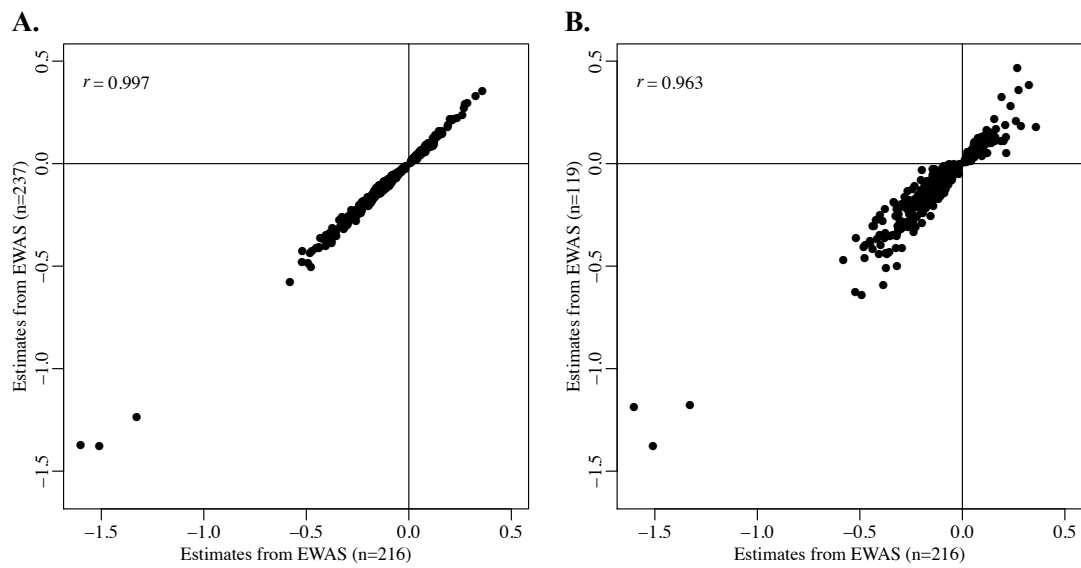


Figure 2.5. Sensitivity analyses for the EWAS of ELEE

The main EWAS of ELEE included 216 women from EPIC-Italy HM450K data (**Table 2.2**) where cases with an age at diagnosis < 50 were excluded. To explore the effects of excluding younger cases the analysis was repeated on the 694 significant probes including all cases and controls ($n=237$), and to explore the effects of including prospective cases in the EWAS the analysis was repeated in controls only ($n=119$). **A.** Comparison of the estimates in the main EWAS ($n=216$) and including all cases and controls ($n=237$) **B.** Comparison of the estimates in the main EWAS ($n=216$) and in controls only ($n=119$). After correction for multiple testing ($FDR Q < 0.05$), all probes remained significant in the first sensitivity analysis ($n=237$) and 563 (81%) in the second sensitivity analysis ($n=119$). Correlation coefficient (r) was calculated using Pearson's correlation coefficient.

enhancers (**Figure 2.6A-C**). With longer ELEE the methylation levels were mostly increasing for promoter regions, CpG islands and active or repressed promoters. A decrease in DNA methylation was observed in gene bodies, intergenic regions, CpG shelves, open sea, and other promoters (**Figure 2.6D-F**).

2.3.3.4 Top hits from the EWAS of ELEE

The significant CpG sites were further filtered on magnitude of change in DNA methylation (regression coefficient from the mixed-effects linear model, denoted as estimate) and variation in DNA methylation levels across individuals (standard deviation of beta-values for each probe). This was to exclude possible false positives or hits with very small changes in DNA methylation that might be biologically irrelevant or difficult to replicate in further validation studies. After this step a total of 212 probes with an absolute estimate above its mean value (0.12%) and with a standard deviation over its mean value (2.39%) were left (**Appendix 2**). Out of these, the top 20 probes with largest estimates (absolute estimate > 0.40%) and the top 20 probes with most significant *P*-values (*P*-values < 1.5×10^{-6}) are presented in **Table 2.4** together with annotation data and gene function for the nearest gene. Three CpG sites were selected from both criteria (cg03785646, cg05422360, and cg13645651) and two CpG sites were located in the pseudogene *HLA-J* and two CpG sites in the gene *MC4R*. For CpG sites located far away from the nearest gene (>10K base pairs), gene function was not reported, leaving 31 genes in the table. At least a third of the genes were associated with cancer or cancer-related functions and three genes have previously been associated with breast cancer (*CHAC1*, *FAT1*, *GALNT9*). The genes *FAT1*, *HLA-J*, *IL6ST*, *SLC1A2* and *TCL1B* interact with oestradiol, and *UNC45A* binds the progesterone receptor. Interestingly, the gene *SMARCC1* regulates transcription by altering the chromatin structure around those genes and has been associated with various cancers.

2.3.4 There are few overlaps between the EWAS of hormonal risk factors

To explore if the combined ELEE model and the individual hormonal risk factors in the model had any common associations with DNA methylation four different EWAS were conducted: for age at menarche, age at menopause, number of pregnancies and breastfeeding duration (in years), with the same adjustments as mentioned above in EPIC-Italy (**Table 2.2**). The number of significant CpG probes (FDR $Q < 0.05$) were less in all four analyses compared to the EWAS of ELEE; 25 probes for age at menarche, 393 for age at menopause, 18 for number of pregnancies and 69 for breastfeeding

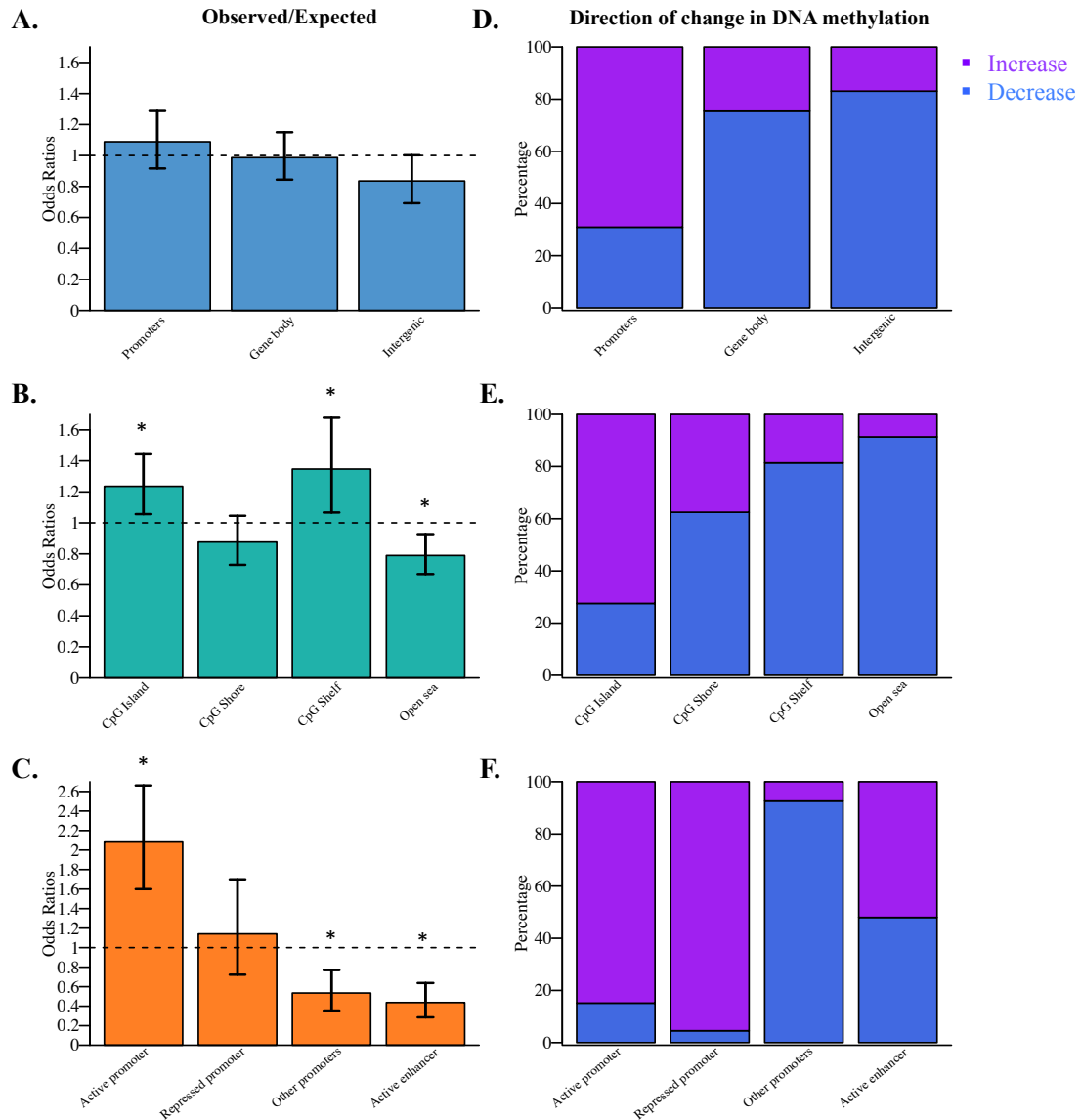


Figure 2.6. Functional genomic regions and histone modifications among EWAS of ELEE hits

Enrichment analysis for the 694 significant CpG probes was conducted by overlapping the CpG sites with annotation data for functional genomic regions, CpG islands, and histone modifications using ChIP-seq data in PBMCs from the ENCODE project. **A-C.** Bar plots of the ORs and 95% CIs, OR > 1 indicates more observed CpG sites than expected in that region. ORs, 95% CIs and *P*-values (Fisher's Exact Test for count data) were estimated using the R function 'oddsratios' from the package 'epitools', a * indicates *P* < 0.05. **D-F.** Bar plots of the direction of change (purple for decrease and blue for increase) in DNA methylation levels with increasing ELEE; the percentage of increased and decreased CpG sites (among the 694 CpG sites associated with ELEE) in each region is shown.

Table 2.4. Top hits from the EWAS of ELEE

The table represents the top CpG sites based on magnitude of change in DNA methylation per year increase in ELEE (estimate > 0.40%) and statistical significance (P -values < 1.5×10^{-6}). Probes in the table are sorted in genetic alphabetic order.

HM450K probe	Nearest gene	Gene function**	UCSC RefGene Group	Selection criteria
cg26252045	ABTB2	Protein heterodimerisation activity and hepatocyte growth	Body	$P=1.36 \times 10^{-6}$
cg03785646	AGAP1	GTPase activity, membrane trafficking, cytoskeleton dynamics. Associated with autism spectrum disorder, schizophrenia and cerebral palsy	Body	Estimate=-0.52, $P=1.62 \times 10^{-7}$
cg01154254	ANKRD53	Regulates chromosomes in mitosis	TSS1500	$P=4.66 \times 10^{-7}$
cg08601628	C8orf74	Uncharacterised protein	Body	$P=4.51 \times 10^{-7}$
cg21153102	CHAC1	Inhibits Notch signalling, regulation of glutathione levels and oxidative balance in the cell. Associated with breast and ovarian cancer recurrence	Intergenic	Estimate=-0.43%
cg07630274	COL13A1	Tissue connectivity. Associate with muscle weakness and non-alcoholic fatty liver disease	Body	$P=2.22 \times 10^{-7}$
cg06968859	CTNNA2	Cell-cell adhesion regulation, tumour suppressor. Associated with bipolar disorder, diabetes type 2, mixed germ cell and laryngeal cancer	Body	$P=2.52 \times 10^{-9}$
cg05566582	DNAJB6*	NA	Intergenic	Estimate=-0.42%
cg09438069	FAT1	Tumour suppressor gene, cell communication. Associated with various cancer types (pancreatic, colon, breast, glioma, leukaemia etc.). Interacts with oestradiol.	Body	Estimate=0.44%
cg03143347	GALNT9	Glycosylation in the Golgi apparatus, expressed in the brain, oncogene. Associated with neuroblastoma, BMI, ovarian and breast cancer	Body	$P=6.36 \times 10^{-7}$
cg24838316	HLA-J	Transcribed pseudogene derived from a histocompatibility complex that plays a central role in the immune system. Interacts with oestradiol.	Intergenic	Estimate=-1.51%
cg23681866			Intergenic	Estimate=-1.33%
cg03495030	HNF1A	Transcription factor required for the expression of several liver-specific genes. Associated with maturity-onset diabetes of the young type 3, diabetes, hepatic adenomas, and pancreatic and lung cancer.	Body	Estimate=-0.41%
cg22968966	HS3ST2*	NA	Intergenic	Estimate=-0.40%
cg21950518	IL6ST	Interleukin 6 signal transducer functions as a part of the cytokine receptor complex. Associated with periodontitis, Kaposi's sarcoma, hepatocellular adenoma. Interacts with oestradiol.	1stExon	$P=1.50 \times 10^{-6}$
cg05422360	MAGEE1	Melanoma Antigen Gene 1, may play a role in brain, muscle, and peripheral nerve. Associated with melanoma and cognitive disability.	1stExon	Estimate=-0.44%, $P=1.72 \times 10^{-7}$
cg13645651	MC4R	Encodes a membrane-bound receptor and member of the melanocortin receptor family. Associated with autosomal dominant obesity.	TSS1500	Estimate=-0.48%, $P=1.41 \times 10^{-7}$
cg18233497			1stExon	Estimate=-0.44%
cg24036523	PAPLN	Associated with peptidase activity and serine-type endopeptidase inhibitor activity	Body	Estimate=-0.45%

cg00443797	PLPP4	Involved in glycerolipid metabolism and innate immune system. Associated with lung cancer.	3'UTR	$P=2.47 \times 10^{-7}$
cg22758104	RPH3AL*	NA	Intergenic	$P=4.62 \times 10^{-7}$
cg19904058	SCART1	Non coding pseudo gene, may play a role in the immune system	Body	Estimate=-0.52%
cg16733643	SCMH1	DNA binding transcription factor activity, associated with body height	Body	Estimate=0.42%
cg13971030	SLC1A2	Glutamate transporter proteins involved in the central nervous system. Associated with neurological disorders, schizophrenia and bipolar disorder. Interacts with oestradiol.	Body	Estimate=-0.49%
cg03723715	SLC8A2	Mediates calcium and sodium levels in the cytoplasm. Associated with glioma and glioblastoma.	Body	$P=1.35 \times 10^{-7}$
cg17567562	SMARCC1	Regulate transcription by altering the chromatin structure around those genes. Associated with various cancers (prostate, colorectal, oral squamous cell carcinoma)	Body	Estimate=-0.48%
cg10298859	SPACA7*	NA	Intergenic	$P=2.26 \times 10^{-7}$
cg15472797	SPRN	Prion-like protein that may modulate prion proteins. Associated with prion disease including Creutzfeldt-Jakob, paranoid personality disorder	Intergenic	Estimate=-0.40%
cg21224380	SSH1	Regulates actin filament dynamics. Associated with polycystic kidney disease 3, porokeratosis, gastric cancer	Body	$P=1.81 \times 10^{-7}$
cg25031061	STK40	Serine/Threonine Kinase 40, may be a negative regulator of NF-kappa-B and p53-mediated gene transcription	Body	$P=4.24 \times 10^{-7}$
cg26888941	SYNGAP1	Negatively regulates Ras, Rap and other receptors that regulate synaptic plasticity and neuronal homeostasis. Associated with intellectual disability and autism spectrum disorder etc.	Body	$P=5.77 \times 10^{-8}$
cg05402891	TCL1B	Enhances the phosphorylation and activation of AKT1 and AKT2. Associated with T-cell leukaemia and lymphoma. Interacts with oestradiol.	TSS200	Estimate=-0.47%
cg25936380	TMEM185B	Transmembrane protein	TSS1500	$P=2.22 \times 10^{-7}$
cg04078118	TNRC6B	RNA-mediated gene silencing by microRNAs and short interfering RNAs. Associated with uterine/prostatic/reproductive organ benign neoplasm, prostate cancer.	TSS1500	$P=3.11 \times 10^{-7}$
cg12608727	UBE2U	Ubiquitin protein ligase activity	Intergenic	$P=1.69 \times 10^{-7}$
cg08551047	UNC45A	Cell proliferation, binds progesterone receptor and regulates progesterone receptor chaperoning pathway	5'UTR	Estimate=-0.58%
cg20858107	XKR6	Kell blood group complex subunit-related family. Associated with autoimmune diseases (such as systemic lupus erythematosus).	Body	Estimate=-0.45%

* CpG site located far away from closest gene (>10K bp), treated as not gene-associated

** Gene function information from UCSC genome browser (<https://genome-euro.ucsc.edu>) and GeneCards Human Gene Database website (<https://www.genecards.org>)

duration. The overlaps between the five EWAS are presented in **Figure 2.7**. Between ELEE and age at menopause 31 probes were overlapping, three probes were shared between ELEE and breastfeeding duration, and one between the number of pregnancies and breastfeeding duration. Out of the 31 CpG sites overlapping the EWAS of ELEE and age at menopause, nine CpG sites were left after filtering on magnitude of change in DNA methylation and standard deviation. Out of these the CpG sites cg25031061 (in gene body of *STK40*) and cg19904058 (in gene body of *SCART1*) can be found in **Table 2.4** with the top hits from the EWAS of ELEE.

2.3.5 Genome-wide DNA methylation and ELEE

Genome-wide DNA methylation levels varied from 51.89% to 54.21% across all subjects in EPIC-Italy with complete information of ELEE (n=216). When stratifying all women in two groups split on the median (**Figure 2.8A**), women with longer ELEE had significantly lower genome-wide DNA methylation compared to women with shorter ELEE (52.96% and 53.05% respectively, $P=0.009$). The analysis was repeated in controls only to exclude any confounding reverse causation of DNA methylation and cancer diagnosis (**Figure 2.8B**). No significant difference was observed between the groups in controls only (n=119, 53.07% and 53.10% respectively, $P=0.940$).

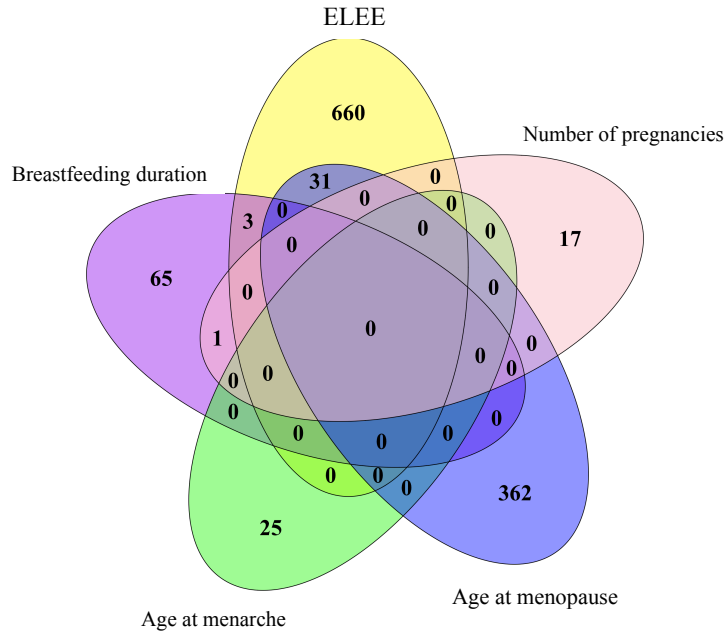


Figure 2.7. EWAS of hormonal risk factors

Venn diagrams of significant (FDR $Q < 0.05$) CpG probes from the EWAS of the hormonal risk factors included in the ELEE model (age at menarche, age at menopause, number of pregnancies and breastfeeding duration) compared with the 694 significant CpG probes identified with the full ELEE model.

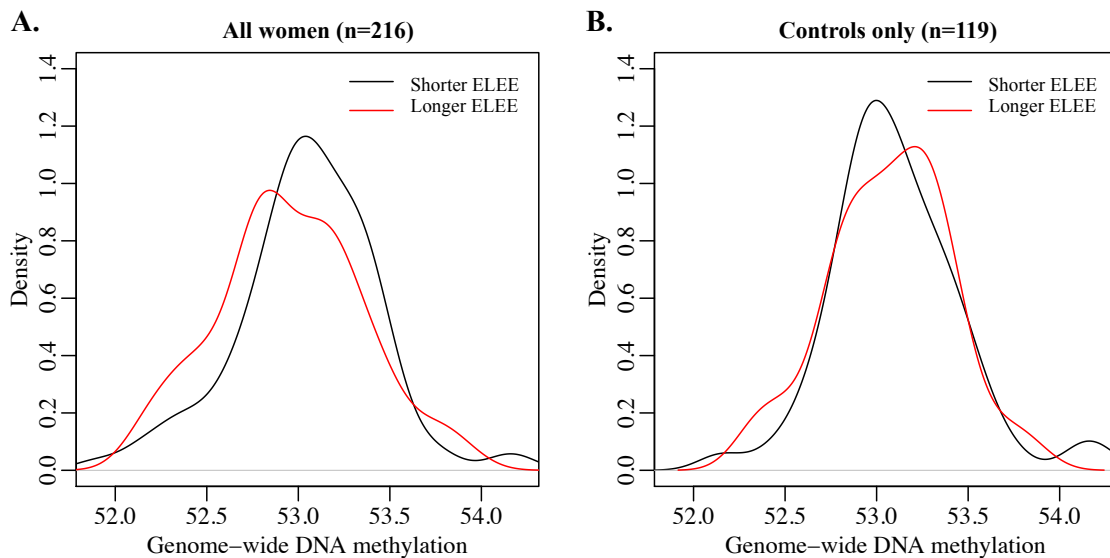


Figure 2.8. Genome-wide DNA methylation and ELEE

Genome-wide DNA methylation was calculated as the average beta-values across the HM450K array in all women (A) and in controls only (B) in two groups split by median ELEE value. For all women there was a significant decrease in genome-wide DNA methylation levels in the group with longer ELEE compared to women in the group with shorter ELEE (52.96% and 53.05% respectively, $P=0.009$). In controls only there was no significant difference in genome-wide levels between the groups (53.07% and 53.10% respectively, $P=0.940$).

2.4. DISCUSSION

In this chapter the association between lifetime endogenous oestrogen exposure and DNA methylation measured in blood has been explored. An ELEE model was defined that showed a 5% increase in breast cancer risk per one year increase in ELEE, and an EWAS of ELEE was conducted using HM450K data in EPIC-Italy. DNA methylation levels at a total of 694 CpG sites were significantly associated with ELEE and enriched in CpG islands and active promoters regions where often an increase in DNA methylation was observed. The results suggest that DNA methylation is altered by the lifetime exposure to oestrogens.

2.4.1 Hormonal risk factors and breast cancer

The associations between known hormonal risk factors and breast cancer were explored in more than 30,000 women in the EPIC-Italy study cohort (**Table 2.1**). Most of the results are consistent with previous findings (**Table 1.1**), but there are a few deviations. No significant associations between nulliparous women and parous women, or women who have been breastfeeding compared to parous women who have not were observed in EPIC-Italy. Previously it has been estimated as a RR of 1.30 for nulliparous women compared to parous women, and RR from 0.61 to 0.97 for women who have breastfed compared to never breastfed [100, 117]. This apparent difference might be a power issue in our dataset due to the low number of nulliparous women (10%) or few women who had never breastfed (18%). A decrease in breast cancer risk was seen for both number of pregnancies and duration of breastfeeding.

Increased BMI was not associated with an increase in breast cancer risk in premenopausal women. It is clear that being overweight, obese, or gaining weight, is one of the biggest risk factors for postmenopausal women since this will increase the aromatase expression and thereby the oestrogen synthesis in the adipose tissue, which is the main source of oestrogen after menopause [42, 50]. There is evidence of an inverse relationship between BMI and breast cancer risk in premenopausal women [43, 44, 46, 55, 56]. The mechanisms are not fully understood, however, one explanation might be an increased number of anovulatory menstrual cycles in obese women as a result from a negative feedback switch from increasing oestrogen levels in adipose tissues, leading to a lower total exposure to high levels of oestrogens and progesterones [48, 56, 330]. In contrast, BMI is associated with

increased free bioavailable oestradiol levels in premenopausal women [55]. Further studies are needed to explore a possible mechanism for this.

It was observed that postmenopausal smokers had decreased risk for breast cancer in EPIC-Italy. This has been observed previously but reported results are inconsistent [73, 76, 331]. A protective anti-oestrogenic effect from changes in the hepatic oestrogen metabolism has been seen for smoking postmenopausal women, which might explain the decrease in risk [73, 81, 331]. Furthermore, the association between smoking and breast cancer risk can be confounded by other exposures, for example BMI and alcohol consumption [61, 331]. Indeed, when the analysis was adjusted for BMI in postmenopausal women the association became weaker, but there was still a trend towards decrease in risk for postmenopausal smokers in EPIC-Italy (HR=0.83, 95% CI: 0.68 - 1.01, $P=0.057$).

No significant association between the use of OC or HRT was seen in EPIC-Italy. Some inconsistent associations could be seen for HRT duration: when dividing into groups of < 1 year, 1-3 years, or 4 years or more, with < 1 year as the reference group, the women who had used HRT for 1-3 years had significantly higher risk for breast cancer risk compared with the reference group (HR=1.25, 95% CI: 1.02 - 1.53, $P=0.033$, **Appendix 1**), however this was not significant for ≥ 4 years (HR=0.90, 95% CI: 0.64 - 1.28, $P=0.564$, **Appendix 1**). This could be due to a small number of cases in the ≥ 4 years group (only 33 cases). Furthermore, using the Student's t-test at baseline, there are significantly less OC users and more HRT users for cases compared to controls, which could be due to the slightly older cases compared to controls. When age was adjusted for in the Cox regression model, this is no longer significant. It is not clear whether OC and HRT use increase the risk for breast cancer in a long-term perspective. An increase in risk is often observed for women while they are taking HRT, but the risk levels are reduced back to normal levels after stopping [95, 121, 122, 126]. A similar trend might be the case for OC use. Also, different combinations of oestrogens and progestins, doses, and formulations affect the risk differently and this is not taken into account in this analysis [121, 136].

2.4.2 The ELEE model

Lifetime oestrogen exposure was estimated using a model including a woman's reproductive time, number of pregnancies and breastfeeding duration. The hormonal risk factors in the model (age at menarche, age at menopause, number of pregnancies, breastfeeding duration) have shown consistent results on breast cancer risk in previous epidemiological studies. It was hypothesised that the total

cumulative exposure over time will cause DNA methylation changes in blood and which later on might reflect a woman's risk for breast cancer.

There are some limitations to this ELEE model; the model is missing some variables that will affect oestrogen exposure, but whose contribution to ELEE is difficult to model. For example, as already mentioned, BMI is associated with postmenopausal breast cancer risk, and will have a larger impact on oestrogen levels in postmenopausal compared to premenopausal women. Variables such as menstrual cycle length and regularity will affect the oestrogen exposure by defining the time a woman is exposed to high levels of oestrogens, but this information is often missing in questionnaire data (including in the data in EPIC-Italy) and was therefore not included in the model. Women using OC are exposed to a constant dose of hormones that might affect oestrogen exposure differently in comparison to the increase and decrease of hormone levels in the menstrual cycle. Also, the oestrogen levels are increased during pregnancy but the long-term effect from having been pregnant is a decrease in oestrogen levels [90, 96, 98]. Since a woman spends a longer time in the “decreased state” (several years) compared to being pregnant (9 months), the long-term effect was only included in the model. The ELEE model was not compared to circulating oestrogen levels, due to small sample size of subjects in EPIC-Italy with measured oestrogen levels (n=55 premenopausal women and 79 postmenopausal women) and difficulties in achieving reliable hormone-measurements (no information on menstrual cycle phase in premenopausal women, low levels in postmenopausal women).

Another limitation is the difference in units between variables (for example kg/m² for BMI) and how to combine these together. Smith *et al.* 1999 [300] solved this by using z-scores for each variable (difference between observed value and mean value, divided by standard deviation) before combining all variables together, and Green *et al.* 2013 [305] tried a more complex mathematical modelling approach. In this project, it was decided to only use time-dependent variables in the model for easy interpretation of the model (years of exposure), and adjusted for other factors affecting oestrogen exposure in the EWAS (age, BMI, alcohol consumption, and smoking).

2.4.3 EWAS of ELEE

The EWAS identified a total of 694 CpG sites that were significantly associated with ELEE. Furthermore, factors that could affect the test statistics were analysed, sensitivity analyses were

performed, genomic region enrichment of the significant CpG sites was explored, and the annotation and gene information of some of the top hits were presented.

Beta regression was used to explore the associations between DNA methylation and ELEE. This is a more statistically valid approach compared to linear regression and has been applied in some of the recent EWAS [312-317]. Furthermore the R package ‘bacon’ was used to analyse inflation and bias. Moderate inflation and bias was observed with lambda of 1.13 and a bias of -0.12. The inflation is often decreased when adjusting for confounders in the analysis and it was expected that the inflation should be higher in an unadjusted model. Higher bias was observed in the unadjusted model and when not adjusting for WBC composition, however, a lower inflation. To note, QQ-plots, inflation factors and bias have not been reported in previous studies using beta regression, so these values cannot be compared with previous studies. The methods might not be suitable to use for beta regression test statistics. However, unmeasured or unknown confounders cannot be excluded in the EWAS. There is a risk for false positives, and in this thesis (i.e. **Chapter 3**) the focus has been on the CpG sites with largest estimates and most statistical significance to avoid this.

Cases from EPIC-Italy in the EWAS were stratified on age at diagnosis ≥ 50 to enrich for ER-positive disease. The effects of excluding younger cases from the study was analysed and the results correlated strongly indicating that this did not create any bias. In addition to this, it was analysed if the inclusion of prospective cases affected the results by repeating the same analysis in controls only. Even though this led to a reduction in the statistical power, with only 119 subjects in the analysis, the results correlated very well with the main EWAS. This suggests that this did not affect the analysis extensively. To note, all subjects were healthy at recruitment and blood draw, and the breast cancer cases have been diagnosed years after recruitment.

2.4.4 Transcriptional mechanisms hypothesis

The mechanisms linking oestrogen exposure and breast cancer risk are not fully understood. Oestrogen binds to the ERs that bind to enhancer elements and increase the gene expression of ER-target genes. If this was the underlying mechanism by which oestrogen exposure affects DNA methylation, it should be expected that the DNA methylation changes would occur at ER-binding sites or ER-target genes, and a subsequent change in expression levels. To our knowledge, there is no data of ER-binding sites measured in blood. However, ChIP-seq data was downloaded for ERs from MCF7s (ER binding events

from Jason Carroll's lab [332]) and overlapped with the 694 CpG sites associated with ELEE, but no enrichment was found among the significant CpG sites ($P=0.969$). This might be due to tissue specificity or could indicate another mechanism leading to DNA methylation changes in blood.

However, as hypothesised, if blood can be used as a surrogate tissue for breast, some up-regulation of genes that increase cell proliferation would be expected. The genes in the top table did not show any direct explanation to why they would be affected by oestrogen exposure or leading to breast cancer; only five genes were reported as interacting with oestradiol and three have previously been associated with breast cancer, and might be the ones increasing the breast cancer risk. The 694 CpG sites significantly associated with ELEE were enriched for CpG islands and active promoters, where DNA methylation levels were mostly increasing. This indicates that oestrogen exposure has an effect on gene expression through inactivation of promoters. ER-target genes was expected to be expressed more following oestrogen exposure, which would result in decreased DNA methylation at promoter regions, but this was not observed in this study. The mechanisms of oestrogen-induced DNA methylation are explored further in **Chapter 5**.

2.4.5 EWAS of hormonal risk factors

Few overlaps between significant CpG sites were observed between the EWAS of ELEE and different hormonal risk factors. As most, a total of 31 CpG sites were shared between the EWAS of ELEE and age at menopause, and only two of the 31 CpG sites are in the top results from the EWAS (**Table 2.4**). Age at menopause has the biggest contribution to the ELEE model with a range between 39-57 (18 years) in the EPIC-Italy data. In comparison, the range for age at menarche is 9-17 (8 years), breastfeeding duration 0-4 years, and number of pregnancies is 0-5. So it is not surprising that age at menopause and ELEE are sharing most CpG sites, however, more overlaps were expected. Either the ELEE model captures oestrogen exposure better than the risk factors alone, or it is modelling additional untested factors that also contribute to breast cancer risk. Also, the low statistical power might not be enough to identify small significant changes in DNA methylation from each individual risk factor.

2.4.6 Genome-wide DNA methylation and ELEE

Lastly, the association between genome-wide DNA methylation and ELEE was explored. Previous studies on breast cancer risk have shown hypomethylation in incident cases compared to controls,

however results are inconsistent [197, 204, 206]. Genome-wide hypomethylation is also common in other cancers. In EPIC-Italy some evidence of decreased DNA methylation levels was observed in women with longer ELEE compared to women with shorter ELEE values between two groups split on the median (52.96% and 53.05%). The previous findings showed a mixture of hypo- and hypermethylated CpG sites with increasing ELEE, so it was not expected to observe any significant difference in genome-wide levels. The analysis was repeated in controls only, to exclude any reverse causation, and no significant difference in genome-wide DNA methylation levels were observed between women with high ELEE or low ELEE. However, when excluding the cases the women with highest ELEE are also excluded, which might affect the results. Furthermore, genome-wide hypomethylation has been observed for cases in this dataset previously [197] and ELEE is associated with breast cancer risk.

2.4.7 Conclusions

The results from this study suggest that ELEE alters DNA methylation in blood; however, replication and validation steps are needed in larger study cohorts. If a methylation signature for oestrogen exposure in blood can be identified this might be used as an easy screening method to identify women at high risk for breast cancer. It can also be used to further explore exposures that can affect the lifetime oestrogen exposure, such as environmental exposures that can mimic oestrogen, and to understand the mechanisms between oestrogen exposure and DNA methylation.

CHAPTER 3:
Methylation Index of ELEE and
Breast Cancer Risk

3 CHAPTER 3: Methylation Index of ELEE and Breast Cancer Risk

3.1. INTRODUCTION

In the previous chapter an EWAS of ELEE was performed and 694 CpG probes were identified as significantly associated with ELEE in EPIC-Italy. In this chapter, this is explored further by developing a Methylation Index (MI) of ELEE and testing its association with breast cancer risk. Furthermore, the association is validated in a large independent study cohort, the Generations Study, using targeted bisulphite sequencing.

3.1.1 Replication and validation in methylation studies

Replication and validation are the processes of repeated analysis in independent samples (and preferable study cohort) using the same or a different methylation sequencing technique. In the previous chapter the statistical challenges in EWAS were discussed as well as how the results might be affected by bias and inflation from small sample sizes and unadjusted confounding factors. The cost of a large genome-wide methylation study is still relatively high and the sample sizes for EWAS are often small (< 1000 subjects), resulting in low statistical power [333]. Furthermore, in comparison to the categorical genotyping in GWAS, DNA methylation levels are quantitative and will have more technical variance depending on the bisulphite-conversion protocols, sequencing technique, and statistical processing [334]. Therefore, replication and validation are needed to establish robust associations between methylation markers, exposure and disease. This will identify false positive hits, find the true biological variability, and determine the generalisation of the methylation signatures, which all are important considerations for development of clinical biomarkers [208, 333, 335]. Replication steps have been included in a majority of the published EWAS and are important for the quality and interpretability of the study [197, 214, 215, 219, 220, 229, 232, 289, 299, 336, 337]. Bisulphite pyrosequencing has often been used for validation in a number of EWAS [195, 203, 205, 232, 273, 334] and is a relatively easy, quick, high-throughput method for detection of methylation levels at the genome-wide level or for a small number of loci for a reasonable cost [183].

As already covered in **Chapter 2**, EWAS for breast cancer risk have so far not replicated any strong methylation signatures, and combined meta-analyses or replication studies have often resulted in smaller effect sizes and lower significance in comparison to the discovery dataset [197, 201, 205, 289]. Therefore, one should be cautious in drawing too strong conclusions from one single EWAS conducted for one study cohort with small sample sizes, before the findings have been replicated and validated. To increase the sample sizes in EWAS the need of larger consortia collaborations has been highlighted and has resulted in recent large published methylation studies in for example BMI [213], alcohol consumption [299], and tea and coffee consumption [338].

3.1.2 Model development in DNA methylation array data

A few EWAS have identified DNA methylation signatures of an exposure or trait of interest by combining CpG sites into a predictive model, as example for the prediction of smoking status [223, 337], alcohol consumption, [299], and aging [214, 215]. The idea is similar to PRS computed from GWAS, where small effects from individual risk-SNPs are combined into a predictive model with a larger combined effect [41, 339].

Statistical modelling of microarray DNA methylation data can be challenging due to the nature of the data, including collinearity (high correlation between groups of CpG sites) and that the number of predictors (CpG sites) are often much larger than the number of observations (sample sizes). Ridge regression is a technique that fits a model by minimising the residual sum of squares of the coefficients using a penalty factor lambda that shrinks the coefficients. The lambda is often tuned using repeated cross-validation in the development data. A lambda equal to zero corresponds to the ordinary least squares estimation, which is not as good at handling highly correlated data and tends to over fit the model [340]. A larger lambda leads to a greater amount of shrinkage, which can decrease the prediction accuracy of the model [341]. Ridge regression is ideal when there are a larger number of predictors and non-zero coefficients is desired, i.e. to keep all predictors in the model [342].

Lasso (least absolute shrinkage and selection operator), is the combination of ridge regression, coefficient shrinkage and subset selection [340]. This method shrinks some coefficients to 0 that are not included in the model. This might be desirable when the number of predictors is large and one wants to identify a subset of the predictors that contributes the most to the model. However, lasso does not handle collinearity very well and is therefore often replaced with the elastic net regression [343].

The elastic net is a compromise between ridge regression and lasso, and encourages a grouping effect where strongly correlated predictors tend to be in or out of the model together. Elastic net has been shown to often outperform lasso in terms of accuracy prediction.

3.1.3 Aims of this chapter

It was hypothesised that lifetime oestrogen exposure leads to DNA methylation changes detectable in blood that are associated with breast cancer risk. From the EWAS of ELEE, CpG sites associated with ELEE were combined into a MI model using HM450K data and tested for association with ELEE and breast cancer risk in case-control pairs in EPIC-Italy. Furthermore, the MI was validated using targeted bisulphite sequencing in a large independent study cohort (the Generations Study, n=880).

3.2. STUDY COHORTS

3.2.1 EPIC-Italy

The same HM450K array data from **Chapter 2** has been used in this chapter. For the development of the MI all women with complete information of ELEE were included (n=237, **Table 3.1**) and in the evaluation of the MI and breast cancer risk, all matched cases and controls were included (n=324, **Table 3.2**).

3.2.2 The Generations Study

The Generations Study cohort is an on-going prospective study from Breast Cancer Now and the Institute of Cancer Research (ICR) and has collected epidemiological data and blood samples from approximately 113,000 women in the UK from the start of recruitment in 2004 with the primary aim to study breast cancer aetiology [344]. In this study three different datasets generated from the Generations Study cohort have been used.

3.2.2.1 Pooled samples

For the test phase in this study, pooled samples for breast and ovarian cancer incident cases and matched controls were included [197]. A total of 548 breast cancer case-control pairs were divided in four pools stratified on breast cancer subtype: 123 cases diagnosed with DCIS, 66 cases with invasive ER-negative disease, 179 cases with invasive ER-positive disease with age at diagnosis < 50 years, and 189 cases with invasive ER-positive disease with age at diagnosis > 50 years. Controls were individually matched to cases on age, ethnicity, and date of recruitment. The average age at recruitment was 52 for all cases and controls with an average time to diagnosis of 2.7 years. The pooled ovarian cancer samples (unpublished data) included 78 women diagnosed with ovarian cancer with age at recruitment of 57 and mean time to diagnosis of 3.1 years.

3.2.2.2 HM450K array data

The Generations Study HM450K dataset included samples from 92 healthy women recruited in 2004 and with an additional blood sample taken in 2010 (5.5 to 6.5 years after first blood collection) with no registered cancer at follow up [207]. For the development of the MI, samples from the first time point were used including women with complete information of ELEE (n=65, **Table 3.1**).

3.2.2.3 *Samples for targeted bisulphite sequencing*

For the targeted bisulphite sequencing 880 independent samples (440 matched case-control pairs) were selected from the Generations Study (**Table 3.3**). Inclusion criteria for incident breast cancer cases, with blood samples taken prior to diagnosis, were the following: invasive ER-positive breast cancer with no previous history of (non-breast) cancer, white ethnicity and completeness of epidemiological data. Controls were individually matched to cases on age at blood draw $\pm$ 5 years.

Table 3.1. Table of characteristics for EPIC-Italy and the Generations Study HM450K data
Subjects with complete information of ELEE included in the development of the MI.

			EPIC-Italy (n=237)	Generations Study (n=65)
Incident cases	n (%)		118 (49.8%)	0
Age	mean (st.dev.), yrs		53.1 (7.2)	51.8 (8.4)
Time to diagnosis	mean (st.dev.), yrs		5.0 (4.2)	NA
Menopausal status	n (%)	<i>Premenopausal</i>	79 (33.3%)	32 (49.2%)
		<i>Postmenopausal</i>	132 (55.7%)	33 (50.8%)
Age at menarche	mean (st.dev.), yrs		12.7 (1.5)	12.8 (1.6)
Age at menopause	mean (st.dev.), yrs		49.7 (3.6)	48.8 (5.7)
Parity	n (%)	<i>Parous</i>	198 (83.5%)	54 (83.1%)
		<i>Nulliparous</i>	39 (16.5%)	11 (16.9%)
Number of pregnancies	mean (st.dev.)		1.7 (1.1)	1.9 (1.1)
Ever breastfed*	n (%)		198 (100%)	47 (87.03%)
Breastfeeding duration*	mean (st.dev.), yrs		0.7 (0.6)	NA
OC ever	n (%)		87 (36.7%)	50 (76.9%)
OC duration	mean (st.dev.), yrs		5.0 (5.8)	NA
HRT ever	n (%)		36 (15.2%)	15 (23.1%)
HRT duration	mean (st.dev.), yrs		2.0 (2.9)	NA
BMI	mean (st.dev.), kg/m ²		25.7 (4.3)	25.4 (3.9)
Alcohol consumption	mean (st.dev.), average g/day		6.6 (8.7)	15.6 (12.9)
Smoking status	n (%)	<i>Smoker</i>	46 (19.4%)	2 (3.1%)
		<i>Former</i>	48 (20.3%)	19 (29.2%)
		<i>Never</i>	143 (60.3%)	44 (67.7%)
Smoking duration	mean (st.dev.), yrs		9.2 (13.1)	NA
Family BRCA history	n (%)		NA	13 (20.0%)
Benign disease history	n (%)		NA	14 (21.5%)

* Breastfeeding variables in parous women only
st.dev = standard deviation; yrs = years

Table 3.2. Table of characteristics for case-control pairs in EPIC-Italy HM450K data

Matched case-control pairs in EPIC-Italy HM450K data used to test the association between the MI and breast cancer risk.

			EPIC-Italy			P value**
			All (n=324)	Cases (n=162)	Controls (n=162)	
Age	mean (st.dev.), yrs		52.9 (7.1)	52.9 (7.2)	53 (7.1)	0.543
Time to diagnosis	mean (st.dev.), yrs		NA	5.3 (4.4)	NA	
Menopausal status	n (%)	<i>Premenopausal</i>	101 (31.2%)	52 (32.1%)	49 (30.2%)	0.838
		<i>Postmenopausal</i>	172 (53.1%)	85 (52.5%)	87 (53.7%)	
Age at menarche	mean (st.dev.), yrs		12.7 (1.5)	12.7 (1.4)	12.7 (1.7)	0.971
Age at menopause	mean (st.dev.), yrs		49.6 (3.8)	50.2 (3.7)	49.1 (3.8)	0.058
Parity	n (%)	<i>Parous</i>	280 (86.4%)	137 (84.6%)	143 (88.3%)	0.589
		<i>Nulliparous</i>	42 (13.0%)	23 (14.2%)	19 (11.7%)	
Number of pregnancies	mean (st.dev.)		1.7 (1.0)	1.6 (1.1)	1.7 (1.0)	0.181
Ever breastfed*	n (%)		215 (76.8%)	103 (75.2%)	112 (78.3%)	1.000
Breastfeeding duration*	mean (st.dev.), yrs		0.7 (0.6)	0.7 (0.6)	0.8 (0.7)	0.213
OC ever	n (%)		126 (38.9%)	59 (36.4%)	67 (41.4%)	0.478
OC duration	mean (st.dev.), yrs		4.9 (5.5)	4.1 (5.1)	5.6 (5.8)	0.283
HRT ever	n (%)		51 (15.7%)	23 (14.2%)	28 (17.3%)	0.558
HRT duration	mean (st.dev.), yrs		2.7 (3.1)	2.8 (2.7)	2.6 (3.4)	0.311
BMI	mean (st.dev.), kg/m ²		25.6 (4.2)	25.8 (4.1)	25.3 (4.3)	0.314
Alcohol consumption	mean (st.dev.), average g/day		6.5 (8.6)	5.5 (7.0)	7.4 (9.9)	0.071
Smoking status	n (%)	<i>Smoker</i>	67 (20.7%)	31 (19.1%)	36 (22.2%)	0.021
		<i>Former</i>	64 (19.8%)	23 (14.2%)	41 (25.3%)	
		<i>Never</i>	191 (59.0%)	106 (65.4%)	85 (52.5%)	
Smoking duration	mean (st.dev.), yrs		9.3 (13.1)	8.1 (12.6)	10.6 (13.4%)	0.075

st.dev = standard deviation; yrs = years

* Breastfeeding variables in parous women only

**P-values indicate differences between cases and controls. Student's paired t-test was used for continuous variables and chi-squared test for categorical variables.

Table 3.3. Table of characteristics for the Generations Study subjects for targeted bisulphite sequencing

		Samples for sequencing			Samples for MI analysis		
		All (n=880)	Cases (n=440)	Controls (n=440)	All (n=678)	Cases (n=339)	Controls (n=339)
Age	mean (st.dev.), yrs	53.2 (10.2)	53.1 (10.2)	53.3 (10.2)	54.0 (10.3)	53.9 (10.3)	54.1 (10.4)
Time to diagnosis	mean (st.dev.), yrs	NA	3.9 (2.4)	NA	NA	4.0 (2.4)	NA
Menopausal status	n (%)	372 (42.3%)	193 (43.9%)	179 (40.7%)	262 (38.6%)	135 (39.8%)	127 (37.5%)
		508 (57.7%)	247 (56.1%)	261 (59.3%)	416 (61.4%)	204 (60.2%)	212 (62.5%)
Age at menarche	mean (st.dev.), yrs	12.7 (1.5)	12.7 (1.5)	12.7 (1.5)	12.7 (1.5)	12.7 (1.5)	12.7 (1.5)
Age at menopause	mean (st.dev.), yrs	50.1 (4.5)	50.2 (4.4)	49.9 (4.5)	50.2 (4.4)	50.3 (4.3)	50.1 (4.5)
Parity	n (%)	730 (83.0%)	365 (83.0%)	365 (83.0%)	566 (83.5%)	283 (83.5%)	283 (83.5%)
		150 (17.1%)	75 (17.1%)	75 (17.1%)	112 (16.5%)	56 (16.5%)	56 (16.5%)
Number of pregnancies	mean (st.dev.)	1.9 (1.2)	1.9 (1.1)	1.9 (1.2)	1.9 (1.2)	1.9 (1.2)	2.0 (1.2)
Ever breastfed*	n (%)	690 (94.5%)	339 (92.9%)	351 (96.2%)	539 (93.6%)	265 (93.6%)	274 (96.8%)
Breastfeeding duration*	mean (st.dev.), yrs	0.8 (1.1)	0.8 (1.0)	0.9 (1.1)	0.8 (1.0)	0.8 (1.0)	0.8 (0.9)
OC ever	n (%)	718 (81.6%)	355 (80.7%)	347 (78.9%)	544 (80.2%)	270 (79.7%)	265 (78.2%)
OC duration	mean (st.dev.), yrs	8.5 (6.5)	8.3 (6.2)	8.7 (6.8)	8.4 (6.6)	8.1 (6.4)	8.8 (6.8)
HRT ever	n (%)	265 (30.1%)	125 (28.4%)	135 (30.7%)	215 (31.7%)	106 (31.3%)	107 (31.6%)
HRT duration	mean (st.dev.), yrs	6.3 (5.2)	6.9 (5.7)	5.7 (4.6)	6.1 (5.1)	6.7 (5.7)	5.4 (4.3)
BMI	mean (st.dev.), kg/m ²	25.6 (4.5)	25.7 (4.5)	25.4 (4.6)	25.4 (4.3)	25.7 (4.3)	25.2 (4.3)
Alcohol consumption	mean (st.dev.), average units/week	13.5 (14.1)	14.43 (14.5)	12.5 (13.6)	13.3 (14.1)	13.8 (14.7)	12.8 (13.5)
Smoking status	n (%)	57 (6.5%)	29 (6.6%)	28 (6.4%)	48 (7.1%)	25 (7.4%)	23 (6.8%)
		243 (27.6%)	132 (30.0%)	111 (25.2%)	184 (27.1%)	98 (28.9%)	86 (25.4%)
Smoking duration	mean (st.dev.), yrs	580 (65.9%)	279 (63.4%)	301 (68.4%)	446 (65.8%)	216 (63.7%)	230 (67.9%)
Family BRCA history	n (%)	4.0 (8.61)	4.1 (8.5)	3.9 (8.7)	4.2 (9.0)	4.4 (9.0)	3.9 (9.1)
Benign disease history	n (%)	181 (20.6%)	108 (24.6%)	73 (16.6%)	144 (21.2%)	86 (25.4%)	58 (17.1%)
		246 (28.0%)	147 (33.4%)	99 (22.5%)	196 (28.9%)	115 (33.9%)	81 (23.9%)

* Breastfeeding variables in parous women only

** P-values indicate differences between cases and controls: Paired student's t-test was used for continuous variables and chi-squared test for categorical variables.
st.dev = standard deviation; yrs = years

3.3. METHODS

3.3.1 Targeted bisulphite sequencing using the Fluidigm 48.48 Access Array

3.3.1.1 *Selection of target CpG sites*

The target CpG sites were selected from the 694 CpG sites identified as significantly associated with ELEE in the EWAS of ELEE based on magnitude of change in DNA methylation and significance (**Appendix 2**). The selection was done in an iterative process; some CpG sites were not included in the target panel due to difficulties with primer design or exclusion after quality control, as described below.

3.3.1.2 *WBC estimation in the targeted sequencing data*

To estimate WBC composition for the sequenced samples, five CpG sites highly correlated with WBC counts for B-cells, CD4+ T-cells, CD8+ T-cells, Monocytes and Natural Killer cells were included in the target panel. The CpG sites were identified from the probes included in the Houseman method and *P*-values were estimated using generalised linear regression models for WBC counts and methylation levels. The CpG probes showing most significant association with the estimated WBC composition in HM450K data from EPIC-Italy, and that replicated in the Generations Study HM450K data, were included in the target panel.

3.3.1.3 *Test phase*

Since targeted bisulphite sequencing using the Fluidigm 48.48 Access Array had not been used previously in our laboratory and only a few publications using this technique for methylation studies had been published at the start of this study [345-347], the entire workflow was tested before sequencing in the 880 Generations Study samples. This included primer design, bisulphite conversion, target amplification using the Fluidigm 48.48 Access Array, library preparation and sequencing, and data processing. Four pools of breast cancer cases (diagnosed with DCIS, ER-negative disease, ER-positive disease with age at diagnosis < 50, and ER-positive disease with age at diagnosis > 50) with matched controls [197] and one pool of ovarian cancer cases with matched controls from the Generations Study cohort, as well as the breast cell lines MCF7, MCF10A and MCF12A were used as samples in the test phase. Controls included 0% methylation DNA (generated from whole-genome amplified DNA [348]), 100% methylated DNA (CpGenome Universal Methylated DNA from

Millipore), and one non-template control. For all samples, 250ng of DNA was bisulphite converted, except for one sample where 60, 125, 250 and 500ng of DNA were included. Quality control of generated sequencing data included batch effects, specificity for alignment of reads in target regions, and sequencing coverage for target CpG sites.

3.3.1.4 Target amplification using the Fluidigm 48.48 Access Array

Before target amplification the DNA was bisulphite converted using the Zymo Research EZ-96 DNA Methylation kit (Cambridge Biosciences) and bisulphite conversion was verified using quantitative PCR with primers specifically designed to only amplify bisulphite converted DNA. Target regions were amplified using the Fluidigm 48.48 Access Array [349], which allows amplification of up to 48 target regions in 48 samples in one single run using integrated fluidic circuits and pooling of up to 384 samples before sequencing by the addition of specific barcode sequences. The Access Array (Juno LP 48.48), reagents for the Access Array and barcode library were ordered from Fluidigm and PCR reagents were ordered from Roche (FastStart High Fidelity PCR System). Amplification of products was confirmed by observing the size-distribution of the resulting 48 sample-pools (15 samples randomly selected per batch) on the BioAnalyzer Agilent 2200 TapeStation using the Agilent High Sensitivity D5000 ScreenTape and reagents (Agilent Technologies). All 48 samples in one batch with 48 PCR products were pooled together and cleaned up using homemade magnetic beads (similar to Agencourt AMPure beads, Beckman Coulter) in a 1:1 ratio to remove unwanted products smaller than 100bp. Clean up was confirmed on the BioAnalyzer.

3.3.1.5 Primer design

Primers were designed using the recommended settings from Fluidigm (product size between 150 to 200 base pairs (bp), annealing temperature 59°C to 61°C) with MethPrimer (v 1.0, <http://www.urogene.org/methprimer>) and Primer3 (v 4.0, <http://primer3.ut.ee>) and ordered from Sigma-Aldrich. If possible, overlap with common SNPs (frequency > 1%) or CpG sites were avoided for the primer sequences. The ‘four-primer amplicon tagging’ protocol was used for target amplification on the Fluidigm 48.48 Access Array platform [350]. Therefore, a common sequence tag was added at the 5’ end of each primer where specific sample barcodes and Illumina sequences can bind, as provided by Fluidigm. For sequencing on the Illumina sequencing systems, sequencing

primers were ordered from Exiqon. Prior to the targeted sequencing, primer specificity was validated by PCR and gel electrophoresis in the test phase.

3.3.1.6 Targeted bisulphite sequencing in the Generations Study

Targeted bisulphite sequencing was conducted for the selected target CpG sites (passing quality control in the test phase) in the Generations Study (n=880, **Table 3.3**) using the Fluidigm 48.48 Access Array. A total of 250ng of DNA was bisulphite-converted before target amplification using the Fluidigm 48.48 Access Array. The 880 samples were divided in 20 batches in four sequencing pools (6, 5, 5 and 4 batches). Each Fluidigm-batch consisted of 44 samples (22 matched case-control pairs), two duplicated pairs, one commercial DNA sample (Promega Human Genomic DNA) to analyse batch effects, and one non-template control. The purified, pooled libraries were sequenced on the Illumina MiSeq system (paired-end 150bp). The case-control status for the samples was blinded before data analysing.

3.3.1.7 Alignment and extraction of DNA methylation levels

FastQ-files for each sample were provided from the sequencing facility. The reads were trimmed for quality, with a Phred quality score of 20 as threshold, and adapter sequences using Trim galore and Trimmomatic. Trim galore uses 'cutadapt' to trim adapter sequences and remove reads of poor quality and reads of length less than 20bp [351]. Trimmomatic uses a sliding window and cut off sequences with a low quality in that window [352]. Reads were paired and aligned using Bismark and Bowtie2 with the four bisulphite-converted DNA strands as template [353]. DNA methylation levels at each CpG sites were extracted using the bismark methylation extractor tool and bedGraph-files were created using the bismark2bedGraph tool. BedGraph-files with methylation levels and coverage were further analysed in R Studio version 3.3.2.

3.3.2 MI of ELEE

Prior to the development of the MI, quality control of the target CpG sites was conducted in sequenced samples from the Generations Study. For each batch and target CpG site the number of pairs with coverage < 30 reads in at least one of the samples was calculated, and if an average of 30% of the pairs had low coverage across the whole cohort the target CpG site was excluded prior to the MI analysis. The duplicated samples were used to estimate variation in DNA methylation at each target CpG site, by

calculating the mean difference in DNA methylation levels in each batch. Target CpG sites with a mean difference of 20% or more between duplicated samples were also excluded. The CpG sites passing these steps were included in the MI model.

3.3.2.1 Generation and pre-processing of the HM450K array data

The HM450K data generation and pre-processing for the Generations Study have been described previously [207]. Briefly, DNA samples were bisulphite converted using the EZ-96 DNA Methylation Gold kit (Zymo Research) and methylation data was generated using the Illumina HumanMethylation 450K BeadChip. The CpG probes with detection P -value > 0.05 in 20% of samples were excluded, and samples with detection P -value > 0.05 in 1% of probes. The beta-values were pre-processed using peak-based correction [324] and ‘ComBat’ to minimize batch effects [326]. The processing of the HM450K data for EPIC-Italy was described in **Chapter 2**.

3.3.2.2 Model development

For model development of the MI, the available HM450K data from the Generations Study and EPIC-Italy were merged together. Samples with missing data for ELEE were excluded ($n=28$ and $n=87$), leaving 65 and 237 samples for the analysis (**Table 3.1**). Duration of breastfeeding was missing for the Generations Study HM450K data subjects and not included in the ELEE-model for these subjects. A MI of ELEE was developed on the CpG sites passing quality control in the targeted sequencing data. For model development ridge regression (without subset selection, to include all CpG sites) was conducted in a 10-fold cross-validation repeated a 100 times (R function ‘glmnet’ in ‘train’ in the R package ‘caret’) to estimate parameters (penalty coefficient lambda for ridge regression and regression coefficients for the model). The final MI model for the predicted ELEE was calculated as a linear function, i.e. intercept plus the sum of the weighted DNA methylation levels at the CpG sites.

3.3.2.3 Model evaluation

The correlation between the MI and ELEE was evaluated in the Generations Study targeted sequencing data using Pearson’s correlation coefficient. The association between the MI and risk for breast cancer was evaluated using conditional logistic regression (R function ‘clogit’ in the package ‘survival’) in matched case-control pairs in EPIC-Italy HM450K data (**Table 3.2**) and the Generations Study targeted sequencing validation data (**Table 3.3**). The ORs were adjusted for age, BMI, alcohol consumption, smoking duration (all variables were reported at the time of recruitment) and estimated WBC

composition in a multivariable model. The association between the MI and breast cancer risk was explored both as continuous variable and by control-quartiles. The discrimination between cases and controls was assessed using the area under the receiver operating characteristic curve (AUC) by bootstrapping 1000 times in the R function 'pROC'. An AUC of 0.5 indicates that the discriminatory accuracy of the model is no better than by chance, and 1.0 for perfect discrimination.

3.3.3 Individual CpG sites and breast cancer risk

For all CpG sites included in the target regions (n=274 CpGs) the individual association with breast cancer risk was explored using conditional logistic regression in the Generations Study targeted sequencing data. Furthermore, all CpG sites overlapping HM450K probes were also analysed for breast cancer risk in matched case-control pairs EPIC-Italy HM450K data. For the analysis of associations between individual sequenced CpG sites and breast cancer risk the DNA methylation levels were converted to z-scores (methylation value minus the mean methylation value divided by standard deviation) to determine ORs per one standard deviation (ORperSD) and allow direct comparison of each site. The ORs were adjusted for age, BMI, alcohol consumption, smoking duration (all variables were reported at the time of recruitment) and WBC composition.

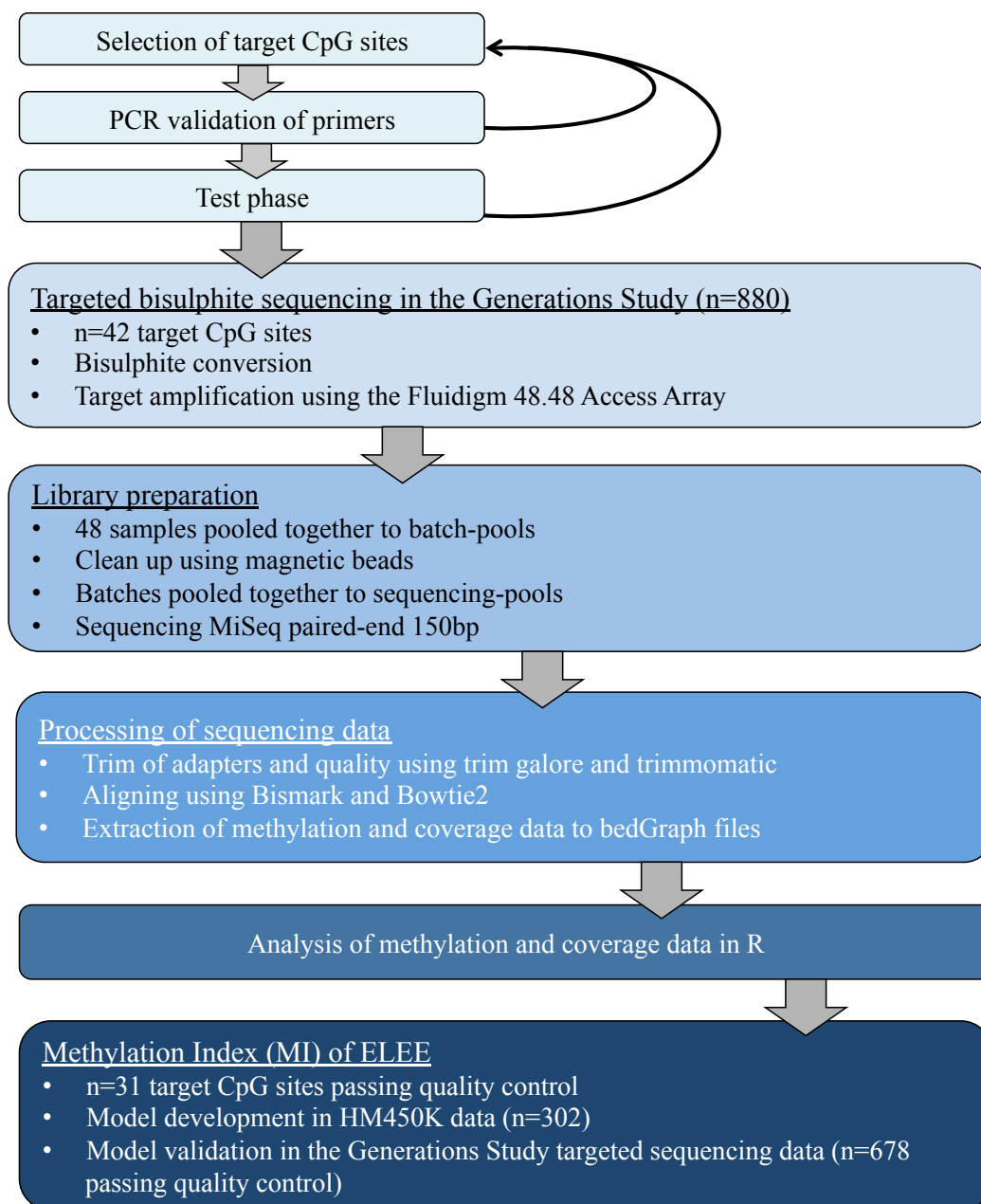


Figure 3.1. Workflow for targeted bisulphite sequencing and data analysis

Target CpG sites were selected in an iterative process after validation of primers and quality control in the test phase. Targeted bisulphite sequencing was conducted for the final 42 target CpG sites in 880 samples in the Generations Study, libraries of batch-pools were prepared and sequenced, and sequencing data was processed before analysed in R. The MI was developed using HM450K data for 31 CpG sites passing quality control in the targeted bisulphite sequencing. The association between the MI and breast cancer risk was validated in the Generations Study targeted sequencing data (n=678 samples passing quality control).

3.4. RESULTS

The workflow including validation of primers, the test phase, targeted bisulphite sequencing using the Fluidigm 48.48 Access Array, library preparation, sequencing data processing and MI analysis, is described in **Figure 3.1**.

3.4.1 Test phase

3.4.1.1 *Primer validation and quality control of targeted sequencing data*

The designed primers were validated by PCR on bisulphite converted Human Genomic DNA (Promega) and gel electrophoresis on the PCR products. Primer pairs were excluded if no PCR product was amplified or if more than one PCR product was visible. Targets were amplified in two batches using the Fluidigm 48.48 Access Array for 46 primer-pairs in the first batch and 47 primer-pairs in the second batch (after excluding some targets with not validated primers). This included five CpG sites for WBC estimation that showed significant association with WBC estimation from the Houseman method in HM450K data (**Figure 3.2**).

The specificity for alignment in target regions was high with 99.9% and 98.2% of the total coverage observed inside target regions in the two batches. Furthermore, the non-template-control had no mapped reads, the 0% methylated control samples had low methylation levels (median values of 0.42% and 0.55%) and the 100% methylated control samples were highly methylated (71.15% and 92.62%) in the two batches.

Samples that clearly failed the sequencing process (67% or more of the target CpG sites with coverage < 500 reads) were excluded before additional analyses. This included two samples in the first batch (88% and 94% CpG sites with coverage < 500 reads) and five samples in the second batch (100%, 67%, 79%, 90% and 100% CpG sites with coverage < 500 reads). A threshold of 500 reads was used for the coverage, estimated from that the coverage should decrease by approximately a factor of 18 (880/48) in the sequencing samples, and a threshold of 30 reads was to be used in the later analysis including 880 samples.

The DNA methylation levels between batches correlated well. After removal of CpG sites with low coverage (coverage < 500 in 20% of the samples) the correlation coefficient was > 0.99 for the CpG sites inside target regions (**Figure 3.3A**). Different concentrations of DNA were tested for one

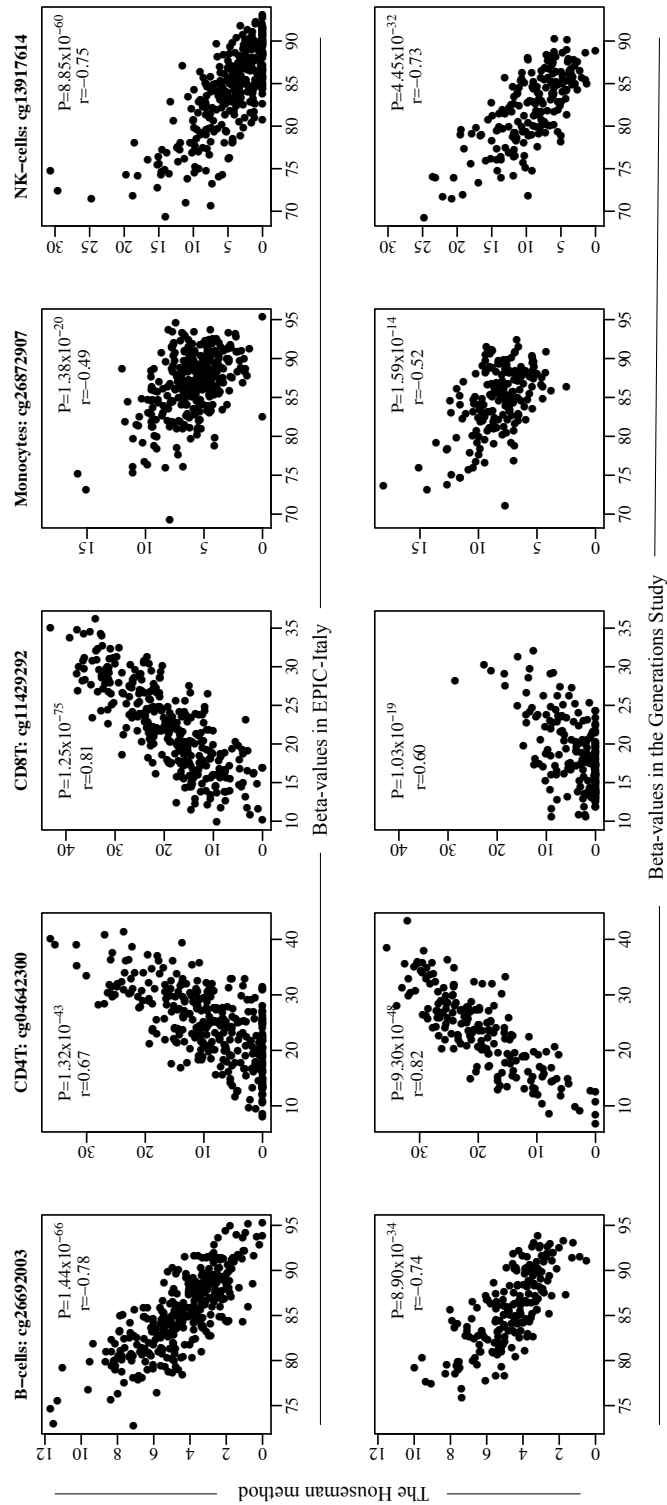


Figure 3.2. CpG probes for estimation of WBC composition
 Five CpG sites used in the Houseman method individually associated with WBC composition for B-cells, CD4+ T-cells, CD8+ T-cells, monocytes and natural killer (NK)-cells were included in the target sequencing panel. The plot shows the correlation between the HM450K beta-values (methylation levels) against WBC composition estimated through the Houseman method in EPIC-Italy and the Generations Study.

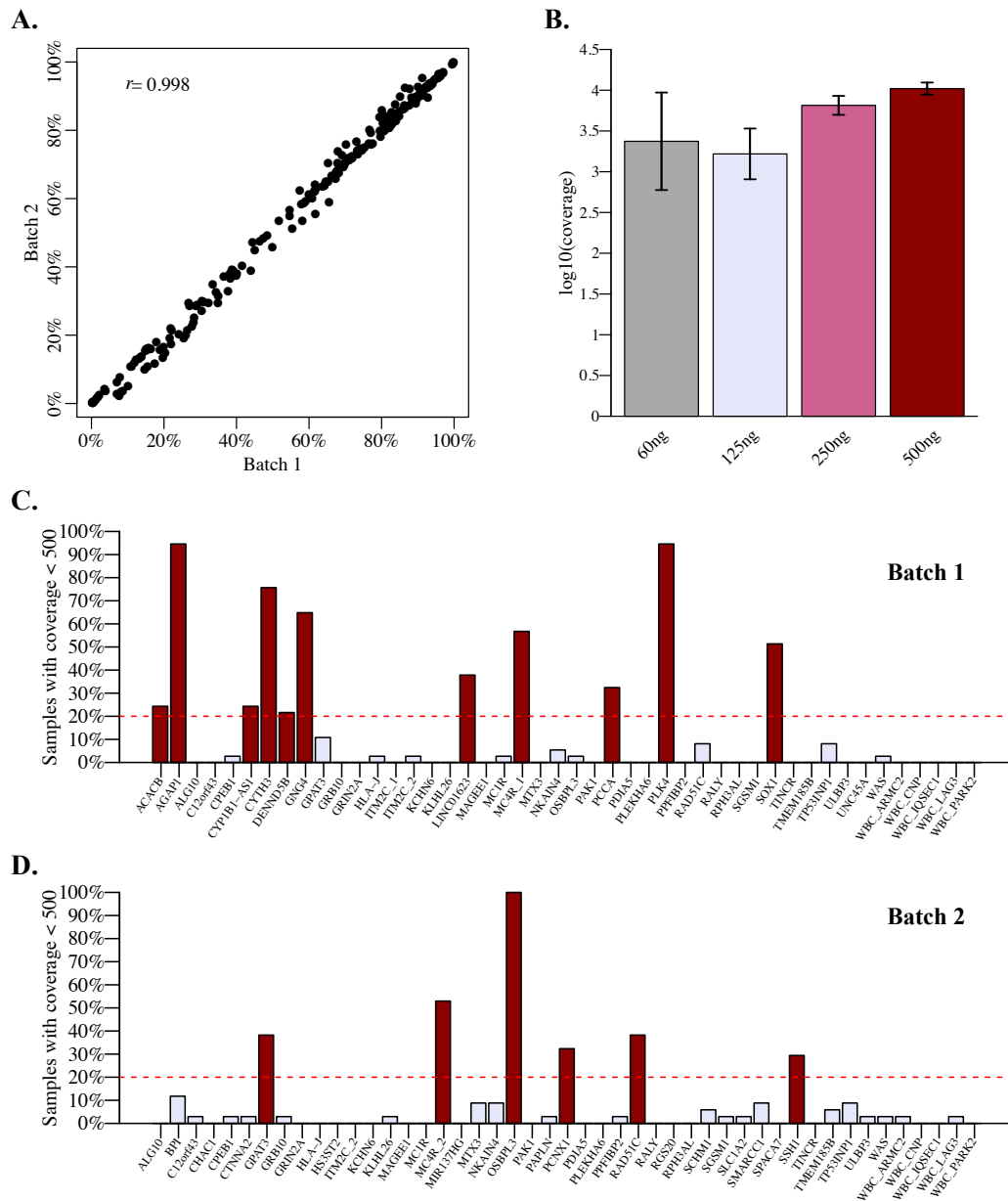


Figure 3.3. Test phase: quality control for targeted bisulphite sequencing using the Fluidigm 48.48 Access Array

A. Analysis of batch effects by correlating methylation levels between the two batches. **B.** For one of the samples four different concentrations of DNA were tested. The bar plots shows the average $\log_{10}$ coverage for 60, 125, 250 and 500ng DNA (before bisulphite conversion). **C.** Percentage of samples with coverage < 500 reads for all target CpG sites, including WBC CpG sites, in batch 1. **D.** Percentage of samples with coverage < 500 reads for all target CpG sites, including WBC CpG sites, in batch 2. Red bars indicates values above the threshold (red dotted lines).

Table 3.4. The 42 CpG sites included in the targeted bisulphite sequencing

Annotated information and result from the EWAS of ELEE in EPIC-Italy for the 42 target CpG sites (probes presented in alphabetic order for the nearest gene). The estimates correspond to regression coefficients from the mixed-effects linear regression model (percentage change in DNA methylation per unit longer ELEE), and *P*-values from the beta regression model, which have been corrected for multiple testing using FDR (*Q*-values).

HM450K probe	Chr	Position	Nearest gene	Distance to gene	EWAS of ELEE			
					Estimate	SE	<i>P</i> value	<i>Q</i> value
cg01893629	chr12	34494825	ALG10	313588	-0.14	0.03	8.15E-05	0.050
cg08254089*	chr20	36933189	BPI	0	-0.28	0.08	7.04E-05	0.048
cg21590238	chr12	121454837	C12orf43	536	-0.29	0.08	2.67E-06	0.014
cg21153102	chr15	41252147	CHAC1	3429	-0.43	0.14	6.16E-05	0.046
cg03340215*	chr15	83315615	CPEB1	0	-0.31	0.08	4.78E-05	0.042
cg06968859	chr2	80724209	CTNNA2	0	-0.26	0.05	2.52E-09	0.001
cg12105860*	chr12	31742801	DENND5B	0	0.14	0.03	1.58E-06	0.012
cg16840364*	chr4	84539569	GPAT3	12541	-0.14	0.06	4.77E-05	0.042
cg08835688	chr7	50849931	GRB10	0	-0.14	0.03	4.50E-05	0.041
cg08349826	chr16	10346403	GRIN2A	69791	-0.11	0.04	4.19E-05	0.040
cg23681866	chr6	29895175	HLA-J	0	-1.33	0.31	2.17E-05	0.031
cg22968966	chr16	22959875	HS3ST2	32215	-0.40	0.12	7.95E-05	0.049
cg15127563	chr2	231729487	ITM2C	132	0.36	0.09	1.77E-05	0.030
cg20020161	chr2	231732669	ITM2C	0	-0.14	0.04	4.49E-05	0.041
cg22097768	chr17	61615913	KCNH6	0	-0.20	0.05	4.88E-05	0.043
cg17969123	chr19	18745971	KLHL26	1865	-0.15	0.04	3.18E-05	0.037
cg05422360	chrX	75648455	MAGEE1	0	-0.44	0.12	1.72E-07	0.005
cg01768446	chr16	89982419	MC1R	1866	-0.13	0.03	4.57E-05	0.041
cg25372296	chr1	98510328	MIR137HG	0	0.33	0.09	3.83E-06	0.017
cg04519403*	chr5	79298951	MTX3	11862	-0.24	0.05	7.51E-06	0.022
cg12091786	chr20	61877942	NKAIN4	0	-0.37	0.1	2.07E-05	0.031
cg25279613*	chr7	24956523	OSBPL3	0	0.20	0.06	4.95E-06	0.020
cg24536703	chr11	77183438	PAK1	0	-0.32	0.09	2.71E-05	0.034
cg24036523	chr14	73712256	PAPLN	0	-0.45	0.11	3.05E-05	0.037
cg16720405	chr3	122790178	PDIA5	0	-0.17	0.05	5.17E-05	0.043
cg13674411	chr1	204232677	PLEKHA6	0	0.11	0.03	3.50E-05	0.038
cg20684174	chr11	7541255	PPFIBP2	0	-0.13	0.03	1.23E-05	0.027
cg01430588*	chr17	56769767	RAD51C	194	-0.32	0.09	6.04E-05	0.045
cg22273487	chr20	32580931	RALY	525	0.19	0.05	1.60E-05	0.028
cg22343083	chr8	54786401	RGS20	0	-0.30	0.07	1.92E-06	0.013
cg22758104	chr17	50465	RPH3AL	11713	-0.23	0.06	4.62E-07	0.006
cg16733643*	chr1	41575522	SCMH1	0	-0.42	0.11	5.00E-05	0.043
cg17588491	chr22	25198892	SGSM1	3242	-0.17	0.04	4.41E-05	0.041
cg13971030	chr11	35366721	SLC1A2	0	-0.49	0.14	9.19E-06	0.024
cg17567562*	chr3	47687980	SMARCC1	0	-0.48	0.11	1.18E-05	0.027
cg10298859	chr13	112883993	SPACA7	146656	-0.20	0.04	2.26E-07	0.005
cg19216791	chr19	5568216	TINCR	210	-0.25	0.07	4.13E-05	0.040
cg25936380	chr2	120981591	TMEM185B	606	-0.29	0.07	2.22E-07	0.005
cg01824466*	chr8	95959531	TP53INP1	0	-0.23	0.08	8.08E-05	0.050
cg26657235	chr6	150378972	ULBP3	4367	-0.12	0.03	1.32E-05	0.028
cg08551047	chr15	91473569	UNC45A	0	-0.58	0.14	2.87E-06	0.015
cg20394620*	chrX	48541924	WAS	260	-0.20	0.05	3.02E-06	0.015

* marked CpG probes did not pass quality control in the targeted sequencing data and were not included in the analysis of the MI

SE=Standard Error

of the samples, using 60, 125, 250 and 500ng of DNA before bisulphite conversion, to evaluate how much DNA was required to generate sufficient sequencing data. All concentrations resulted in relatively high coverage, but with larger variations for 60 and 125ng (**Figure 3.3B**). Therefore 250ng was used in the following experiments.

3.4.1.2 Excluded target CpG sites in the test phase

Target CpG sites with low coverage (< 500 reads) in more than 20% of the samples were excluded from the target panel after the first batch (**Figure 3.3C**). Therefore, ten target CpG sites were excluded in the first batch (the CpG sites nearest the genes *ACACB*, *AGAP1*, *CYP11B1-AS1*, *CYTH3*, *DENND5B*, *GNG4*, *LINC01623*, *MC4R*, *PCCA*, *PLK4* and *SOX1*). New primers were designed for *MC4R*, but better primers were difficult to design for the other assays. Instead new target CpG sites were included in the second sequencing batch to replace the excluded ones. In the second batch the new assay for *MC4R* again generated low-coverage sequencing data, as well as the CpG sites nearest the genes *GPAT3*, *OSBPL3*, *PCNX1*, *RAD51C* and *SSH1* (**Figure 3.3D**). To fill up the target panel, the target CpG sites nearest the genes *OSBPL3*, *GPAT3*, *RAD51C* were included in the target panel (since these assays worked fine in the first batch), as well as *DENND5B* (since this assay had the lowest percentage of samples with low coverage among the rest of the excluded assays). All target CpG sites for the WBC estimation passed quality control. The sequences for all final primer pairs are found in **Appendix 3** and annotation information for the 42 target CpG sites are found in **Table 3.4**.

3.4.2 Quality control of target sequencing data in the Generations Study

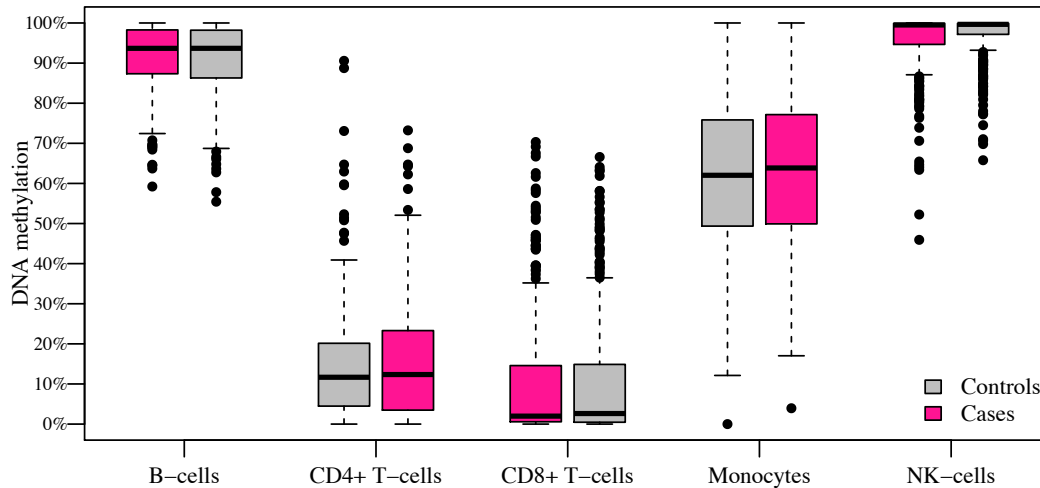
Targeted bisulphite sequencing of 42 target regions using the Fluidigm 48.48 Access Array was conducted in 880 samples (440 matched case-control pairs) from the Generations Study (**Table 3.3**). After quality control of the generated sequencing data, two batches were rerun (batch 12 and 19) due to poor quality, likely a result of decreased performance of the barcodes used (the same barcodes were used in both batches). In the analyses only the sequencing data from the new batches 12 and 19 are used, referred to as sequencing pool 5. The targeted sequencing data was of high quality: high bisulphite conversion efficacy was observed (> 99%), average sequencing depth per CpG site per sample was 1740, more than 97% of the reads were assigned to a sample (i.e. included a barcode sequence) and more than 99% of the paired reads were aligned to target regions (**Table 3.5**). Furthermore, although some variation between DNA methylation levels was observed, a high

Table 3.5. Summary of the targeted sequencing data in the Generations Study

Targeted bisulphite sequencing was conducted in 880 samples in the Generations Study, divided in 20 batches and four sequencing pools. Two batches were rerun due to insufficient data, in a fifth sequencing pool. The table provides the total number of reads before pre-processing, reads to samples (i.e. reads that included barcode sequences), undetermined reads (with no barcode sequence), coverage observed in and outside target regions, and bisulphite conversion efficacy calculated as average bisulphite-converted cytosines in a CHG and CHH context (where H = A, T or C).

	SeqPool1 (n=288) Batches 1 - 6	SeqPool2 (n=240) Batches 7 - 11	SeqPool3 (n=240) Batches 12 - 16	SeqPool4 (n=192) Batches 17 - 20	SeqPool5 (n=96) Batches 12 and 19
Total nr of reads	41,630,730	35,243,020	37,343,220	41,851,010	33,058,310
Undetermined reads	733,686 (1.8%)	818,244 (2.3%)	838,572 (2.3%)	894,656 (2.1%)	646,478 (2.0%)
Reads to samples	40,897,040 (98.2%)	34,424,770 (97.7%)	36,504,650 (97.8%)	40,956,350 (98.9%)	32,411,830 (98.0%)
Coverage in targets*	1,254 (99.8%)	1,242 (99.8%)	1,311 (99.8%)	1,754 (99.7%)	3,137 (99.8%)
Coverage outside target*	2.8 (0.2%)	2.6 (0.2%)	2.7 (0.2%)	5.5 (0.3%)	5.1 (0.2%)
Bisulphite conversion efficacy	99.70%	99.40%	99.30%	99.50%	99.10%

* Coverage inside and outside target regions was calculated by total coverage at each CpG site / frequency (observation in number of samples) / total number of CpG sites

**Figure 3.4. Estimation of WBC composition in the Generations Study targeted sequencing data**

Five CpG sites from the Houseman method were included in the target panel for estimation of WBC counts in the targeted bisulphite sequencing data for B-cells, CD4+ T-cells, CD8+ T-cells, monocytes and natural killer (NK)-cells. The differences between cases and controls were analysed in 328 pairs with coverage > 30 reads for all five CpG sites.

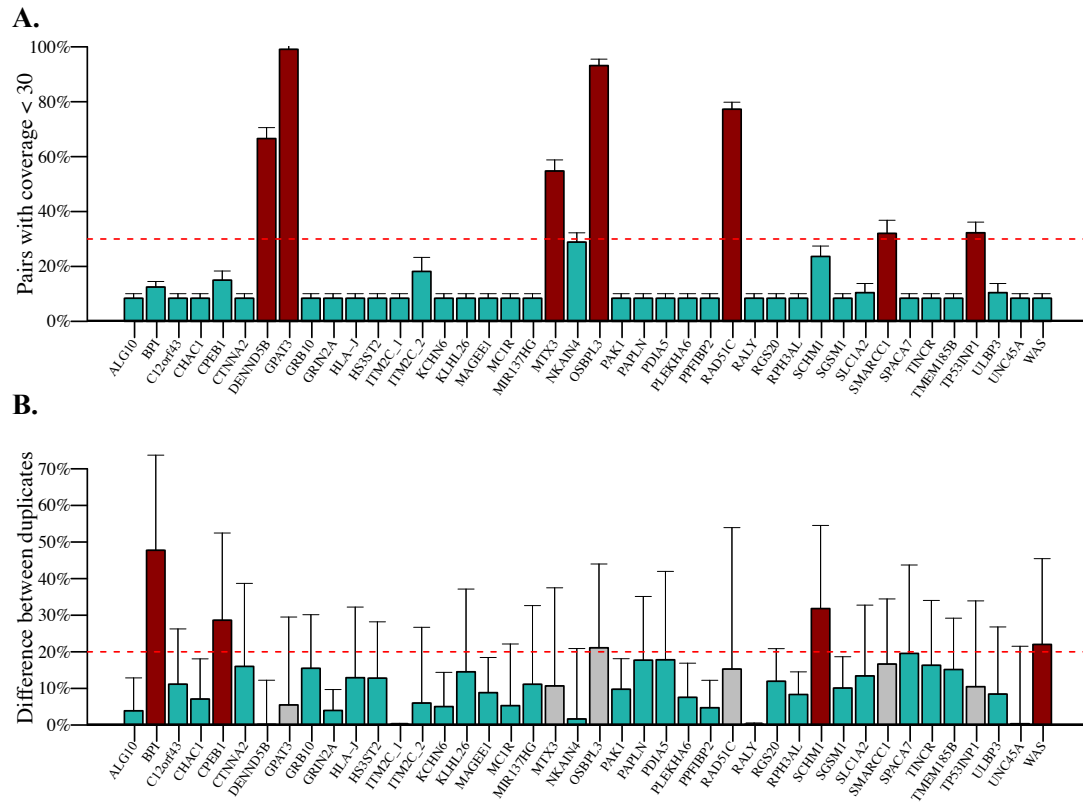


Figure 3.5. Excluded target CpG sites in the Generations Study targeted sequencing data prior to development of the MI

A. Bar plots of the average percentage of pairs in each batch where at least one of the case or control had a coverage < 30 reads. A threshold of 30% was used to filter out target CpG sites with low coverage (red line). **B.** Bar plots of the mean differences in DNA methylation levels between duplicated samples. Target CpG sites already excluded are in grey. A threshold of 20% was used as a threshold to filter out target CpG sites before development of the MI (red line). Red bars correspond to excluded CpG sites.

correlation in DNA methylation values was observed between the 20 batches (mean correlation coefficient $r=0.88$).

3.4.3 No observed differences in WBC composition between cases and controls

Five probes were selected from the Houseman method to estimate the WBC counts for B-cells, CD4+ T-cells, CD8+ T-cells, monocytes and natural killer cells (**Figure 3.2**). All five WBC target CpG sites passed all quality control steps in the Generations Study samples with an average coverage of 230 to 2800 reads per CpG site per sample. A total of 328 pairs had coverage > 30 reads for at least one of the samples in the pair for the five CpG sites and were used to analyse differences in WBC composition between cases and controls. No significant difference ($P > 0.05$) in DNA methylation levels between cases and controls were observed for any of the five WBC target CpG sites (**Figure 3.4**).

3.4.4 Exclusion of target CpG sites and samples in the Generations Study targeted sequencing data prior to MI development

In the Generations Study targeted sequencing data, eleven out of the 42 target CpG sites were excluded prior to the MI analysis due to either low coverage in $> 30\%$ of the pairs or large differences ($> 20\%$) in DNA methylation levels between duplicated pairs (**Figure 3.5**). Out of the 880 samples in the Generations Study targeted sequencing data, one pair was excluded due to incorrect case-control status and 100 pairs were excluded due low coverage (< 30 reads) in more than 10% of the remaining target CpG sites of at least one sample in the pair, leaving a total of 678 samples for the analyses (339 matched case-control pairs, **Table 3.3**).

3.4.5 The MI is correlated with ELEE in EPIC-Italy but not in the Generations Study

The MI to predict ELEE was developed on HM450K data from EPIC-Italy and the Generations Study ($n=237$ and 65, **Table 3.1**) using ridge regression on the 31 target CpG sites passing quality control in the targeted sequencing data (**Table 3.4**). As expected, the MI correlated well with ELEE in the development data; a high correlation was observed in EPIC-Italy and a moderate correlation in the Generations Study HM450K data ($r=0.60$, $P=5.89 \times 10^{-25}$ and $r=0.27$, $P=0.027$, respectively, **Figure 3.6A**). No association between the MI and ELEE was observed in the Generation Study targeted sequencing data ($n=678$, $r=-0.04$, $P=0.340$, **Figure 3.6B**). Furthermore, no significant association between the ELEE and breast cancer risk was observed in the sequenced Generations Study subjects

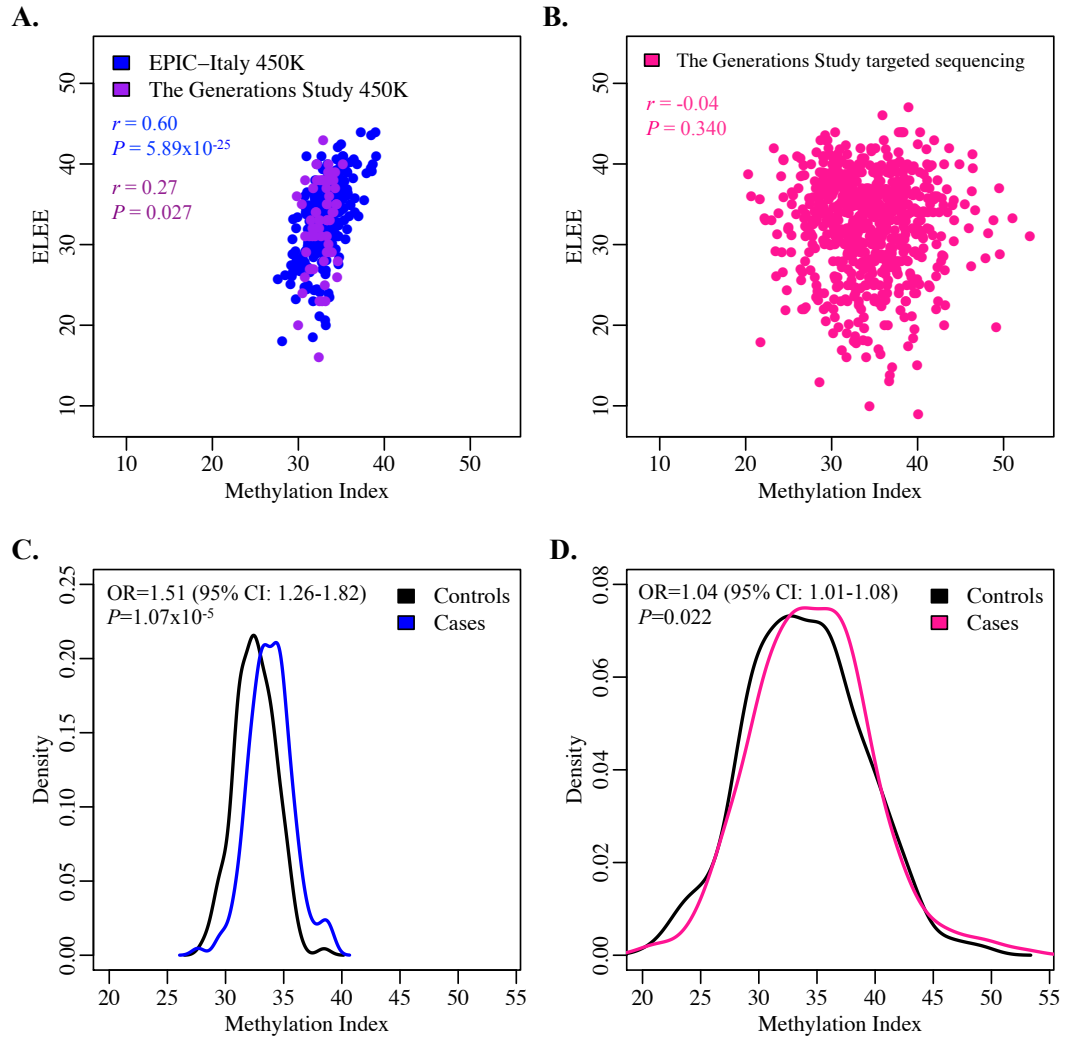


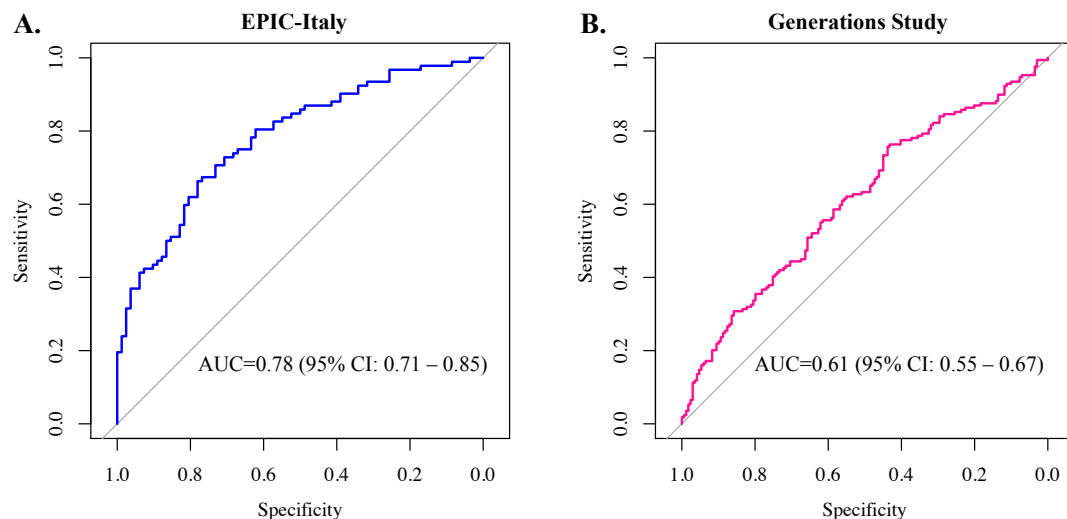
Figure 3.6. The Methylation Index of ELEE

The MI was developed in combined HM450K data from EPIC-Italy (n=237) and the Generations Study (n=65) using ridge regression on 31 target CpG sites that passed quality control. The correlation between the MI and ELEE and the association between the MI and breast cancer risk were evaluated. **A.** The correlations between the MI and ELEE in the development HM450K data were as following: $r=0.60$, $P=5.89 \times 10^{-25}$ for EPIC-Italy and $r=0.27$, $P=0.027$ for the Generations Study. **B.** The MI and ELEE were not correlated in the Generations Study targeted sequencing data ($r=-0.04$, $P=0.340$). **C.** Density plot of the MI values in matched controls and cases in EPIC-Italy HM450K data (n=162 pairs). The MI was associated with breast cancer risk in EPIC-Italy (OR=1.51, 95% CI: 1.26 – 1.82, $P=1.07 \times 10^{-5}$). **D.** Density plot of the MI values in matched controls and cases in the Generations Study targeted sequencing data (n=339 pairs). The MI was associated with breast cancer risk in the Generations Study (OR=1.04, 95% CI: 1.01 – 1.08, $P=0.022$). ORs were adjusted for age, BMI, alcohol consumption, smoking duration and WBC composition.

Table 3.6. The MI by quartile and breast cancer risk

Analysis by control-quartiles of the MI and risk of breast cancer in matched case-controls pairs in EPIC-Italy HM450K data, which the model partly was developed in (n=162 pairs), and in the Generations Study targeted sequencing data for validation (n=339 pairs). The association with breast cancer risk was analysed using conditional logistic regression adjusted for age, BMI, alcohol consumption, smoking duration and WBC composition in the multivariable model.

Development data: EPIC-Italy (n=162 pairs)							
Quartile	Range	Controls	Cases	Unadjusted model		Multivariable model	
				OR (95% CI)	P value	OR (95% CI)	P value
Q1	27.6 – 31.3	41	11	Ref	Ref	Ref	Ref
Q2	31.3 – 32.5	40	24	2.00 (0.85 – 4.70)	0.111	1.87 (0.71 – 4.94)	0.203
Q3	32.5 – 33.8	40	46	3.37 (1.54 – 7.40)	0.002	4.14 (1.66 – 10.33)	0.002
Q4	33.8 – 39.1	41	81	5.88 (2.72 – 12.68)	6x10⁻⁶	5.45 (2.17 – 13.67)	3x10⁻⁴
Validation data: The Generations Study (n=339 pairs)							
Quartile	Range	Controls	Cases	Unadjusted model		Multivariable model	
				OR (95% CI)	P value	OR (95% CI)	P value
Q1	20.3 – 30.3	84	66	Ref	Ref	Ref	Ref
Q2	30.3 – 33.8	85	88	1.33 (0.86 – 2.05)	0.193	1.52 (0.95 – 2.42)	0.083
Q3	33.8 – 37.0	85	82	1.26 (0.80 – 2.01)	0.320	1.60 (0.96 – 2.64)	0.069
Q4	37.0 – 53.1	85	103	1.63 (1.03 – 2.60)	0.038	1.77 (1.07 – 2.93)	0.027

**Figure 3.7. Receiver operating characteristic curves for the MI and breast cancer risk**

The discrimination for the highest and lowest quartiles of the MI in EPIC-Italy and the Generations Study. The estimated AUC values were 0.78 in EPIC-Italy (95% CI: 0.71 – 0.85) and 0.61 in the Generations Study (95% CI: 0.55 – 0.67).

(n=339 matched case-control pairs, OR=1.01, 95% CI: 0.98 – 1.04, $P=0.562$), in contrast to EPIC-Italy subjects (n=162 matched case-control pairs, OR=1.10, 95% CI: 1.03 – 1.17, $P=0.007$).

3.4.6 The MI is associated with breast cancer risk

The association between the MI and breast cancer risk was tested in matched case-control pairs in EPIC-Italy HM450K data and the Generations Study targeted sequencing data (n=162 and 339 pairs respectively, **Table 3.2** and **Table 3.3**). The MI, ranging from 27.6 to 39.1, showed a 51% increase in breast cancer risk per unit increase in MI in EPIC-Italy in a multivariable model (OR=1.51, 95% CI: 1.28 – 1.82, $P=1.07 \times 10^{-5}$, **Figure 3.6C**). In the Generations Study a 4% increase in risk per unit increase in MI, ranging from 20.3 to 53.1, was observed (OR=1.04, 95% CI: 1.01 – 1.08, $P=0.022$, **Figure 3.6D**). To explore a non-linear association between the MI and breast cancer risk the MI was divided into control-quartiles. A significantly higher breast cancer risk was observed for women in the highest quartile compared to women in the lowest quartile in EPIC-Italy (OR_{Q4_vs_Q1}=5.45, 95% CI: 2.17 – 13.67, $P=3 \times 10^{-4}$) and the Generations Study (OR_{Q4_vs_Q1}=1.77, 95% CI: 1.07 – 2.93, $P=0.027$, **Table 3.6**). The AUC values for the prediction between the highest and lowest quartiles were estimated to 0.78 in EPIC-Italy (95% CI: 0.71 – 0.85) and 0.61 in the Generations Study (95% CI: 0.55 – 0.67) (**Figure 3.7**).

To explore if the inclusion of incident breast cancer cases in the model development dataset affected the results, the model development was also conducted in controls-only in the combined EPIC-Italy and the Generations Study HM450K data (**Table 3.1**, total n=184). The control-only MI model showed very similar results as previously and a significant association with breast cancer risk in the Generations Study targeted sequencing data was observed (n=339 pairs, OR=1.04, 95% CI: 1.01 – 1.08, $P=0.012$, **Figure 3.8**).

The MI was calculated in the pooled breast- and ovarian cancer cases and control samples included in the test phase. For all breast cancer pools, there was a 2.26% increase in the MI for pooled cases compared to pooled controls (**Table 3.7**). Stratifying by subgroup, an increase in the MI was observed for pooled cases diagnosed with DCIS, ER-negative disease, and ER-positive disease with age at diagnosis over 50, but not for ER-positive cases with age at diagnosis before 50.

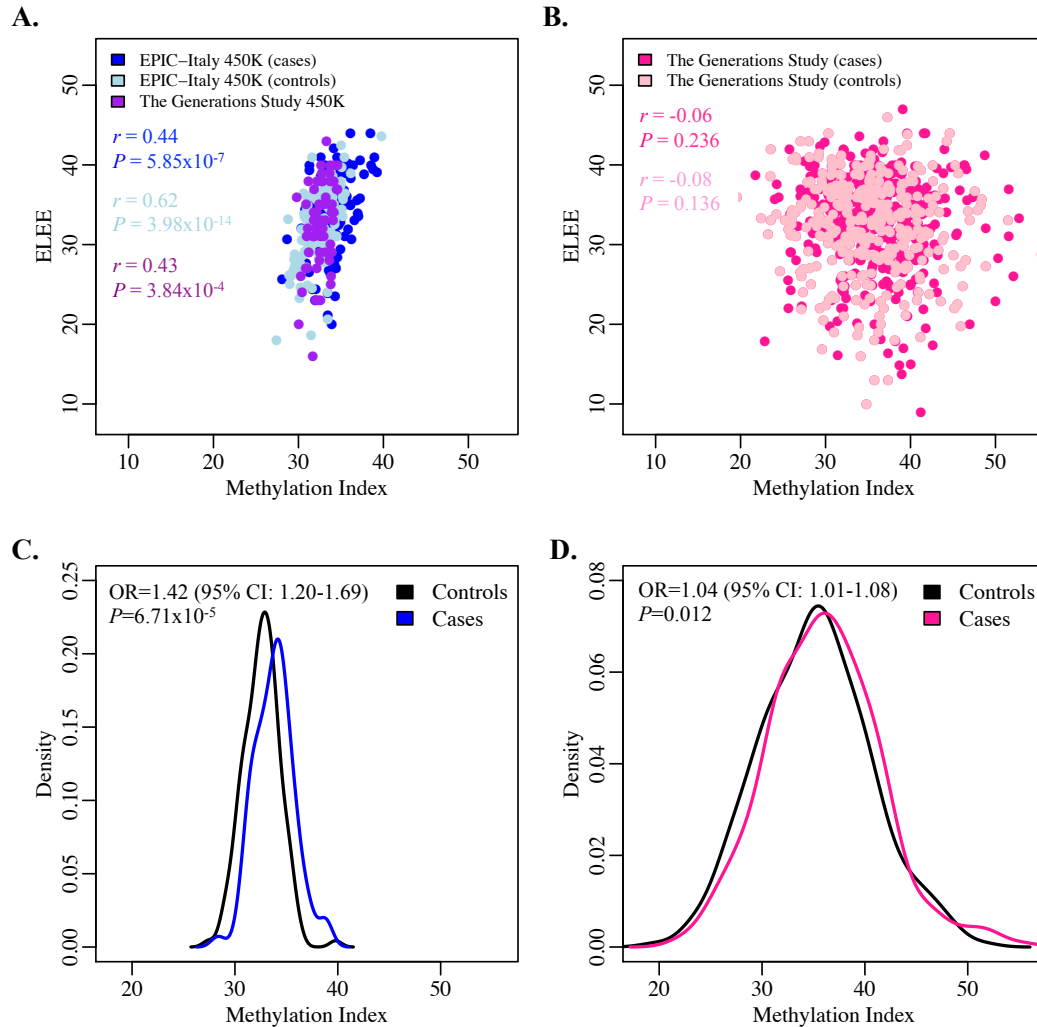


Figure 3.8. Development of the MI in controls-only

To explore if including incident cases in the model development of the MI affected the results, model development was conducted in controls only in the combined EPIC-Italy and the Generations Study HM450K data (n=184 in total). **A.** The correlations between the controls-only-MI and ELEE in the development data were as following: $r=0.62$, $P=3.98 \times 10^{-14}$ for EPIC-Italy controls (n=119), $r=0.44$, $P=6.85 \times 10^{-7}$ for EPIC-Italy cases (n=118), and $r=0.43$, $P=3.84 \times 10^{-4}$ for the Generations Study (n=65). **B.** The controls-only-MI and ELEE were not correlated in the Generations Study controls or cases ($r=-0.08$, $P=0.136$ and $r=-0.07$, $P=0.063$). **C.** Density plot of the controls-only-MI values in matched cases and controls in the EPIC-Italy HM450K data (n=162 pairs). The controls-only-MI was associated with breast cancer risk in EPIC-Italy (OR=1.42, 95% CI: 1.20 – 1.69, $P=6.71 \times 10^{-5}$). **D.** Density plot of the controls-only-MI values in matched controls and cases in the Generations Study targeted sequencing data (n=339 pairs). The MI was associated with breast cancer risk in the Generations Study (OR=1.04, 95% CI: 1.01 – 1.08, $P=0.012$). ORs were adjusted for age, BMI, alcohol consumption, smoking duration and WBC composition.

Table 3.7. The MI in pooled breast and ovarian cancer cases and controls

The MI was calculated in pooled cases and controls (included in the test phase), and compared between the cases and controls.

Group	n pairs	MI value Cases	MI value Controls	Difference
All breast cancer pools	548	36.50	34.24	2.26
DCIS	123	35.32	32.52	2.80
ER-negative	66	39.28	36.63	2.66
ER-positive, age at diagnosis < 50	176	34.17	36.45	-2.29
ER-positive, age at diagnosis > 50	183	36.51	32.07	4.44
Ovarian cancer pools	78	37.89	35.52	2.38

Table 3.8. The MI, ELEE and breast cancer risk in pre- and postmenopausal women

The correlation between the MI and ELEE, and the association between the MI and breast cancer risk in pre- and postmenopausal women. The ORs were adjusted for age, BMI, alcohol consumption, smoking duration and WBC composition.

	MI and ELEE			MI and breast cancer risk		
	n subjects	r	P value	n pairs	OR (95% CI)	P value
Premenopausal women						
EPIC-Italy HM450K	79	0.53	7x10⁻⁷	45	1.61 (1.17 – 2.22)*	0.004*
The Generations Study HM450K	32	0.36	0.04			
The Generations Study targeted sequencing	262	-0.06	0.315	111	1.01 (0.94 – 1.09)	0.713
Postmenopausal women						
EPIC-Italy HM450K	132	0.72	2x10⁻²²	76	1.91 (1.29 – 2.82)	0.001
The Generations Study HM450K	33	0.20	0.270			
The Generations Study targeted sequencing	416	-0.01	0.840	188	1.07 (1.02 – 1.12)	0.006

* unadjusted model for EPIC-Italy premenopausal women due to too few pairs

3.4.7 The MI in pre- and postmenopausal women

The association between the MI and ELEE and MI and breast cancer risk was stratified by menopausal status. The correlations between the MI and ELEE were similar in both pre- and postmenopausal women in the Generations Study targeted sequencing data, but stronger for postmenopausal women in EPIC-Italy ($r=0.72$, $P=2.34 \times 10^{-22}$, **Table 3.8**). No heterogeneity was observed for the association between MI and ELEE in pre- and postmenopausal in either EPIC-Italy ($Q=0.79$, $I^2=0$) or the Generations Study ($Q=0.42$, $I^2=0$). The association between the MI and breast cancer risk appeared to be stronger in postmenopausal women in both EPIC-Italy and the Generations Study (OR=1.91, 95% CI: 1.29 – 2.82, $P=0.001$ and OR=1.07, 95% CI: 1.02 – 1.12, $P=0.006$, respectively, **Table 3.8**). However, the test for heterogeneity between pre- and postmenopausal women in EPIC-Italy and Generations Study suggested no significant heterogeneity ($Q=0.19$, $I^2=43$ and $Q=0.24$, $I^2=27$, respectively).

3.4.8 The association between the MI and breast cancer risk is stronger in cases with longer time to diagnosis

In matched case-control pairs in EPIC-Italy HM450K data and the Generations Study targeted sequencing data, the pairs were stratified into two groups split by median time to diagnosis (3.8 years in EPIC-Italy and 4 years in Generations Study). In both study cohorts the association between the MI and breast cancer risk appeared to be stronger for pairs with a longer time to diagnosis. In EPIC-Italy the association between the MI and breast cancer risk was significant in both groups but with a higher OR in pairs with a longer time to diagnosis ($n=81$ pairs in both groups, OR=1.35 vs 2.24, **Figure 3.9A-9B**). In the Generations Study, for pairs with shorter time to diagnosis there was no significant association with breast cancer risk ($n=149$ pairs, OR=1.03, 95% CI: 0.97 – 1.09, $P=0.315$, **Figure 3.9C**), however, the MI was significantly associated with breast cancer risk in pairs with time to diagnosis above the median ($n=190$ pairs, OR=1.06, 95% CI: 1.01 – 1.12, $P=0.024$, **Figure 3.9D**).

3.4.9 Individual CpG sites and association with breast cancer

The association between each CpG sites included in the target regions ($n=274$) and breast cancer risk was explored in the Generations Study targeted sequencing data (**Appendix 4**). A total of 14 CpG sites showed association with breast cancer risk with a P -value < 0.05 , but no CpG sites remained

significant after correction for multiple testing (all FDR $Q > 0.05$). Out of these, four CpG sites were targets in the sequencing (CpG sites nearest the genes *CTNNA2*, *GRB10*, *RPH3AL* and *TINCR*). The four target CpG sites were associated with breast cancer risk in both the Generations Study and EPIC-Italy ($P < 0.05$, **Table 3.9**). Hypomethylation was observed for all four CpG sites except for the CpG site nearest *GRB10* where hypermethylation was observed in the Generations Study and hypomethylation in EPIC-Italy.

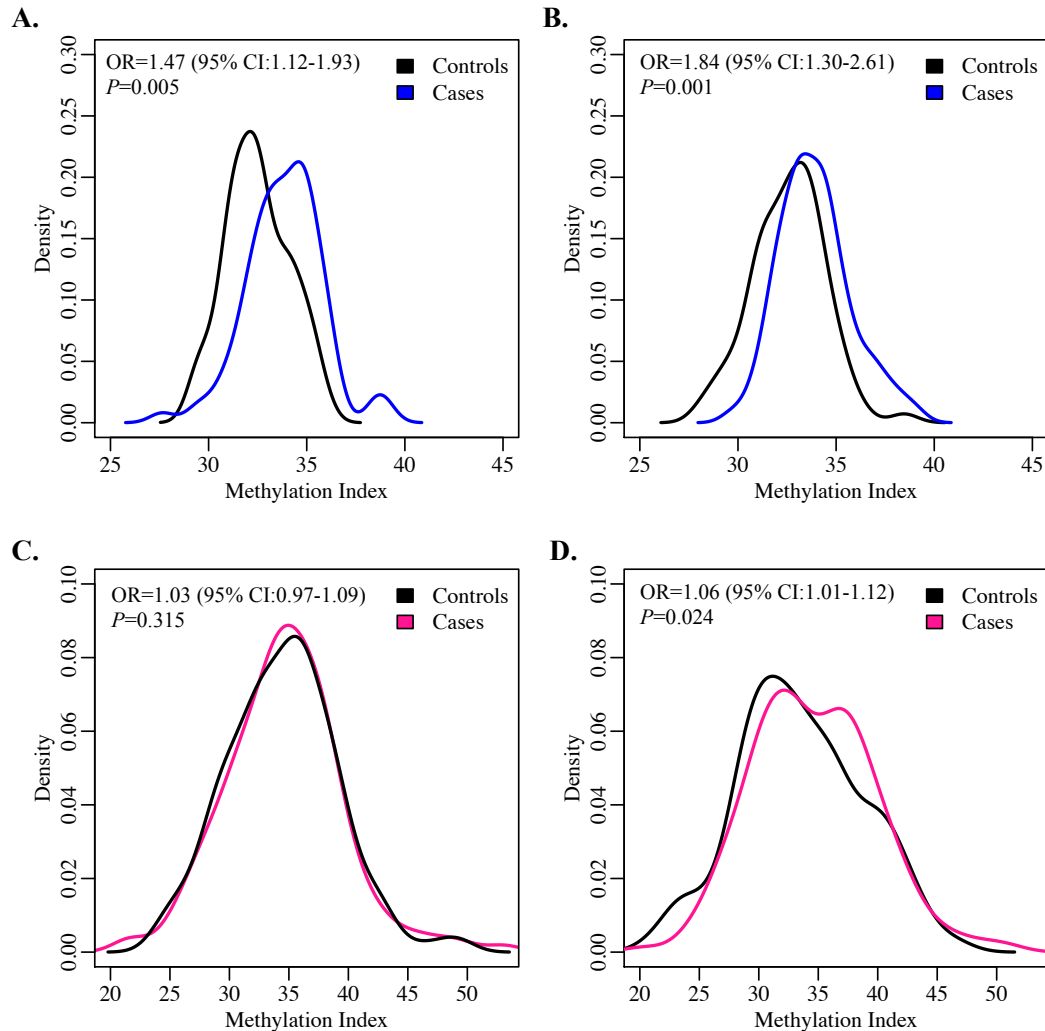


Figure 3.9. The MI, breast cancer risk and time to diagnosis

The association between the MI and breast cancer risk was analysed in two groups in matched case-control pairs stratified on median time to diagnosis in EPIC-Italy HM450K data and the Generations Study targeted sequencing data. **A.** The MI was associated with breast cancer risk in EPIC-Italy pairs with a shorter time to diagnosis ($n=81$ pairs, OR=1.47, 95% CI: 1.12 – 1.93, $P=0.005$). **B.** The MI was associated with breast cancer risk in EPIC-Italy pairs with a longer time to diagnosis ($n=81$ pairs, OR=1.84, 95% CI: 1.30 – 2.61, $P=0.001$). **C.** In the Generations Study, the MI was not associated with breast cancer risk in the 149 pairs with shorter time to diagnosis (OR=1.03, 95% CI: 0.97 – 1.09, $P=0.315$). **D.** The MI was associated with breast risk in the 190 pairs with a longer time to diagnosis (OR=1.06, 95% CI: 1.01 – 1.12, $P=0.024$). ORs were adjusted for age, BMI, alcohol consumption, smoking duration and WBC composition.

Table 3.9. Target CpG sites associated with breast cancer risk

The target CpG sites associated with breast cancer risk (P -value < 0.05) in matched case-control pairs in both the Generations Study targeted sequencing data and EPIC-Italy HM450K data. The methylation values were converted to z-scores to estimate OR per standard deviation (ORperSD). The ORs were adjusted for age, BMI, alcohol consumption, and smoking duration at recruitment and WBC composition.

HM450K probe	Nearest gene	The Generations Study targeted sequencing data						EPIC-Italy Illumina 450K data					
		n pairs	Meth cases	Meth controls	Difference	ORperSD (95% CI)	P value	n pairs	Meth cases	Meth controls	Difference	ORperSD (95% CI)	P value
cg15907392	CTNNA2	360	74.5	78.9	-4.4	0.80 (0.67 - 0.95)	0.012	162	73.7	75.6	-1.9	0.68 (0.53 - 0.87)	0.002
cg10298859	GRB10	409	80.4	78.1	2.4	1.23 (1.02 - 1.49)	0.033	162	54.5	55.4	-0.9	0.62 (0.47 - 0.82)	0.001
cg12091786	RPH3AL	417	85.7	86.6	-1.0	0.83 (0.69 - 0.99)	0.036	162	92.4	93.5	-1.1	0.70 (0.54 - 0.91)	0.007
cg26657235	TINCR	413	47.6	50.1	-2.4	0.77 (0.64 - 0.93)	0.006	162	47.2	48.2	-1.0	0.79 (0.62 - 1.00)	0.048

3.5. DISCUSSION

In this chapter it was aimed to further explore the associations between ELEE, DNA methylation in blood, and breast cancer risk. Targeted bisulphite sequencing was performed on 42 CpG sites selected from the results in the EWAS of ELEE (**Chapter 2**) in a large independent study cohort (the Generations Study, n=880). A MI model of ELEE was developed using HM450K methylation levels at 31 CpG sites and tested for association with breast cancer risk, using the Generations Study targeted sequencing data for validation. The observed association between the MI and breast cancer risk in EPIC-Italy was validated in the Generations Study targeted sequencing data, but not the correlation between the MI and ELEE. Women in the highest quartile of the MI in the Generations Study had 77% higher risk for breast cancer compared with women in the lowest quartile and an estimated AUC of 0.61.

There are several explanations of why the correlation between the MI and ELEE was not seen in the Generations Study. It is possible that the MI model is over-fitted in the EPIC-Italy data, and additional study cohorts would be needed to assess this further. Alternatively, population specific differences between the Italian and UK based cohorts or technical differences between the HM450K and the targeted bisulphite sequencing data cannot be excluded. Importantly, the ELEE model itself is not significantly associated with breast cancer risk in the Generations Study sub-cohort analysed in this study, which is an important assumption in the analyses. Of the lifestyle factors, the Generations Study had higher alcohol consumption, less smokers, more number of pregnancies, longer breastfeeding duration, and considerably more reported users of OC and HRT, which may all significantly alter the epigenome and oestrogen exposure. Also, different breast cancer screening procedure between Italy and the UK might have affected how early the cancer was detected (time to diagnosis was longer in EPIC-Italy with an average time to diagnosis of 8.5 years compared to 4.0 years in the Generations Study).

There was a shorter average time to diagnosis for women in the Generations Study (mean value 4 years compared with 5.3 years in EPIC-Italy). However, the distribution of time to diagnosis was left-skewed (17% of the women in EPIC-Italy had a time to diagnosis ≤ 2 years) and the distribution in the Generations Study was mostly uniform. There was also a difference in follow-up time in the both study cohorts (up to 16 years for EPIC-Italy, and 10 years for the Generations Study).

A stronger association between the MI and breast cancer risk was observed for cases with a longer time to diagnosis in both cohorts. Therefore, with a longer follow-up time, the association might be stronger in the Generations Study overall. Why the association between the MI and breast cancer risk was stronger in pairs where cases had a longer time to diagnosis needs to be explored further. One possibility is that it might be due to changes in methylation signatures occurring during early stage cancer development (leading up to diagnosis) that might be adversely affecting the methylation of CpG sites that contribute to the MI. This would suggest that the increase in MI might not be due to the reverse effect from early cancer development, but rather that early cancer development adversely affects the MI.

Advantages in this study include the validation in a large independent study cohort using targeted bisulphite sequencing and the inclusion of an extended test phase prior to the sequencing in the Generations Study. Targeted bisulphite sequencing using the Fluidigm 48.48 Access Array gave the opportunity to explore a larger number of CpG sites in comparison to bisulphite pyrosequencing. Theoretically, 10-15 CpG sites could have been analysed at the same time using pyrosequencing. Since the protocol and workflow had not been used in our laboratory previously, a test phase was included before sequencing of the 880 samples in the Generations Study. This probably reduced errors and mistakes in the procedure and improved the experimental design, by the opportunity to validate the target assays before sequencing in the Generations Study. The targeted bisulphite sequencing generated high-throughput data with high coverage for most of the target regions. Even so, 11 out of the 42 target CpG sites did not yield sufficient sequencing data to be included in the MI analysis. This could be due to pipetting error or bubbles in the wells when loading the Access Array, or other random events. In the MI analysis, sufficient data was produced for 31 target CpG sites in 339 matched case-control pairs.

A total of 274 CpG sites were included in the target regions. Fourteen of these CpG sites showed modest association with breast cancer risk in the Generations Study and six overlapped a HM450K probe. Of these, four CpG sites showed association ($P < 0.05$) with breast cancer risk in EPIC-Italy as well. For future breast cancer risk studies, these would be the CpG sites to focus on.

Higher MI values in pooled breast cancer cases compared to controls were observed. When the pools were stratified into subtypes, the largest differences between cases and controls were observed for ER-positive cases with age at diagnosis > 50 , but a decrease in the MI was observed for ER-positive cases with age at diagnosis < 50 . Hypothetically, there might be a shorter time to diagnosis

from recruitment for younger cases than older cases, which would confirm the previous findings for time to diagnosis and the association with breast cancer risk. However, this data is missing for these pooled samples so it cannot be confirmed. Furthermore, a higher MI value for ovarian cancer cases compared to controls was observed. ELEE as an exposure could be important in ovarian cancer as well [354, 355], however, this is out of the scope for this study.

CpG sites to estimate WBC composition were included in the targeted bisulphite sequencing data and no significant differences for any of the five WBC types included were observed. Even though no strong difference between cases and controls was observed, it might be necessary to adjust for WBC composition since this can be affected by other exposures and confound the results [329, 356-358].

3.5.1 Conclusions

In this chapter a DNA methylation signature in blood associated with breast cancer risk was identified and validated in 339 matched case-control pairs in the Generations Study targeted sequencing data. However, an association between the MI and ELEE was not observed in the Generations Study. Further studies are needed to clarify this and to explore the association between the MI and breast cancer risk in other study cohorts.

CHAPTER 4:

Meta-analysis of the MI and ELEE

4 CHAPTER 4: Meta-analysis of the MI and ELEE

4.1. INTRODUCTION

In the two previous chapters an EWAS of ELEE was conducted and DNA methylation levels at 694 CpG sites associated with ELEE were identified. From this, 42 CpG sites were selected for validation using targeted bisulphite sequencing in 880 independent samples, whereof 31 CpG sites and 678 samples passed quality control. Using HM450K data, a MI of ELEE was developed including DNA methylation levels at the 31 CpG sites. The MI was associated with breast cancer risk in matched case-control pairs in the discovery study cohort EPIC-Italy ($OR_{Q4_vs_Q1}=5.45$, 95% CI: 2.17 – 13.6, $P=3.00 \times 10^{-4}$) and this was validated in the Generations Study ($OR_{Q4_vs_Q1}=1.77$, 95% CI: 1.07 – 2.93, $P=0.027$). In this chapter a larger meta-analysis of ELEE, the MI, and breast cancer risk was performed including three new datasets.

4.1.1 Meta-analysis

Meta-analysis is a systematic approach to combine effect estimates from observational studies of the same exposure and outcome and obtain a summary estimate of the effect. This can be particularly statistically challenging for methylation studies due to the often small sample sizes, study cohort bias, study specific inflation and differences in study design [359]. Two main approaches are used for meta-analysis: fixed- or random-effects models. Fixed-effects models assume that all studies are estimating the same effect size, meaning that there is no heterogeneity in the true effect and that the observed variances are only due to sampling error [360]. Random-effects models allow for true differences in effects between studies, which can be due to true population differences, for example different age distributions or other characteristics. Random-effects models are designed to handle this heterogeneity between studies [360, 361]. Briefly, fixed-effects models test under the null hypothesis that all effect sizes are equal to zero and random-effects model that the mean effect size is equal to zero [362].

Statistical heterogeneity in effects between studies can be estimated by a number of methods, for example Cochran's Q -statistics and the inter-study-variance (often denoted I^2) [361, 363]. The I^2 statistic measures the percentage of variability in effect that is due to true heterogeneity between studies and not by chance [360, 362, 363]. This has often been calculated by the commonly used

DerSimonian & Laird estimator [364], but has recently been replaced by different likelihood techniques, for example the restricted maximum likelihood method. This method has been shown to be more efficient, with lower bias and good performance for both larger and smaller number of studies included in the meta-analysis [361, 362, 365, 366]. When the heterogeneity estimator I^2 is equal to zero the restricted maximum likelihood method behaves as a fixed-effects model [362].

4.1.2 Aims for this chapter

In this chapter it was aimed to perform a meta-analysis for the association between the MI (developed in **Chapter 3**) and breast cancer risk, as well as explore the association between the MI and ELEE model (from **Chapter 2**), in four different nested case-control study cohorts: the Generations Study, EPIC-Italy, EPIC-IARC and the Melbourne Collaborative Cohort Study (MCCS). The study was conducted by sharing the R-script for the analysis (**Appendix 5**) to each of the study co-ordinators, and then compiling the results in a meta-analysis.

4.2. STUDY COHORTS

4.2.1 The Generations Study

Targeted bisulphite sequencing including CpG sites in the MI was conducted on 880 samples from the Generations Study, whereof 678 passed quality control (**Table 4.1**). The dataset has been described previously in **Chapter 3**.

4.2.2 EPIC-Italy

EPIC-Italy has been described previously in **Chapter 2** and **Chapter 3**. HM450K methylation data for new breast cancer cases and controls have been generated and 236 (118 case-control pairs) subjects that had not been used previously in this study were available for the meta-analysis (**Table 4.1**). Case-control pairs were matched on age $\pm$ 5 years and study centre.

4.2.3 EPIC-IARC

Participants from the EPIC cohort from IARC (International Agency for Research on Cancer) were selected for a nested case-control study. This included subjects from Germany, Greece, Italy, Spain, The Netherlands and UK [206]. Subjects overlapping with EPIC-Italy were excluded prior to the analysis. Questionnaire data, anthropometric measurements and blood samples were collected for all study participants at time at recruitment between 1993 and 1998, after ethical review by the IARC. The cancer diagnosis was updated from cancer registries for Italy, Spain, the Netherlands and UK, and by active follow-up for Germany and Greece. The case-control pairs were matched on centre, age $\pm$ 2 years, time of blood collection, fasting status, menopausal status, current pill and hormone replacement therapy use and menstrual cycle. A total of 840 (420 case-control pairs) subjects were included in the meta-analysis (**Table 4.1**). The EPIC-IARC analysis was conducted by Dr Srikant Ambatipudi at IARC.

4.2.4 MCCS

Participants from the MCCS were selected for a nested case-control study. The MCCS is a prospective cohort study including 24,469 women recruited between 1990 and 1994 [204]. Cancer incidences were identified via the Victorian cancer registry (VCR) and the Australian Cancer Database, and a total of 680 cases (2.8%) were diagnosed with invasive breast cancer between enrolment and before 2008. In

the meta-analysis, 310 cases-control pairs are included, matched on year of birth, year of baseline attendance, and country of origin (**Table 4.1**). The DNA was extracted from dried blood spots. The MCCS analysis was conducted by Dr Laura Baglietto at University of Pisa.

Table 4.1. Table of characteristics for the meta-analysis
Subject characteristics at recruitment for all women across the four study cohorts included in the meta-analysis of MI and breast cancer risk.

	The Generations Study			EPIC-Italy (new)			EPIC-IARC			MCCS		
	All (n=678)	Cases (n=339)	Controls (n=339)	All (n=236)	Cases (n=118)	Controls (n=118)	All (n=870)	Cases (n=435)	Controls (n=435)	All (n=620)	Cases (n=310)	Controls (n=310)
Age												
mean (st.dev.), yrs	54.0 (10.3)	53.9 (10.3)	54.1 (10.4)	51.2 (7.1)	51.3 (7.3)	51.1 (7.0)	52.2 (9.0)	52.2 (9.0)	52.2 (9.0)	56.7 (8.1)	56.7 (8.2)	56.6 (8.1)
Time to diagnosis												
mean (st.dev.), yrs	NA	4.0 (2.4)	NA	8.5 (3.7)	NA	NA	NA	7.5 (3.2)	NA	NA	7.9 (3.8)	NA
Menopausal status												
n (%)												
Premenopausal	262 (38.6%)	135 (39.8%)	127 (37.5%)	113 (47.9%)	61 (51.7%)	52 (44.1%)	364 (41.8%)	182 (41.8%)	182 (41.8%)	267 (43.1%)	131 (42.3%)	136 (43.9%)
Postmenopausal	416 (61.4%)	204 (60.2%)	212 (62.5%)	123 (52.1%)	57 (48.3%)	66 (55.9%)	506 (58.2%)	253 (58.2%)	253 (58.2%)	353 (56.9%)	179 (57.7%)	174 (56.1%)
Age at menarche												
mean (st.dev.), yrs	12.7 (1.5)	12.7 (1.5)	12.7 (1.5)	12.6 (1.5)	12.4 (1.4)	12.8 (1.5)	13.1 (1.6)	13.0 (1.7)	13.2 (1.6)	13.1 (1.6)	13.1 (1.6)	13.1 (1.5)
Age at menopause												
mean (st.dev.), yrs	50.2 (4.4)	50.3 (4.3)	50.1 (4.5)	47.8 (4.8)	47.0 (5.0)	48.5 (4.6)	48.8 (4.9)	49.1 (4.7)	48.5 (5.1)	49.7 (4.2)	50.3 (3.8)	49.2 (4.6)
Number of pregnancies												
mean (st.dev.)	1.9 (1.2)	1.9 (1.2)	2.0 (1.2)	1.9 (0.9)	1.9 (1.0)	1.9 (0.7)	1.6 (1.1)	1.5 (1.1)	1.6 (1.2)	2.4 (1.7)	2.5 (1.8)	2.40 (1.5)
Ever breastfed												
n (%)	539 (79.5%)	265 (78.2%)	274 (80.8%)	180 (76.3%)	88 (74.6%)	92 (78.0%)	573 (68.2%)	283 (65.0%)	290 (66.6%)	481 (77.6%)	241 (77.7%)	240 (77.4%)
Breastfeeding duration												
mean (st.dev.), yrs	0.9 (1.0)	0.9 (1.0)	0.8 (0.9)	0.6 (0.7)	0.7 (0.8)	0.6 (0.6)	0.6 (0.9)	0.7 (0.9)	0.6 (0.8)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
BMI												
mean (st.dev.), kg/m ²	25.4 (4.3)	25.7 (4.3)	25.2 (4.3)	26.4 (4.1)	26.3 (4.7)	26.4 (5.0)	25.8 (4.5)	26.0 (4.8)	25.6 (4.2)	26.8 (4.7)	27.2 (5.2)	26.3 (4.0)
Alcohol consumption												
mean (st.dev.), *	13.3 (14.1)	13.8 (14.7)	12.8 (13.5)	3.0 (1.2)	3.0 (1.0)	3.1 (1.3)	9.1 (12.5)	9.9 (13.0)	8.2 (11.9)	7.4 (12.5)	6.8 (10.5)	8.0 (14.3)
Smoking status												
n (%)												
Smoker	48 (7.1%)	25 (7.4%)	23 (6.8%)	58 (24.6%)	27 (22.9%)	31 (26.3%)	189 (21.7%)	90 (20.7%)	99 (22.8%)	42 (6.8%)	21 (6.8%)	21 (6.8%)
Former	184 (27.1%)	98 (28.9%)	86 (25.4%)	44 (18.6%)	26 (22.0%)	18 (15.3%)	186 (21.4%)	95 (21.8%)	91 (20.9%)	152 (24.5%)	70 (22.6%)	82 (26.5%)
Never	446 (65.8%)	216 (63.7%)	230 (67.9%)	134 (56.8%)	65 (55.1%)	69 (58.5%)	491 (56.4%)	249 (57.2%)	242 (55.6%)	426 (68.7%)	219 (70.6%)	207 (66.8%)
Smoking duration												
mean (st.dev.), yrs	4.2 (9.0)	4.4 (9.0)	3.9 (9.1)	10.5 (14.0)	10.5 (13.8)	10.5 (14.2)	9.5 (13.3)	9.4 (13.2)	9.6 (13.4)	8.0 (13.9)	7.5 (13.5)	8.5 (14.3)
OC ever												
n (%)	544 (80.2%)	270 (79.7%)	265 (78.2%)	88 (37.3%)	43 (36.4%)	45 (38.1%)	474 (54.5%)	237 (54.5%)	237 (54.5%)	348 (56.1%)	163 (52.5%)	185 (59.7%)
OC duration												
mean (st.dev.), yrs	8.4 (6.6)	8.1 (6.4)	8.8 (6.8)	4.1 (3.7)	4.5 (4.1)	3.6 (3.3)	NA	NA	NA	6.9 (6.5)	6.9 (6.6)	6.9 (6.4)
HRT ever												
n (%)	215 (31.7%)	106 (31.3%)	107 (31.6%)	46 (19.5%)	22 (18.6%)	24 (20.3%)	223 (25.6%)	112 (25.7%)	111 (25.5%)	189 (30.4%)	90 (29.0%)	99 (31.9%)
HRT duration												
mean (st.dev.), yrs	6.1 (5.1)	6.7 (5.7)	5.4 (4.3)	1.9 (2.2)	1.8 (2.2)	2.0 (2.3)	NA	NA	NA	4.3 (4.8)	4.4 (4.9)	4.2 (4.8)
MI												
mean (st.dev.), units	34.3 (5.0)	34.6 (5.1)	33.9 (5.0)	33.1 (2.0)	33.1 (2.0)	33.2 (2.1)	36.2 (2.1)	36.1 (2.2)	36.2 (2.0)	36.9 (2.4)	37.0 (2.4)	36.8 (2.4)
ELEE												
mean (st.dev.), yrs	32.6 (6.2)	32.6 (6.2)	32.5 (6.1)	31.8 (5.1)	31.9 (5.5)	31.7 (4.8)	32.0 (6.0)	32.1 (6.0)	32.0 (5.9)	33.8 (6.5)	34.2 (6.1)	33.5 (6.7)

* alcohol consumption was reported as average units/week for the Generations Study, average self-reported intensity (numbers 1 to 9) at ages 20, 30, 40 and 50 for EPIC-Italy, and average gram per day for EPIC-IARC and MCCS
st.dev = standard deviation; yrs = years

4.3. METHODS

4.3.1 Pre-processing of the HM450K data

The following protocol was used for HM450K methylation data for the new EPIC-Italy data, EPIC-IARC and MCCS. Pre-processing was conducted in the R package ‘minfi’ [367]. Raw intensity data IDAT files were imported into R and background subtraction and normalisation were conducted using the ‘preprocessIllumina’ function, followed by correction for type I and II probe bias by subset-quantile within array normalisation using the ‘preprocessSWAN’ function [368]. Samples with missing information (P detection value > 0.01) for more than 5% of the CpG sites were excluded, as well as CpG sites with missing information in 20% or more of the samples. Furthermore, CpG probes located on the Y chromosome and cross-hybridising probes [328] were excluded using the R package ‘DMRcate’. WBC composition was estimated using the Houseman method [329].

4.3.2 Statistical analysis

The associations between the MI and breast cancer risk were individually analysed in the four studies using conditional logistic regression (function ‘clogit’ in R package ‘survival’) for breast cancer case-control status and the MI, adjusted by age, BMI, smoking duration, alcohol consumption (all variables were reported at the time of recruitment) and WBC composition. Alcohol consumption was reported as average units/week for the Generations Study, average self-reported intensity (numbers 1 to 9) at ages 20, 30, 40 and 50 for EPIC-Italy, and average gram per day for EPIC-IARC and MCCS. The association was analysed for the MI as a continuous variable and also by control-quartiles in each study cohort. In **Chapter 3** a stronger association between the MI and breast cancer risk for pairs including cases with a longer time to diagnosis was observed. Therefore, the pairs were stratified into two groups on median time to diagnosis to explore differences in the risk association between a shorter or a longer time between blood sampling and diagnosis. The correlation between the MI and ELEE was explored using Pearson’s correlation coefficient.

The meta-analyses were conducted for log odds ratios and standard errors using weighted random-effects model with the restricted-maximum likelihood method (function ‘rma.uni’ in the R package ‘metafor’). The Cochran’s Q statistics and I^2 statistics were used to estimate heterogeneity

between the studies; $Q < 0.05$ and $I^2 > 50\%$ were defined as heterogeneous estimations. If the heterogeneity I^2 was equal to 0, the meta-analysis behaved as a fixed-effects model.

4.4. RESULTS

4.4.1 Baseline characteristics

The analysis included 2374 women (1187 matched case-control pairs) from four prospective study cohorts (The Generations Study, EPIC-Italy, EPIC-IARC and MCCS) with mean time to diagnosis of 4.0, 8.5, 7.5, and 7.9 years (**Table 4.1**). Women in EPIC-Italy were youngest at recruitment with a mean age of 51.2 and women were oldest in MCCS with a mean age of 56.7. This was reflected in lowest and highest percentage of postmenopausal women (52.1% and 61.4%), ELEE values (mean values of 31.8 and 33.8) and MI values (mean values of 33.1 and 36.9) for the two study cohorts. In relation to the ELEE model, women in the Generations Study and EPIC-Italy had slightly younger ages at menarche (mean ages of 12.7 and 12.6) compared to EPIC-IARC and MCCS (13.1), and age at menopause varied from 47.8 for EPIC-Italy and 50.2 in the Generations Study. MCCS had the highest percentage of women who had reported that they have been breastfeeding (79.5%) as well as the most number of pregnancies (mean 2.4 pregnancies per woman). EPIC-IARC had the lowest percentage of women who had ever breastfed (68.2%) and number of pregnancies (1.6). Furthermore, there were substantially more smokers in EPIC-Italy and EPIC-IARC (24.6% and 21.7%) compared to MCCS and the Generations Study (6.8% and 7.1%), the highest percentage of OC and HRT users was in the Generations Study (80.2% and 31.7%, respectively) and lowest percentage in EPIC-Italy (37.3% and 19.5%, respectively).

4.4.2 The MI is associated with breast cancer risk

The association between the MI and breast cancer risk as continuous and categorical variable was explored across the four study cohorts and combined in meta-analyses. All estimates were adjusted for age, BMI, smoking duration, alcohol consumption, and WBC composition. Smoking duration was not adjusted for in EPIC-IARC due to missing data for a majority of the subjects ($n=506$).

The MI was significantly associated with breast cancer risk in the Generations Study (OR=1.04, 95% CI: 1.01 – 1.08, $P=0.022$) but not individually in the other study cohorts, EPIC-Italy (OR=1.13, 95% CI: 0.85 – 1.51, $P=0.394$), MCCS (OR=1.06, 95% CI: 0.96 – 1.17, $P=0.225$) and EPIC-IARC (OR=0.98, 95% CI: 0.90 – 1.06, $P=0.616$). The combined meta-analysis, however, showed low heterogeneity across study cohorts ($Q=0.45$, $I^2=0\%$) and a significant association with

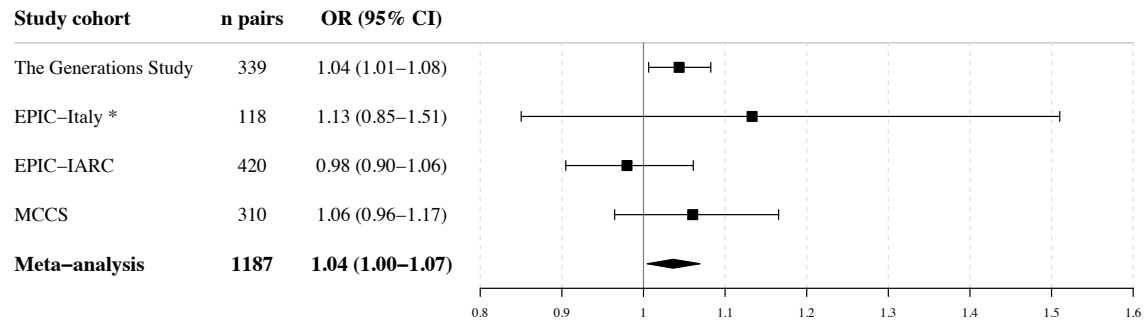


Figure 4.1. Forest plot of the MI and breast cancer risk

The association between the MI (as continuous variable) and risk for breast cancer was estimated in the four studies using conditional logistic regression adjusted for age, BMI, alcohol consumption, smoking duration, and WBC composition (all variables were reported at the time of recruitment). The log odds ratios represent the odds ratios (ORs) and the lines the 95% confidence intervals (CIs).

* EPIC-Italy corresponds to the new EPIC-Italy samples, not included in the development of the MI

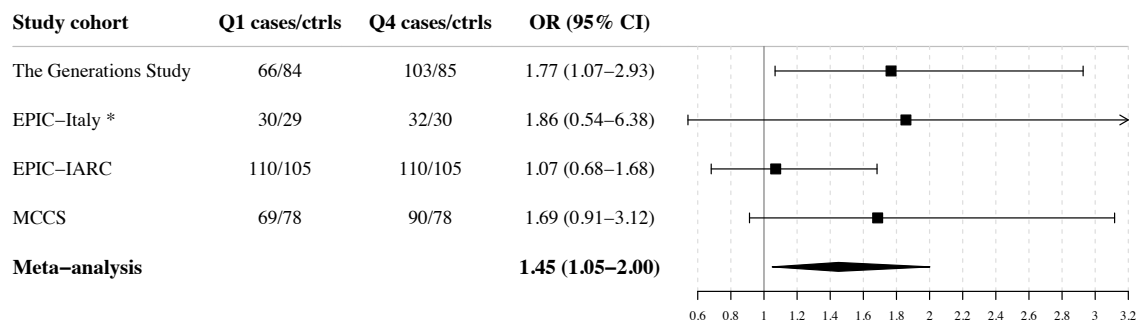


Figure 4.2. Forest plot of the MI by the quartile and breast cancer risk

The association between the MI (as categorical variable) and risk for breast cancer was estimated in the four studies using conditional logistic regression adjusted for age, BMI, alcohol consumption, smoking duration, and WBC composition (all variables were reported at the time of recruitment). The log odds ratios were combined in a meta-analysis using restricted-maximum likelihood model. The square boxes represent the Odds Ratios (ORs) between the highest and lowest quartile in each study and the lines the 95% Confidence Intervals (CIs).

* EPIC-Italy corresponds to the new EPIC-Italy samples, not included in the development of the MI

Table 4.2. Meta-analysis of the MI and breast cancer risk

The association between MI (as continuous variable) and risk for breast cancer was estimated in the four studies using conditional logistic regression adjusted for age, BMI, alcohol consumption, smoking duration and WBC composition (at recruitment). The log odds ratios were combined in a meta-analysis using restricted-maximum likelihood model.

Study cohort	n pairs	OR (95% CI)	P value	Test for heterogeneity
The Generations Study	339	1.04 (1.01 - 1.08)	0.022	
EPIC-Italy *	118	1.13 (0.85 - 1.51)	0.394	
EPIC-IARC	420	0.98 (0.90 - 1.06)	0.616	
MCCS	310	1.06 (0.96 - 1.17)	0.225	
Meta-analysis	1187	1.04 (1.00-1.07)	0.024	$Q=0.45$; $I^2=0\%$

* EPIC-Italy corresponds to the new EPIC-Italy samples, not included in the development of the MI

Table 4.3. Meta-analysis of the MI by the quartile and breast cancer risk

The MI was divided in four control-quartiles and the association between MI (as categorical variable) and risk for breast cancer was estimated using conditional logistic regression adjusted for age, BMI, alcohol consumption, smoking duration and WBC composition (at recruitment). The log odds ratios were combined in a meta-analysis using restricted-maximum likelihood model.

	min	max	n cases	n controls	OR (95% CI)	P value	Test for heterogeneity
The Generations Study	Q1	20.3	30.3	66	84	Ref	
	Q2	30.3	33.8	88	85	1.52 (0.95 - 2.42)	0.083
	Q3	33.8	37.0	82	85	1.60 (0.96 - 2.64)	0.069
	Q4	37.0	53.1	103	85	1.77 (1.07 - 2.93)	0.027
EPIC-Italy *	Q1	24.7	31.9	30	29	Ref	
	Q2	31.9	33.0	30	30	1.86 (0.57 - 6.07)	0.303
	Q3	33.0	34.1	26	29	0.83 (0.22 - 3.08)	0.775
	Q4	34.1	43.7	32	30	1.86 (0.54 - 6.38)	0.325
EPIC-IARC	Q1	29.4	34.9	110	105	Ref	
	Q2	34.9	36.2	107	105	0.95 (0.63 - 1.44)	0.816
	Q3	36.2	37.5	93	105	0.86 (0.56 - 1.31)	0.476
	Q4	37.5	44.0	110	105	1.07 (0.68 - 1.68)	0.767
MCCS	Q1	30.5	35.1	69	78	Ref	
	Q2	35.1	36.5	63	77	1.04 (0.63 - 1.69)	0.891
	Q3	36.5	38.2	88	77	1.58 (0.95 - 2.63)	0.080
	Q4	38.2	46.4	90	78	1.69 (0.91 - 3.12)	0.095
Meta-analysis	Q1		275	296	Ref		
	Q2		288	297	1.16 (0.88 - 1.52)	0.290	$Q=0.40$; $I^2=9\%$
	Q3		289	296	1.22 (0.83 - 1.80)	0.303	$Q=0.17$; $I^2=44\%$
	Q4		335	298	1.45 (1.05 - 2.00)	0.024	$Q=0.44$; $I^2=16\%$

* EPIC-Italy corresponds to the new EPIC-Italy samples, not included in the development of the MI

breast cancer risk with an OR of 1.04 per one unit increase in the MI (95% CI: 1.00 – 1.07, $P=0.024$).

The results are summarised in **Table 4.2** and **Figure 4.1**.

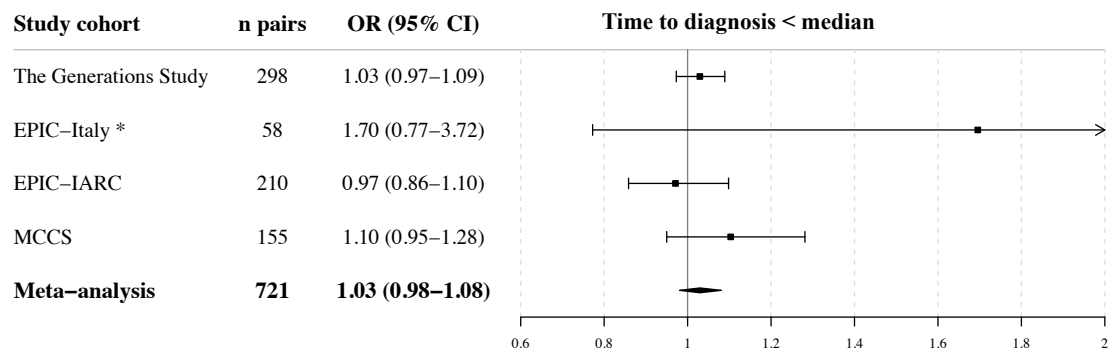
Women in the highest quartile compared to the lowest quartile of MI had significantly higher breast cancer risk in the Generations Study ($OR_{Q4_vs_Q1}=1.77$, 95% CI: 1.07 – 2.93, $P=0.027$). The combined meta-analysis showed low heterogeneity across study cohorts ($Q=0.44$, $I^2=16\%$) with a combined OR of 1.45 between the highest and lowest quartile ($OR_{Q4_vs_Q1}=1.45$, 95% CI: 1.05 – 2.00, $P=0.024$). The results are summarised in **Table 4.3** and **Figure 4.2**.

There was no significant association between the MI and breast cancer risk in pairs with time to diagnosis less than median in the combined meta-analysis ($OR=1.03$, 95% CI: 0.98 - 1.08, $P=0.241$), but there was a significant association in pairs with time to diagnosis above the median ($OR=1.05$, 95% CI: 1.01 - 1.10, $P=0.021$). Low heterogeneity was observed in the two meta-analyses ($Q=0.36$, $I^2=13\%$ and $Q=0.55$, $I^2=0\%$, respectively). The results are summarised in **Table 4.4** and **Figure 4.3**.

4.4.3 The MI is not associated with ELEE

The association between the ELEE and MI was analysed in women with no missing information of the variables included in the ELEE model (age at menarche, age at menopause, number of pregnancies and breastfeeding duration), leaving 678 women in the Generations Study, 233 women in EPIC-Italy, 510 women in EPIC-IARC and 436 women in MCCS for the analysis. Breastfeeding duration was missing for many of the women in EPIC-IARC ($n=297$) and therefore not included when calculating the ELEE. Overall, the MI did not correlate with the ELEE in any of the four study cohorts (**Table 4.5**).

A.



B.

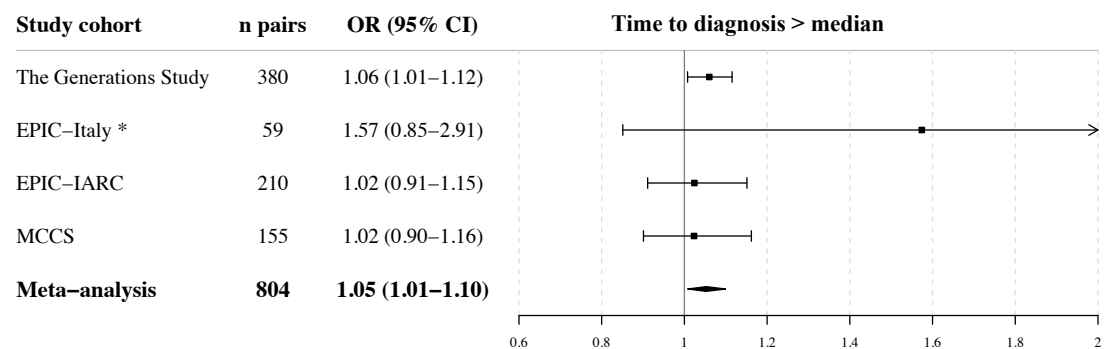


Figure 4.3. Forest plot of the MI, breast cancer risk and time to diagnosis

The pairs in all four study cohorts were divided in two groups by the median time to diagnosis. The association between the MI (as continuous variable) and risk for breast cancer was estimated using conditional logistic regression adjusted for age, BMI, alcohol consumption, smoking duration and WBC composition (all variables were reported at the time of recruitment). The log odds ratios for the two groups were combined in two meta-analyses using restricted-maximum likelihood model. The square boxes represent the odds ratios (ORs) and the lines the 95% confidence intervals (CIs). **A.** The results in groups with shorter time to diagnosis **B.** The results in groups with longer time to diagnosis

* EPIC-Italy corresponds to the new EPIC-Italy samples, not included in the development of the MI

Table 4.5. Correlation between the MI and ELEE across study cohorts

The correlation between the MI and ELEE was analysed using Pearson's correlation coefficient.

	n subjects	Correlation coefficient	P value
The Generations Study	678	-0.04	0.340
EPIC-Italy *	233	0.07	0.263
EPIC-IARC	510	0.06	0.154
MCCS	436	-0.04	0.449

* EPIC-Italy corresponds to the new EPIC-Italy samples, not included in the development of the MI

4.5. DISCUSSION

In this chapter the association between the MI and breast cancer risk has been explored in a meta-analysis across four different prospective study cohorts, excluding the original discovery dataset in EPIC-Italy. A significant combined association between the MI and breast cancer risk was observed, with an OR of 1.04 per one unit increase in the MI, and an OR of 1.45 between women in the highest quartiles compared to the lowest quartiles. The association seemed to be strongest in case-control pairs with a longer time to diagnosis. Furthermore, an association between the ELEE model and the MI was not observed in any of the four studies.

Even though the combined analyses showed a significant association between the MI and breast cancer risk, there are differences between the study cohorts. The association between the MI and breast cancer risk was strongest in the Generations Study ($OR_{Q4_vs_Q1}=1.77$, 95% CI: 1.07 - 2.93, $P=0.027$) and weakest in EPIC-IARC ($OR_{Q4_vs_Q1}=1.07$, 95% CI: 0.68 - 1.68, $P=0.767$). Step-wise increase in the odds ratios in the MI by quartile analysis was observed in the Generations Study and MCCS, as well as in the combined meta-analysis, but not for EPIC-IARC and EPIC-Italy (**Table 4.3**). The associations might have been influenced by the range of MI values in the study cohorts with the largest range in the Generations Study (20.3 to 53.1) and smallest range in EPIC-IARC (29.4 to 44.0).

Furthermore, EPIC-IARC has a shorter average time to diagnosis for cases, which might have affected the results given the association was stronger with a longer time to diagnosis. This association with a longer time to diagnosis might be due to changes in methylation signatures occurring at early stage cancer development (before diagnosis) that might be adversely affecting the methylation of CpG sites that contribute to the MI. This suggests that the increase in MI is not due to the reverse effect from early cancer development, but rather that early cancer development adversely affects the MI, however, this needs to be explored further.

The weights in the meta-analysis are by default calculated as the inverse of the standard errors. Estimates with smaller standard errors will therefore contribute more to the meta-analysis in comparison to estimates with larger standard errors, such as EPIC-Italy that had large confidence intervals across all analyses due to the smallest sample size ($n=236$). The combined analyses are therefore less influenced by the large odds ratios observed for EPIC-Italy in these analyses.

Furthermore, differences between the sequencing techniques used cannot be excluded (targeted bisulphite sequencing and the HM450K array). Also, differences in the study cohorts cannot be excluded that might affect this association, including study design. For example there are more smokers in the EPIC studies compared to the Generations Study and MCCS, which considerably affects DNA methylation. Also, there are different patterns in breastfeeding, number of pregnancies, and OC and HRT use across the studies, which will affect the total lifetime oestrogen exposure.

Unfortunately, not all 31 CpG sites included in the MI model are included in the new HM850K array, which meant that a fifth study cohort could not be included in the analysis (the Prostate, Lung, Colorectal, and Ovarian Cancer Screening trial, PLCO). These are the CpG probes cg08349826, cg15127563, cg25372296, cg22273487 nearest the genes *GRIN2A*, *ITM2C*, *MIR137HG* and *RALY*. Further optimisation of the MI including only the 27 CpG sites located on the new array is needed to include future HM850K array data sets.

In summary, a combined significant association between the MI and breast cancer risk was observed, but there were differences between the study cohorts included in the meta-analysis. Also, no association between the ELEE model and the MI was observed across any of the four studies. This suggests that a signature for lifetime oestrogen exposure has not been identified, but a signature that is associated with breast cancer risk. Further optimisation of the ELEE and the MI is needed to establish the relationship and mechanisms between oestrogen exposure, DNA methylation and breast cancer risk.

CHAPTER 5:
Molecular Mechanisms Between
Oestrogen Exposure and DNA
Methylation

5 CHAPTER 5: Molecular Mechanisms Between Oestrogen Exposure and DNA Methylation

5.1. BACKGROUND

In the previous chapters the associations between lifetime oestrogen exposure, DNA methylation in blood and breast cancer were explored, and a methylation signature associated with breast cancer risk was identified. It is well established that oestrogens are involved in the development of breast cancer, however, the mechanisms underlying the association between oestrogen exposure and breast cancer are not fully understood [137]. In this chapter the mechanisms between oestrogen and DNA methylation in blood will be further explored.

5.1.1 Molecular mechanisms for oestrogen-induced DNA methylation

The two main mechanisms described are a transcriptional response from oestradiol via the ERs and genotoxic effects from other different oestrogen metabolites (**Figure 1.2**). Oestrogens bind to and activate the ER- α and ER- β [243, 244], which translocate to the nucleus and interact with the DNA at ER-binding sites, which often function as enhancers [11]. This increases the gene expression of ER-target genes, which often are involved in cell proliferation, particularly for the ER- α that is associated with cancer development [11, 243-245]. Another outcome from oestrogen exposure is a genotoxic effect. The catechols 2-OHE2, 4-OHE2 that are formed in the oestrogen metabolism can oxidise to 2,3-quinones and 3,4-quinones, which can bind to the DNA and form DNA adducts. The 3,4-quinones are the most carcinogenic compounds, and can cause unstable, depurinating adducts leading to DSBs [249-253]. Additionally, the formation of ROS in the oestrogen metabolism leads to oxidative damage of the DNA [253, 255]. The error-prone DNA repair mechanisms following DNA damage can lead to mutagenesis as well as changes in histone modifications and DNA methylation by the recruitment of the maintenance DNMT1 and the *de novo* DNMT3B [256-260]. The increase in DNA methylation is thought to act as a protection against replication of the damaged sites [257-260].

Table 5.1. The expression of oestrogen receptors in blood

Summary of the referenced papers that have explored the expression of ER- α and ER- β and GPER1 in different blood cell types, the methods they have used, and the results.

Study	Cell type	Method(s)	Result
Pelekanou <i>et al.</i> 2016 [374]	Monocytes from healthy women and men	Western blot and qPCR	ER- α -36 isoform and GPER1
Pierdominici <i>et al.</i> 2010 [372]	Peripheral blood lymphocytes (CD4+ and CD8+ T lymphocytes, B lymphocytes and NK cells) from healthy men	Flow cytometry and western blot	Intracellular expression of ER- α and ER- β (ER- α -46 isoform most represented ER)
Zhang <i>et al.</i> 2009 [367]	T-lymphocytes from women	Western blot	ER- β > ER- α
Scariano <i>et al.</i> 2008 [373]	PBMNCs from healthy women and men	Flow cytometry, western blot, qPCR	ER- α and ER- β expression in the nuclei
Rider <i>et al.</i> 2006 [366]	T-cells from systemic lupus erythematosus patients	Western blot	ER- β > ER- α
Shim <i>et al.</i> 2006 [368]	CD19+ T-cells and CD3+ T-cells from healthy men and women	Western blot	ER- β only
Nancy <i>et al.</i> 2005 [371]	PBMNCs from myasthenia gravis patients	Flow cytometry and qPCR	ER-alpha and ER-beta expression
Phiel <i>et al.</i> 2005 [369]	CD4+ T-cells, CD19+ B-cells, CD8+ T-cells, monocytes from healthy women and men	qPCR	ER- α > ER- β (CD4+ T-cells) ER- β > ER- α (CD19+ T-cells) low but detectable levels of ER- α and ER- β (CD8+T-cells and monocytes)
Kassi <i>et al.</i> 2001 [370]	PBMNCs from systemic lupus erythematosus patients	qPCR	ER-alpha and ER-beta expression

5.1.2 ER- α and ER- β expression in blood

The oestrogen responsiveness is tissue-specific due to the expression of different ER subtypes, the expression of different co-activators and repressors, and the expression of FOXA1 and GATA3 [11, 244]. In early breast cancer development, the expression of ER- α is increased and ER- β is decreased [246, 248]. The expressions of ER- α and ER- β have previously been observed in blood with some inconsistent results [369-378]. The expression levels seem to differ between cell types, for example relatively high expression levels were observed for CD4+ T-cells and B-cells, but low, yet still detectable, levels in CD8+ T-cells and monocytes [373]. Higher levels of ER- α compared to ER- β were observed in CD4+ T-cells and higher levels of ER- β compared to ER- α in CD8+ T-cells and monocytes. Another study found the GPER1 and ER- α -36 isoform to be expressed in monocytes [378]. The expressions of both the ER- α and the ER- β in PBMNCs have been confirmed in several studies [374, 376, 377], but one study reported only ER- β expression in CD19+ B-cells and CD3+ T-cells [372], and two studies report higher levels of ER- β compared to ER- α in T- and B-cells [370, 373]. Note that for some these studies, the blood samples were collected from healthy men only [376], or patients with autoimmune diseases [370, 374, 375], which can affect the results. In summary, the evidence for the expression of the ERs in blood is not entirely conclusive. The previous studies are summarised in **Table 5.1**.

5.1.3 Aims of this chapter

In this chapter two hypotheses of oestrogen-induced DNA methylation were explored; a DNA repair mechanism following DNA damage from oestrogens, and a transcriptional mechanism via the ERs (**Figure 1.2**). As a model system a lymphocyte cell line (LCL) was used with three different mutation phenotypes: wild-type samples (with no pathogenic *BRCA1/2* mutations), samples with *BRCA1* mutations, and samples with *BRCA2* mutations. Following oestrogen treatment it was hypothesised that different DNA methylation profiles would be observed between the three phenotypes, as a result from different DNA repair-responses to the DNA damage. The genome-wide DNA methylation levels were measured using the LINE-1 assay, and four CpG sites that showed association with breast cancer risk in **Chapter 3** were analysed (**Table 3.9**). Furthermore, a transcriptional mechanism was explored by analysing the protein levels of ER- α and ER- β in LCLs.

5.2. MATERIALS & METHODS

5.2.1 Cell culture

5.2.1.1 LCLs

All experiments were conducted in Epstein-Barr Virus (EBV)-transformed isolated B-cells (LCLs). The EBV-transformation increases the proliferation of the cells, leads to immortalisation without becoming tumorigenic, and affects minimal genetic changes, making them an excellent renewable source of genetic material [379-381]. The LCLs have been used as a model system in a number of studies, are relatively easy to grow and requires minimal maintenance [381]. The cells are non-adherent and grow as clusters in clumps due to the expression of adhesion molecules (**Figure 5.1**).

Nine samples of EBV-transformed LCLs were provided by the Kathleen Cuningham Foundation Consortium for research into Familial Breast Cancer (kConFab). Of these, three samples were from women who have tested negative for pathogenic mutations in *BRCA1* or *BRCA2* (wild-type), three samples from women carrying a pathogenic mutation in *BRCA1* and three samples from women carrying a pathogenic mutation in *BRCA2* (**Table 5.2**). All women were premenopausal, had not been diagnosed with cancer, and not used HRT or Tamoxifen (to reduce exogenous hormonal exposure). The shipment from Australia took approximately three days and the cells were sent in tightly packed T25 flasks. All cells were clear, yellow and with clumps of cells at the bottom of the flasks. The cells were incubated overnight at 5% CO₂ at 37°C and split to new flasks on the second day upon arrival. The LCLs were grown in RPMI-1640 (Sigma Life Science) supplemented with 10% Fetal Calf Serum (FCS, First Link UK), 1% 2mM L-glutamine and 1% Penicillin-Streptomycin. When the cell suspension turned yellow 2-5 ml of media was added to the samples and they were split to new flasks every 3-5 day. The cells were regularly tested for Mycoplasma using the Lonza MycoAlert Mycoplasma Detection kit.

5.2.1.2 ZR-75-1 and U2OS-ER- β cell lines

The cell lines ZR-75-1 (breast epithelial cells) and U2OS-ER- β (bone osteosarcoma epithelial cells) were kindly provided from Professor Simak Ali's laboratory, Hammersmith Campus. Both cell lines were grown in T75 flasks in Dulbecco's Modified Eagle Medium (DMEM, Sigma Life Science)

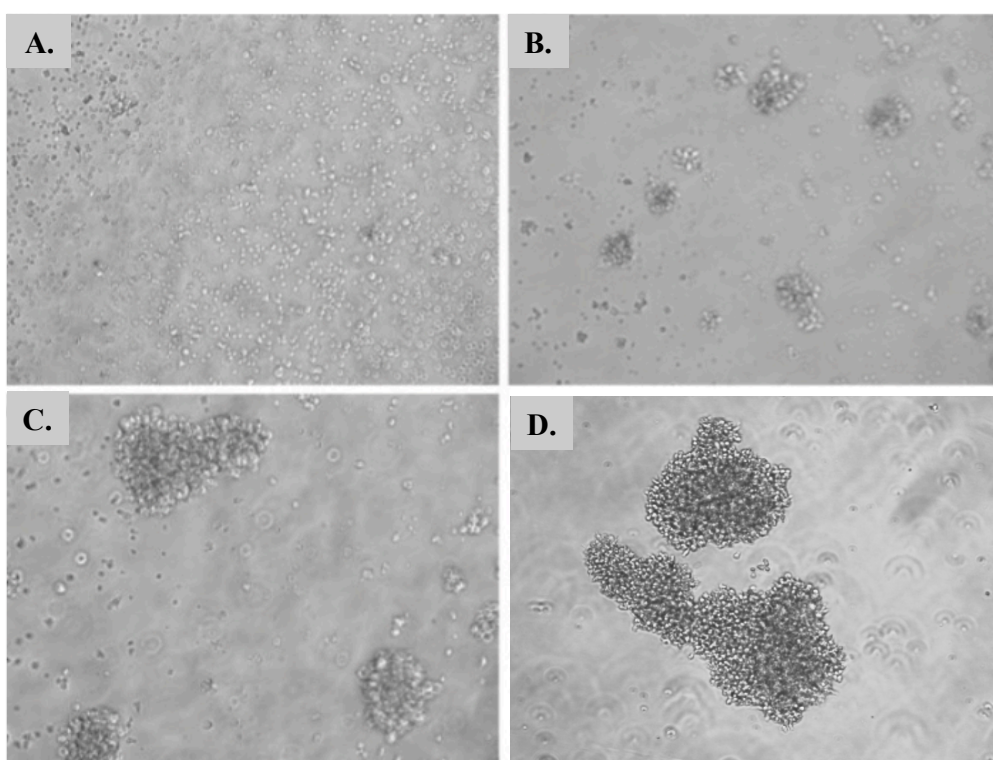


Figure 5.1. EBV-transformed Lymphocyte Cell Lines (LCLs)

Picture A-C from Hui-Yuen *et al.* [376] shows the transformation of EBV-infected LCLs from uninfected cells (A), one week after EBV-transformation (B) and after five weeks (C). Picture D is a picture of cells from one LCL sample used in this study.

Table 5.2. Summary of LCL samples from kConFab

All experiments in this chapter was conducted in EBV-transformed lymphocyte cell lines (LCLs) from kConFab. This included three samples from women who have tested negative for pathogenic mutations in *BRCA1* or *BRCA2* (wild-type), three samples from women carrying a pathogenic mutation in *BRCA1* and three samples from women carrying a pathogenic mutation in *BRCA2*. All women were premenopausal, had not been diagnosed with cancer, and not used HRT or Tamoxifen (to reduce exogenous hormonal exposures).

Sample	Age at blood sampling	Age at start OC use	Age at last OC use	Mutation	Mutation Information
BRCA1 (1)	44	20	20	BRCA1 del exon 20 ins G	Copy number deletion variant allele predicted to encode a non-functional protein.
BRCA1 (2)	38	15	38	BRCA1 5382_5383 ins C (STOP 1829)	Variant allele predicted to encode a truncated non-functional protein.
BRCA1 (3)	38	20	29	BRCA1 IVS 19-12 G>A	IARC class based on posterior probability from multifactorial likelihood analysis
BRCA2 (1)	38	17	20	BRCA2 IVS 16-2 ins AG	IARC class based on posterior probability from multifactorial likelihood analysis
BRCA2 (2)	28	17	27	BRCA2 8216 A>T (E2663V)	IARC class based on posterior probability from multifactorial likelihood analysis
BRCA2 (3)	22	17	22	BRCA2 6633_6637 del CTAA (STOP 2137)	Variant allele predicted to encode a truncated non-functional protein.
Wild-type (1)	49	21	22		
Wild-type (2)	33	18	29		
Wild-type (3)	44	18	41		

supplemented with 10% FCS, 1% 1mM L-glutamine and 1% Penicillin-Streptomycin. Cells were split every 3-5 days and incubated in 5% CO₂ at 37°C.

The cell lines were used as positive and negative controls for the western blots of ER- α and ER- β . ZR-75-1 expresses both ER- α and ER- β , however stronger for ER- α [382-384]. The expression of ER- β in the U2OS cells were regulated via the T-Rex system (Invitrogen) that regulates the expression of the gene of interest under the control of tetracycline-operator sequences. In the absence of doxycycline the gene expression is silenced from the tetracycline repressor, and the addition of doxycycline to the culture media released the tetracycline-repressor and allows expression [385].

5.2.1.3 Doxycycline treatment for the U2OS-ER- β cells

The U2OS-ER- β cells were plated on 6-well plates and incubated for 24 hours before the addition of doxycycline to the growth media. Doxycycline (Doxycycline Hyclate $\geq$ 98%, Sigma Aldrich) was dissolved in 500 ml water to a final concentration of 2 mg/ml and filtered via 0.22 μ M filter system. Cells for ER- β positive control were treated with 100 ng/mg of doxycycline for 24 hours before protein extraction.

5.2.1.4 Oestrogen treatment of the LCLs

Before oestrogen treatment of the LCLs, the growth media was changed to RPMI-1640 without phenol-red (Sigma Life Science) supplemented with charcoal-stripped FCS (Thermo Fisher) to reduce the hormonal background effects [386, 387]. The cells were washed with PBS twice before they were re-suspended in the oestrogen-depleted media and incubated for 24 hours before the addition of 17- β -oestradiol (E2, Sigma Life Science). Cells were treated with E2 in the concentrations of 0, 10 or 100nM for 6 hours before γ H2AX immunofluorescence staining, and 48 hours before DNA extraction and analysis of DNA methylation.

5.2.2 γ H2AX immunofluorescence imaging

For analysis of DNA damage in the oestrogen treated cells, the phosphorylation of the histone H2AX at serine 139 (γ H2AX) was used as a marker of double strand breakages [388]. After 6 hours of E2 treatment, the cells were washed with PBS twice and re-suspended in ~200ml of PBS. Cells were dried on cover slips and fixed by incubation in -20°C methanol for 5 minutes followed by -20°C acetone for 2 minutes, adapted from an earlier protocol for immunofluorescence staining [389]. The cells on the

cover slips were blocked in a buffer of PBS with 5% FCS and 0.03% Triton-X for 15 minutes and washed with PBS. The cells were incubated in 1:200 γ H2AX primary antibody (Cell Signaling) for 30 minutes, washed with PBS three times for five minutes, and incubated in 1:400 secondary antibody (Alexa Fluor 488 goat anti-rabbit, Invitrogen) for 30 minutes in the dark. The cover slips were washed with PBS three times for five minutes and mounted on microscope slides using ProLong Gold antifade reagent with DAPI (Invitrogen). The cells were imaged using the Leica SP5 Confocal microscope with 40X oil objective at wavelengths 405 nm for DAPI and 488 nm for γ H2AX at zoom factor 4.

5.2.2.1 Analysis of DNA damage

For analysis of the images of the γ H2AX staining, CellProfiler 3.1.5 and their pipeline for Speckle Counting was used, with a few modifications (**Figure 5.2**). In the first step the nuclei objects were identified from the DAPI with sizes 50-200 pixels of diameter. An adaptive threshold method was used to distinguish the nuclei, which adapts the threshold between the foreground and background intensities across the images. For each pixel, the threshold was calculated based on the pixels within a given window surrounding that pixel. Objects touching the border were discarded. The nuclei were distinguished via their shape, separating two cells if they were clumped together. For analysis of the images for γ H2AX, the intensities were first enhanced and overlapped with the DAPI objects. The γ H2AX foci objects with sizes 2-35 pixels were identified by the use of an adaptive method with the intensities divided in three classes instead of two, where the middle intensity was classified as background. Also, the lower threshold was set to an intensity of 0.1 to filter out low-intensity staining. This was due to the observation of background noise, or low-intensity non-specific cell staining of γ H2AX in the images. The foci were identified by intensity where one single peak of brightness was assigned to one focus. The number of γ H2AX foci per cell was used to determine the level of DNA damage, and also the mean intensity level per foci per cell. The DNA damage was compared between treatments using Student's t-test.

5.2.3 Bisulphite pyrosequencing

Five DNA methylation assays were analysed by bisulphite pyrosequencing [183], including LINE-1 as a surrogate measurement for genome-wide DNA methylation levels, and the four CpG sites that were found to be associated with breast cancer risk in both the Generations Study targeted sequencing data and EPIC-Italy 450K data (CpG sites near the genes *CTNNA2*, *GRB10*, *RPH3AL* and *TINCR*, **Table**

3.9). Primers were designed using the PyroMark Assay Design 2.0 software including two PCR primers (one biotinylated at the 5' end to be able to bind to the streptavidin-coated beads in the pyrosequencing process) and one sequencing primer, and ordered from Sigma-Aldrich (**Appendix 6**). Validation of primers included annealing temperature optimisation by running a temperature gradient PCR for 53 to 60°C and gel electrophoresis, and the temperature with the strongest PCR product was selected for each assay. All nine LCL samples were treated with 0, 10 and 100nM E2 for 48 hours and DNA was extracted using the Puregene Core Kit B (Qiagen) and quantified using the Nanodrop 1000. A total of 300 ng DNA from each sample was bisulphite converted in triplicates using the Zymo Research EZ-96 DNA Methylation kit (Cambridge Biosciences) and analysed on the PyroMark sequencer (PyroMark Q96 MD, Qiagen). Controls for the pyrosequencing included sequencing primer only, biotinylated primer only, PCR product without primers, 0% and 100% methylated DNA, and non-template control. Comparison of the DNA methylation levels between untreated and treated cells was analysed per dose (using all LCLs), per mutation phenotype, and per individual sample, using Student's t-test.

5.2.4 Western blot for ER- α and ER- β

Cell suspensions containing between 5×10^6 to 1×10^7 cells were harvested and lysed to break up the cell membrane using a RIPA buffer including the protease and phosphatases-inhibitors Aprotinin and Leupeptin. The cells were left to lyse for 30 minutes on ice and afterwards centrifuged at max speed for 15 min at 4°C. The supernatant containing the proteins were transferred to a new 1.5 ml tube and the proteins were quantified using the Bradford protein assay dye reagent (Bio-Rad) measured on wavelength 595nm.

For immunoblotting 10% and 5% acrylamide resolving and stacking gels were used for sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE). For each sample 15-30 μ g of protein was mixed with 3X Laemmli loading buffer (BioVision), boiled to unfold the proteins, and loaded onto the gels. Gels were run at 80V for 15min followed by 110V for 1.5-2h. The proteins were transferred to a PVDF membrane using the iBlot System (Invitrogen) and blocked in 5% milk for one hour followed by primary antibody incubation overnight at 4°C. The membranes were washed with TBST three times for 30 minutes and incubated in secondary antibody for one hour at room temperature. The membranes were washed with TBST six times for five minutes and covered in Pierce

ECL Western Blotting Substrate mix before detection using the Chemiluminescence Imager (GE Healthcare ImageQuant LAS 4000). Membranes were stripped with Restore PLUS Western Blot Stripping buffer (Thermo Scientific) and re-blotted with beta-actin for loading control. The antibodies used included 6F11 for ER- α (monoclonal, Invitrogen) diluted 1:1000 in 5% milk, MC10 for ER- β (monoclonal, Invitrogen) diluted 1:500 in 5% milk [390], and anti-mouse diluted 1:2000 and 1:500 in 5% milk as secondary antibody.

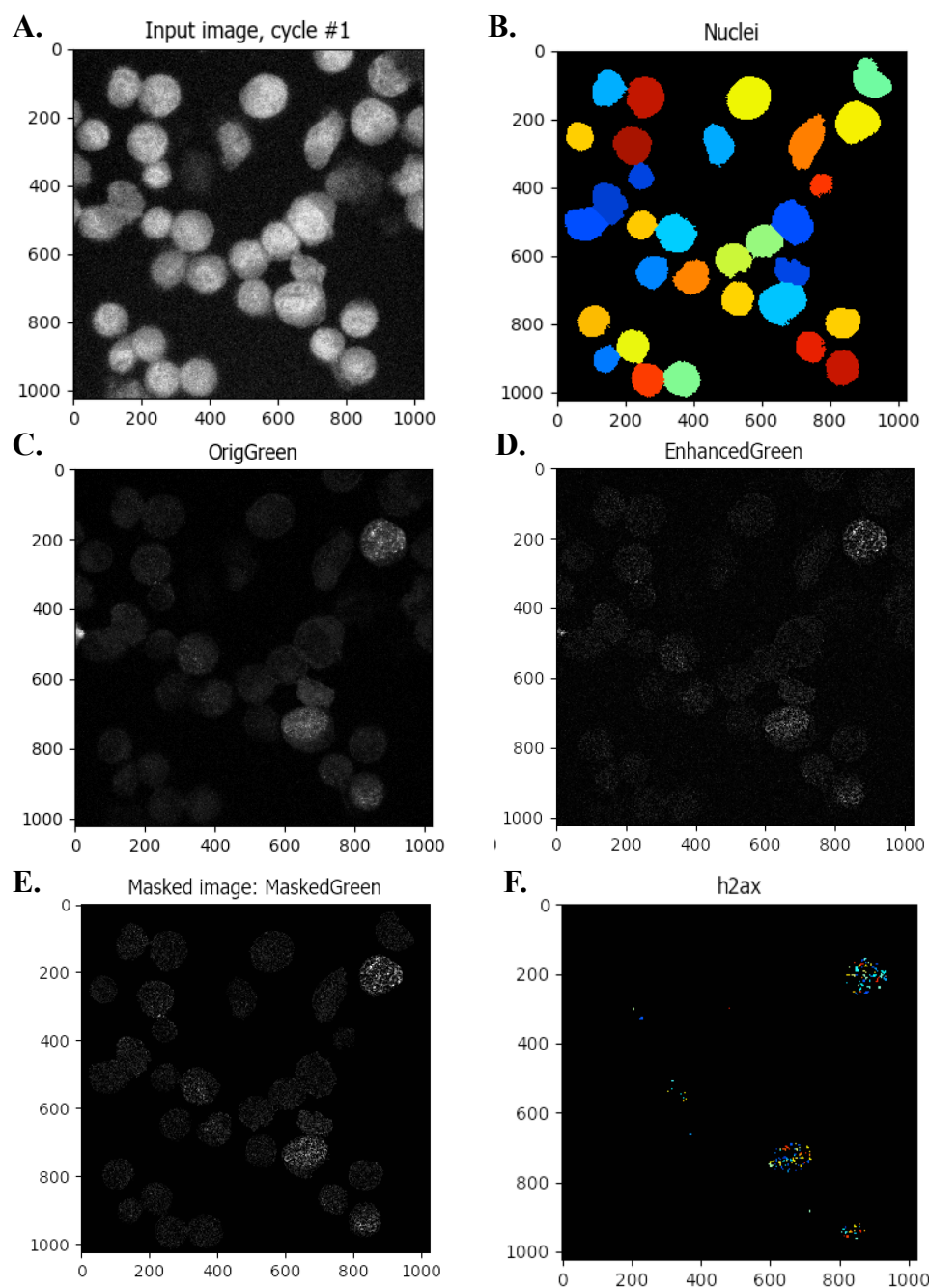


Figure 5.2. Analysis of γ H2AX immunofluorescence using CellProfiler

A. The image for DAPI is used as input for detection of nuclei objects. **B.** Nuclei objects are identified by their shape. **C.** The image for γ H2AX is used as input for foci (speckle) detection. **D.** The intensity are enhanced. **E.** The enhanced intensities are overlapped with the image of the nuclei (B). **F.** The γ H2AX foci are identified by intensity thresholds. Number of foci (F) per cell (B) was analysed in this study, as well as the mean intensity per foci per cell.

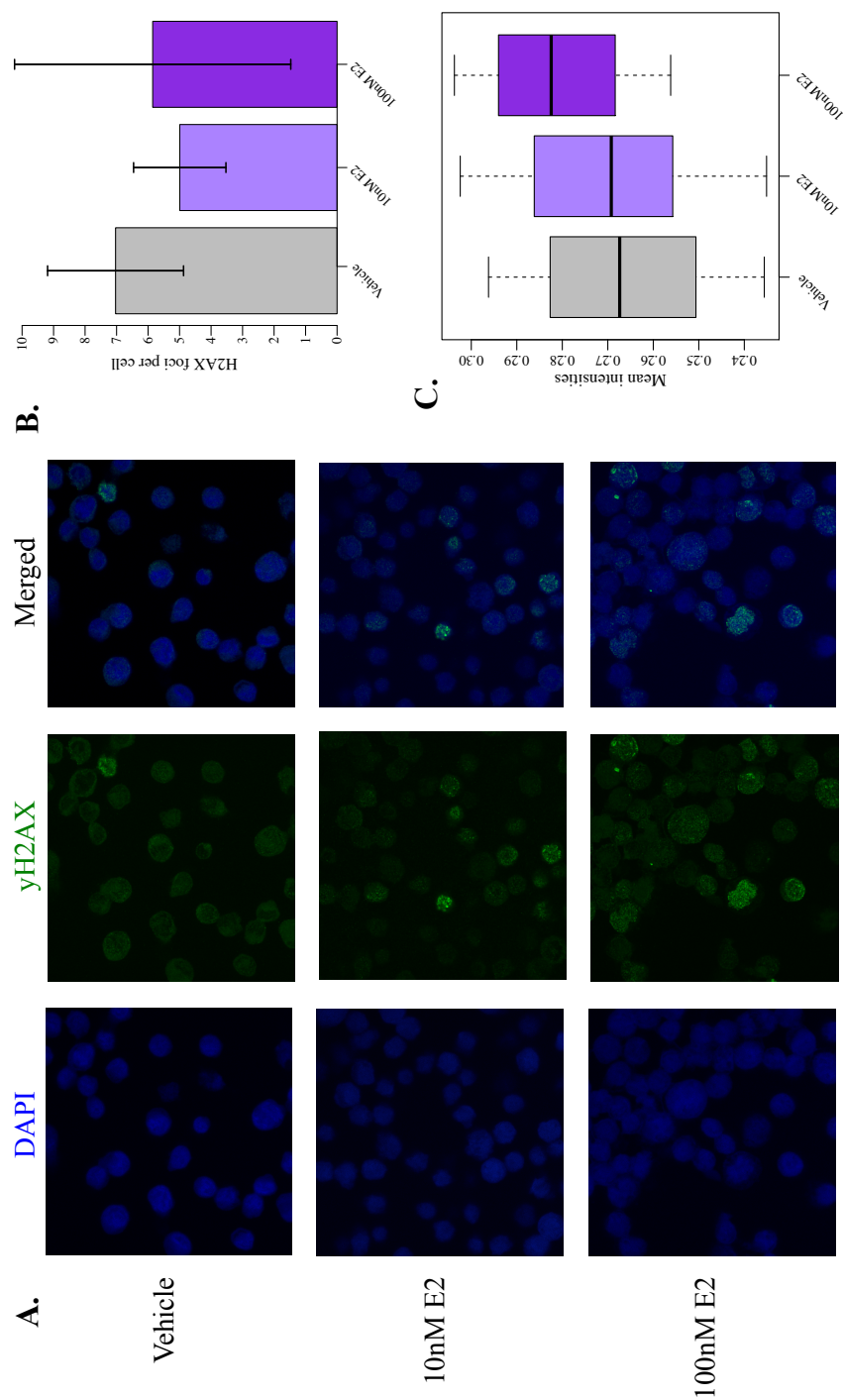


Figure 5.3. DNA damage following oestrogen treatment in LCLs
A. Representative figure of cells treated with 0 (vehicle), 10nM and 100nM 17- β -oestradiol (E2) stained with γ H2AX immunofluorescence as a marker for DNA damage. **B.** The number of γ H2AX foci per cell was quantified for five images taken randomly per treatment using CellProfiler for the different E2 treated cells. **C.** The mean intensities per foci per cell was further analysed. Differences between treatments were compared using Student's t-test with a *P*-value threshold of 0.05.

5.3. RESULTS

5.3.1 DNA damage following oestrogen treatment in LCLs

One LCL (wild-type) was treated with 0, 10 and 100nM E2 for six hours and analysed for DNA damage by γ H2AX immunofluorescence imaging (**Figure 5.3A**). Five images were taken randomly for each treatment. No significant difference in γ H2AX foci per cell was observed between the three different treatments (**Figure 5.3B**). However, low-intensity non-specific background binding of γ H2AX was observed for the images, making them difficult to analyse without stringent threshold settings (described in methods), and omitting low intensity foci. Mean intensity levels of the identified γ H2AX objects were also analysed for each treatment and image, and a stepwise increase in the intensity was observed, however, not significant (P -value > 0.05 , **Figure 5.3C**). This might suggest that DNA damage was higher for the LCLs with increased oestrogen treatment.

5.3.2 DNA methylation following oestrogen treatment in LCLs

All nine samples of LCLs were treated with 0, 10 and 100nM E2 for 48 hours and DNA methylation levels of five assays were analysed using bisulphite pyrosequencing. The designed primers are listed in **Appendix 6**. No significant (P -value > 0.05) difference was observed between untreated and treated samples in any of the five assays (**Figure 5.4**), neither when stratifying the samples by their *BRCA1/2* mutation phenotypes (**Figure 5.5A-C**). However, significant differences for *CTNNA2*, *GRB10*, *RPH3AL* and *TINCR* mean methylation values were observed when grouping all treated and untreated samples between the mutation phenotypes (**Figure 5.5D**). For the *CTNNA2* assay DNA methylation levels were lowest for the wild-type samples; for the *GRB10* assay the levels were highest for the wild-type samples compared to the *BRCA2*-mutation samples; for the *RPH3AL* assay the levels were highest for *BRCA1*-mutation samples, and for the *TINCR* assay the levels were highest for *BRCA2*-mutation samples.

Sample-specific DNA methylation changes following E2 treatment were observed for some of the individuals (**Figure 5.6**). Significant increase in DNA methylation following oestrogen treatment was seen for LINE-1 for one of the *BRCA2*-mutated individuals and treatments (10nM), and two of the individuals (wild-type and *BRCA1*) in one of the treatments (10nM and 100nM) in the *CTNNA2* assay. Significant decrease in DNA methylation following oestrogen treatment was observed for both

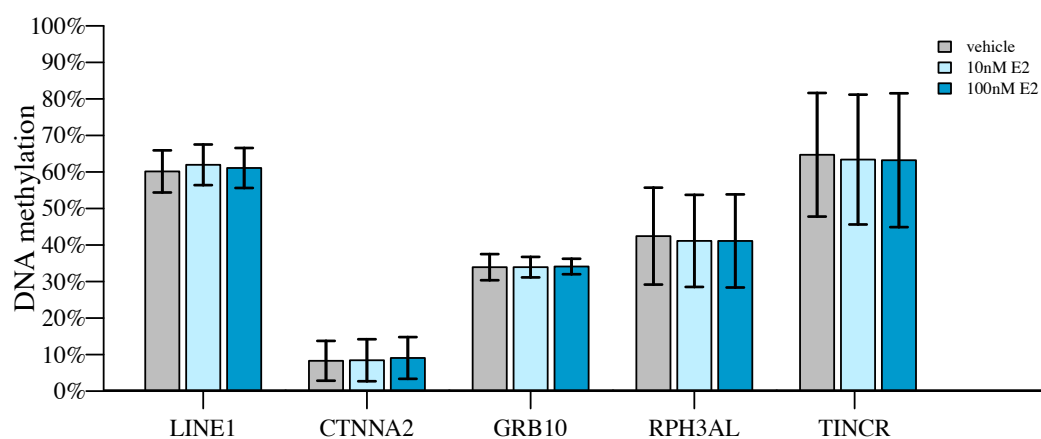


Figure 5.4. DNA methylation in LCLs following oestrogen treatment

Nine samples of LCLs were treated with 0 (vehicle), 10nM and 100nM 17- β -oestradiol (E2) for 48 hours, DNA was extracted and bisulphite converted, and pyrosequencing for five assays was conducted: LINE-1 as a surrogate measurement for genome-wide DNA methylation levels and four CpG sites that showed significant association with breast cancer risk in the Generations Study (**Chapter 3, Table 10**). No significant (P -value > 0.05, Student's t -test) difference in DNA methylation levels between treated and untreated samples were observed for any of the five assays.

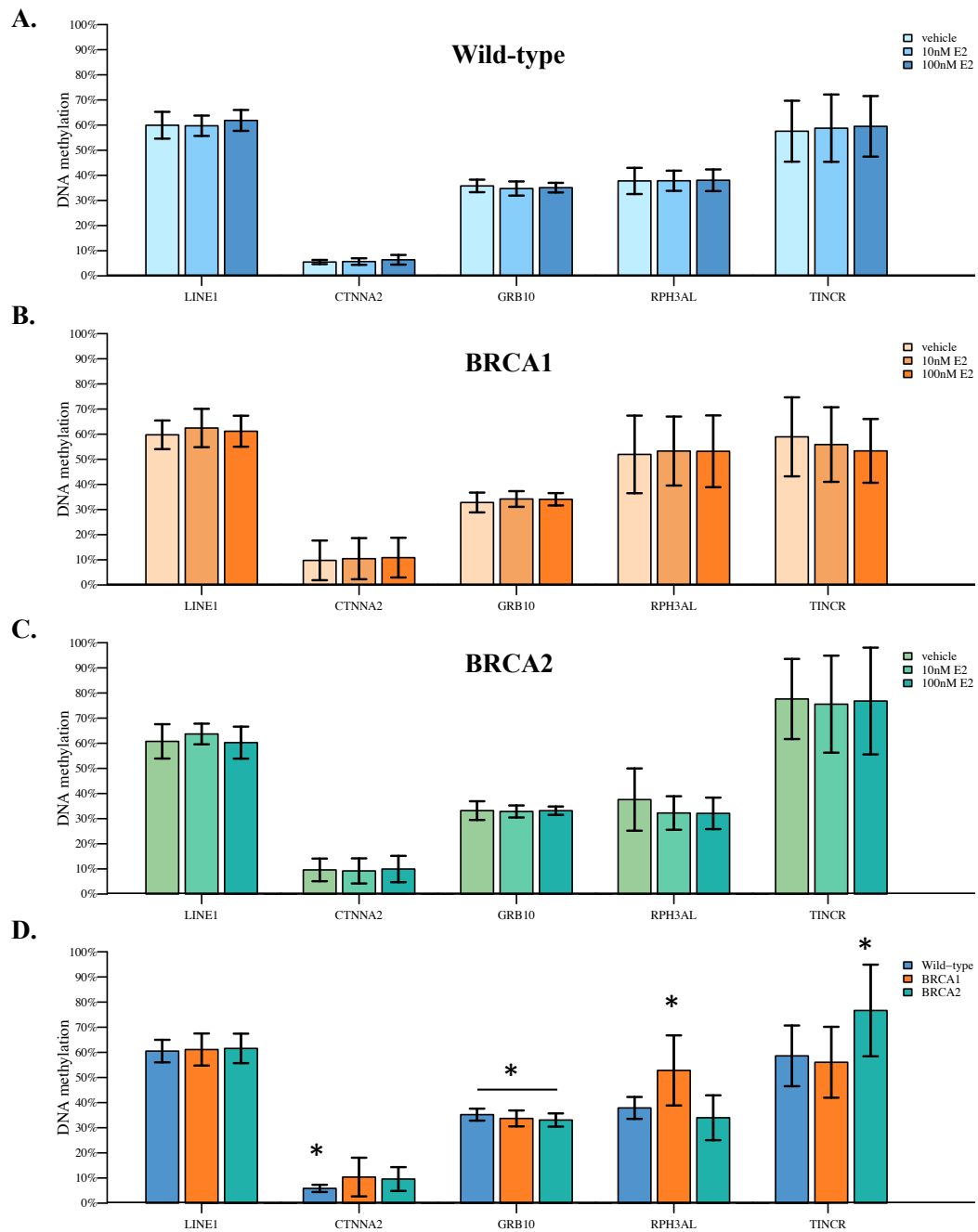


Figure 5.5. DNA methylation in LCLs following oestrogen treatment, phenotype specific

The DNA methylation levels from pyrosequencing for treated and untreated cells in the five assays for (A) wild-type samples from women with no pathogenic mutation in *BRCA1* or *BRCA2*, (B) samples from women with pathogenic *BRCA1* mutation and (C) samples from women with pathogenic *BRCA2* mutation. No significant differences between the untreated or treated samples were observed. D. The DNA methylation levels for the five assays were compared between the three mutation-phenotypes (grouping untreated and treated samples together), using Student's t-test and significant differences (P -value < 0.05 , marked by *) was observed for the CTNNA2, GRB10, RPH3AL and TINCR-assays.

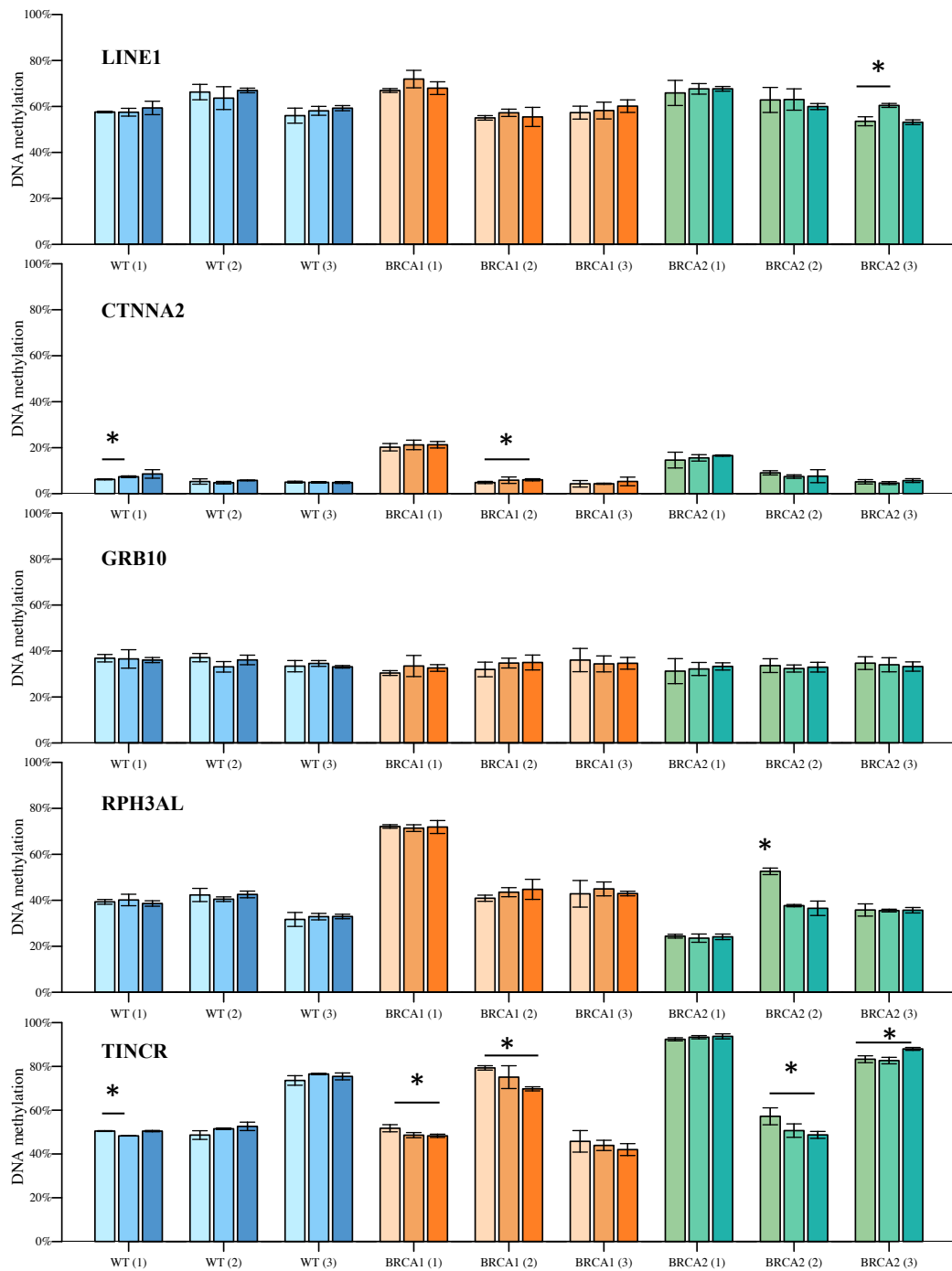


Figure 5.6. DNA methylation in LCLs following oestrogen treatment, per individual sample
Sample-specific DNA methylation levels for treated and untreated cells for the five pyrosequencing assays. DNA methylation levels between treatments were compared using Student's t-test with a *P*-value threshold of 0.05. Significant differences between treated and untreated cells are marked by *.

treatments in the *RPH3AL* assay for one of the *BRCA2*-mutated individuals, and for the *TINCR* assay for two of the *BRCA1*- and one of the *BRCA2*-mutated individuals for untreated and 100nM E2-treated cells, but an increase in another *BRCA2*-mutated individual. No significant difference was observed for the LINE-1 assay for any of the samples, except for an increase in 10nM E2-treated samples for one of the *BRCA2*-mutated individuals.

5.3.3 The ERs are not detectable in LCLs

The specificity for the ER- α antibody was high with a strong band at 66kDa only for the ER- α positive control (proteins extracted from ZR-75-1 cells). The ER- α was not detected in any of the nine LCLs, even when a total of 30 μ g of protein was loaded (**Figure 5.7A**). The specificity for the ER- β antibody was however questionable with two strong bands for the ER- β positive control around 60 and 55kDa (U2OS-ER- β cells treated with doxycycline), one faint band around 60kDa for the ER- β negative control, and two strong bands around 60 and 62kDa for the ER- α positive control (the ER- β was expected to also be expressed in ZR-75-1, but not as strong as ER- α). A few faint bands were observed on the blot for the nine LCLs, but not enough to conclude significant expression of ER- β (**Figure 5.7B**). A total of three blots per antibody were analysed, showing the same results.

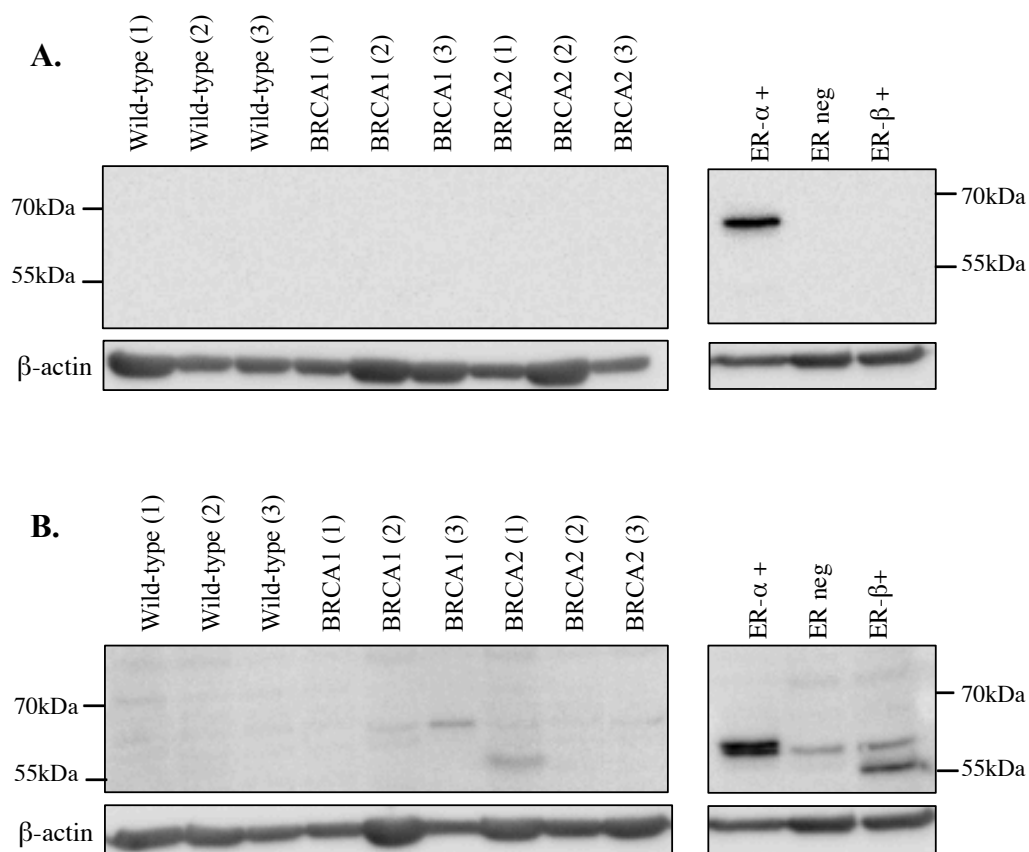


Figure 5.7. Western blot of ER- α and ER- β in LCLs

A. Western blot of ER- α in LCLs. Expected size for the antibody 6F11 was 66kDa. **B.** Western blot of ER- β in LCLs. Expected size for the antibody MC10 was 59kDa. In both blots 30 μ g of protein was loaded. Beta-actin was used as a loading control. Positive and negative controls were ZR-75-1 and U2OS-ER- β +/- doxycycline. A total of three blots per antibody were analysed, showing the same results.

5.4. DISCUSSION

In this chapter the mechanisms of oestrogen-induced DNA methylation have been explored in lymphocyte cell lines (LCLs). No expression of either ER- α or ER- β was observed in any of the samples, suggesting that the proteins are not highly expressed in LCLs, which would exclude the hypothesis of a transcriptional mechanism leading to DNA methylation changes. No significant difference in γ H2AX foci per cell was observed between untreated and E2 treated cells, however the intensity of the marker seemed to be increasing with higher E2 doses, which might suggest an increase in DNA damage following higher oestrogen exposure similar to previous reports [249]. Furthermore, no significant difference in DNA methylation levels in untreated or E2 treated samples was observed for any of the five analysed pyrosequencing assays, however, significant differences were observed for some of the E2 treated samples in some individuals. Also, *BRCA1/2*-mutation specific DNA methylation levels were observed for the four assays previously shown to be associated with breast cancer risk.

The expression of ER- α and ER- β was explored in LCLs since previous studies have showed somewhat inconsistent results of ER expression in blood (**Table 5.1**). Even though it was expected that the results might differ depending on the cell type, experimental design, and antibody specificity [376], the expression of the ERs in blood is not fully clear. The ERs might have been detectable in other WBCs, and the result is expected to differ in different cell types. Unfortunately, the ER- β antibody in this study did not bind specifically to its target, even though the antibody was selected from a recent review of ER- β antibodies [390]. Even so, no high expression of the ERs was observed in any of the LCLs, which suggests that the transcriptional mechanism hypothesis is not correct. There is still a possibility of an ER-independent effect on the transcription, for example via the GPER1 that was not been explored in this study. To further explore a transcriptional mechanism, it would be interesting to perform ChIP-seq of ERs in blood to identify any ER-binding events and ER-target genes, which are expected to be cell type specific. However, to do this the ERs need to be expressed at a higher level, so it would not be suitable to use LCLs for this, and fresh blood samples with cell sorting may be required.

Low-intensity background binding was observed in the images for γ H2AX staining, and DNA damage was observed even for untreated LCLs (ratio of γ H2AX foci and nuclei was on average 7 for

untreated cells), which might have complicated this analysis. Even though LCLs have several advantages and have been used in a number of published studies as a model system for blood [381], there are some limitations. The increased proliferation rate increases chromosomal instability and DNA damage events in the LCLs, particularly at early stages of EBV-transformation [391, 392]. Large inter-experimental variability has been observed in LCLs following DNA damage experiments, including the γ H2AX assay [392, 393]. Furthermore, differences in DNA methylation between LCLs, whole blood and peripheral blood cells including leukocytes have been observed, suggesting that LCLs should be used with caution as a model system in DNA methylation experiments [394-397]. However, it is not unexpected that DNA methylation will differ between LCLs and other cell types. These are important considerations when designing experiments in LCLs, however, the purpose of this study was not to compare with other blood cell types, but to explore changes in DNA methylation following oestrogen treatment. To note, the mean DNA methylation levels in the LCLs at the four target CpG sites differed greatly from the Generations Study targeted sequencing data (*CTNNA2*: 9% in the LCLs vs 77% in the Generations Study, *GRB10*: 34% vs 79%, *RPH3AL*: 42% vs 86%, *TINCR*: 64% vs 49%).

No significant changes in DNA methylation were observed between untreated or treated LCLs in the five assays, but significant individual-specific differences were observed for some of the samples for the *CTNNA2*, *RPH3AL* and *TINCR* assays with mainly decrease in DNA methylation following oestrogen treatment. This is in line with previous results where hypomethylation was observed in incident breast cancer cases in the Generations Study targeted sequencing data in **Table 3.9**. Interestingly, *BRCA1* and *BRCA2*-mutation phenotypic DNA methylation levels were observed for the *CTNNA2*, *RPH3AL* and *TINCR* assays. This might suggest that the DNA methylation levels are different in samples with an impaired DNA repair pathway, but will need further investigation.

There are several limitations to this study that should be acknowledged. Firstly, more samples should be analysed for DNA damage induced changes, as this was only performed on one individual. Further improvements of the experiments would involve optimisation of the right treatment dose and treatment time of E2 for the LCLs, further optimisation of ER- β western blotting, and optimisation of the protocol for γ H2AX immunofluorescence staining. Other further experiments could be to use whole blood or PBMCs instead of LCLs, to analyse the expression of other ER isoforms and the GPER1, and to analyse DNA methylation using a genome-wide array or other experimental technique to explore a larger number of CpG sites following oestrogen treatment.

5.4.1 Conclusions

In this chapter, two main mechanisms of oestrogen-induced DNA methylation were explored in LCLs. The results might suggest a potential role for increased DNA damage response following oestrogen treatment, and limited transcriptional response due to the non-detectable expression levels of the ERs. However, no change in DNA methylation was observed after oestrogen-treatment at the five analysed pyrosequencing assays, but modest individual specific changes were observed. These are preliminary results and further optimisation and replication are needed to draw stronger conclusions.

CHAPTER 6:
Discussion, Future Research
Directions, and Conclusions

6 CHAPTER 6: Discussion, Future Research Directions and Conclusions

6.1. GENERAL DISCUSSION

The hypothesis for this project was that lifetime oestrogen exposure accumulates DNA methylation changes detectable in blood, which can be used for breast cancer risk assessment. The main research question can be divided into three parts. Firstly: does lifetime oestrogen exposure affect DNA methylation in blood? Secondly: can a blood DNA methylation signature be used to assess a woman's risk of developing breast cancer? And lastly, by combining the first two questions: can a DNA methylation signature reflecting a woman's lifetime oestrogen exposure be used to assess her breast cancer risk? In addition to this, the mechanisms between oestrogen exposure and blood DNA methylation have been explored.

6.1.1 Lifetime oestrogen exposure and DNA methylation

It was hypothesised that lifetime oestrogen exposure accumulated DNA methylation changes detectable in blood. An EWAS of the Estimated Lifetime (o)Estrogen Exposure (ELEE) model was conducted, and this is, to our knowledge, the first EWAS of lifetime oestrogen exposure. A total of 694 CpG sites were found to be associated with ELEE after correction for multiple testing (**Figure 2.3A**). Somewhat unexpectedly, these CpG sites were enriched in active promoter regions and CpG islands with mainly an increase in DNA methylation levels associated with longer ELEE (**Figure 2.6**), which would mean that the outcome of longer ELEE is more repression of genes. This was partly unexpected since it was thought that increasing ELEE would lead to higher expression of ER-target genes. However, promoter hypermethylation is in line with previous observations in breast cancer [197, 204, 206], and might support the hypothesis of changes in DNA methylation following oestrogen-induced DNA damage in blood (further discussed in section 6.1.3). Furthermore, the CpG sites associated with ELEE were also enriched in CpG shelves (regions located 2-4K base pairs outside of CpG islands) with mainly a decrease in DNA methylation levels (**Figure 2.6**). It is not clear what affect this will have on the transcription or the genomic stability, especially since the opposite, hypermethylation, was found in CpG islands. In the other genomic regions (gene body, intergenic, and open sea) the DNA methylation

generally decreased with longer ELEE, and the DNA methylation levels for a majority of the 694 CpG sites (61%) were decreasing with longer ELEE (**Figure 2.3B**). This is in accordance with the genome-wide DNA hypomethylation that has been recognised in breast cancer, even though results are inconsistent [195, 197, 204-206]. In summary, the results from the EWAS of ELEE are similar to what has been found in the breast cancer epigenome, and might mean that DNA methylation following lifetime oestrogen exposure can be used as a biomarker for breast cancer risk.

Little emphasis has been put on the interpretation of the genes identified in this project. This is mostly relevant, and easier to interpret, if the methylation is located in gene promoters or enhancer regions, where an increase in DNA methylation would lead to silencing of the target gene. The top hits reported in **Table 2.4**, however, consisted of CpG sites in several different functional genomic regions, and some CpG sites were located far away from the nearest gene, which makes it difficult to link to a gene function.

As summarised in **Chapter 2**, the findings from previous EWAS for blood DNA methylation and breast cancer risk have not replicated across study cohorts [197, 204, 205, 288]. It was hypothesised that a different approach compared with the traditional case-control design might be more successful in identifying blood DNA methylation changes associated with breast cancer [208]. Therefore, a risk factor focused EWAS was conducted with ELEE as the risk exposure and DNA methylation as the outcome. The idea was to first identify CpG sites where DNA methylation levels were associated with ELEE, and then link this with breast cancer risk in a second step. This has been assessed previously with smoking as the exposure, where DNA methylation signatures can be used to discriminate smokers from never-smokers with high accuracy, and can be used to predict and mediate the risk for lung cancer [211, 220]. Furthermore, the DNA methylation signature could be used as an intermediate molecular phenotype that might be more accurate in estimating the risk exposure, compared with questionnaire-based data. This signature could be used in further experiments to test whether other exposures contribute to the ELEE, such as xenoestrogens. Unfortunately, no DNA methylation signature of lifetime oestrogen exposure that replicated across study cohorts was identified in this project (**Table 4.5**), and further work is needed to replicate the analysis in larger datasets.

The risk factor analysed in the EWAS was a model for the estimated lifetime oestrogen exposure (ELEE). Different hormonal factors, all known to affect oestrogen exposure, have been strongly associated with breast cancer risk, as summarised in the introduction (**Table 1.1**) and

replicated in the EPIC-Italy questionnaire data in this study (**Figure 2.2 and Appendix 1**). By combining risk factors together, it was hypothesised that a woman's cumulative oestrogen exposure up to the time at blood draw could be estimated. The ELEE model in this study included a woman's reproductive time, from which was subtracted one year for each full-term pregnancy and total lifetime breastfeeding duration (in years). This will roughly estimate a woman's total lifetime oestrogen exposure, given that parous women have decreased oestrogen levels in older ages compared to nulliparous women, and thereby a shorter ELEE value (**Figure 6.1**). Women with more pregnancies and longer breastfeeding duration will have a shorter ELEE, compared with nulliparous women.

The ELEE-model in this project is not a perfect model, and does not include all factors that affect a woman's exposure to oestrogens. However, even though it is well established that all risk factors contribute to breast cancer risk, it is not clear how all the different hormonal risk factors would contribute to the total oestrogen exposure. For example, only current or former use of HRT or OC at time of blood drawn might add into the total measured exposure, but not in a long-term perspective if the exposure was several years ago. Increase in breast cancer risk is only seen for current use and up to 10 years after cessation, while the oestrogen exposure would only be temporarily increased [95, 121, 126, 132, 133]. Relevant to this project, it is also possible that the DNA methylation changes following OC and HRT use are reversible as well after cessation. Also, as already discussed, BMI should be modelled differently in pre- and postmenopausal women [43, 46], and the oestrogen exposure is further affected by menstrual cycle regularity, length, and this is often changed after FFTP [110, 111]. During pregnancy the oestrogen levels are increased dramatically, but the long-term effect is a decrease in oestrogen levels compared to nulliparous women [110, 111]. It is not clear what the total effect in oestrogen exposure from have being pregnant will be, and the ELEE model in this project only included the long-term decrease.

In conclusion, lifetime oestrogen exposure is complex to model, and there are yet many unanswered questions in the mechanisms for different hormonal risk factors. The perfect ELEE model might include age at menarche, regularity and length of menstrual cycles before and after FFTP, number of pregnancies, duration breastfeeding, current or recent OC and HRT use, premenopausal BMI, postmenopausal BMI, and adult weight gain (**Figure 6.1**). Furthermore, lifestyle factors such as average lifetime alcohol consumption, average lifetime physical activity, and postmenopausal smoking will contribute to the exposure. How much each factor should be weighted to contribute to the ELEE

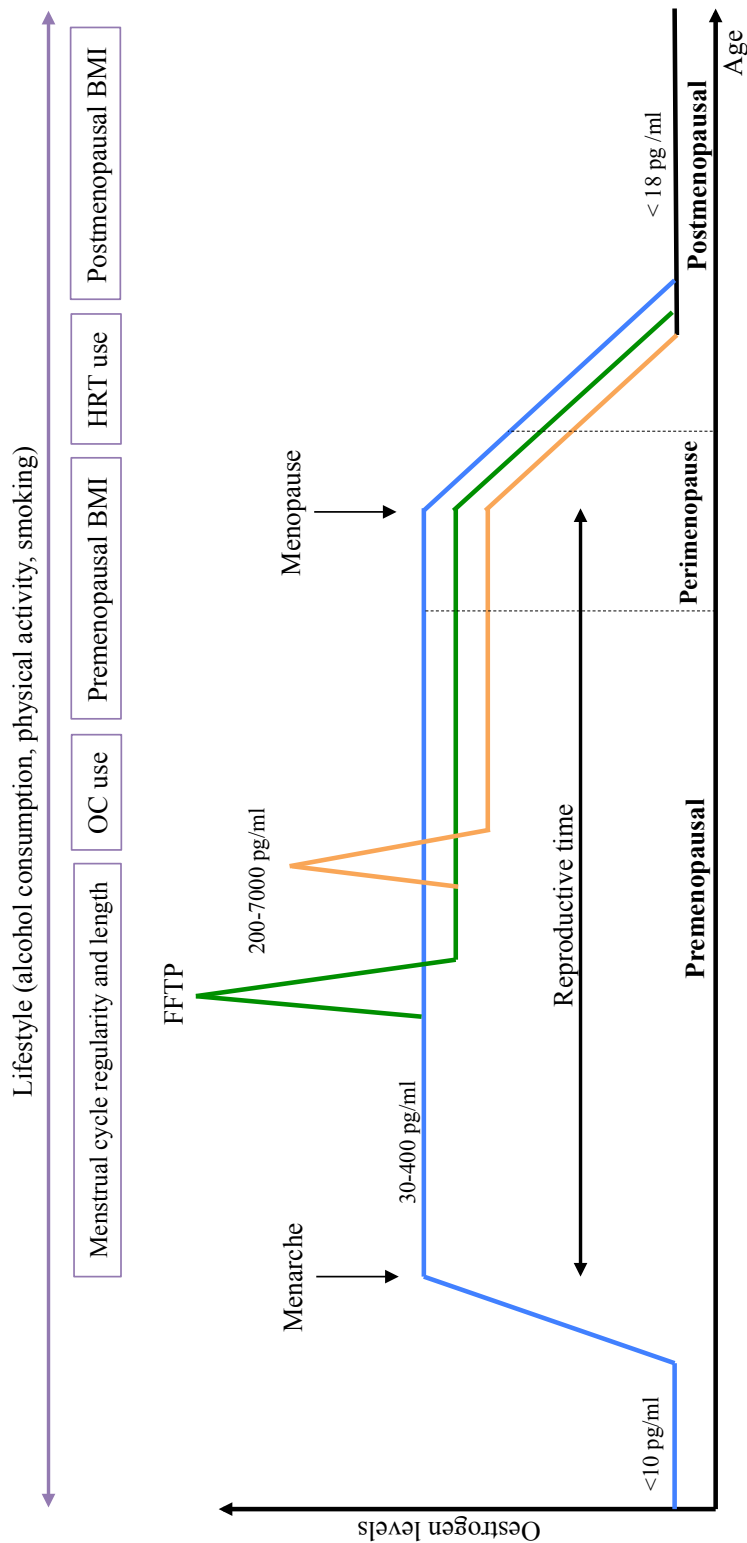


Figure 6.1 Estimating a woman's total lifetime oestrogen exposure

A woman is exposed to high levels of circulating oestrogens produced in the ovaries between age at menarche and age at menopause (the reproductive time). During pregnancy, a woman is exposed to high levels of oestrogens, but the long-term effect from having pregnant is a decrease in oestrogen levels compared to nulliparous women (the figure demonstrates a nulliparous woman in blue, one pregnancy in green, and two pregnancies in orange). After menopause, the oestrogen levels are decreased dramatically. The lifetime oestrogen exposure is affected by menstrual cycle regularity and length, use of OC or HRT, and BMI, which has different outcomes in pre- and postmenopausal women. Furthermore, different lifestyle factors will affect a woman's exposure to oestrogens, such as alcohol consumption, physical activity, and smoking. Oestrogen levels in young women before menarche, in premenopausal women, during pregnancy, and postmenopausal women from Abbassi-Ghanavati *et al.* 2009 [113].

model is, however, also not known. While the ELEE model used in this project might not be perfect, it captures the most important life events to estimate differences in lifetime oestrogen exposure between women, and includes the hormonal risk factors with the strongest knowledge about how they contribute to the lifetime oestrogen exposure.

6.1.2 Blood DNA methylation and breast cancer risk

It was hypothesised that a blood DNA methylation signature could be used to estimate a woman's risk for developing breast cancer. The Methylation Index (MI) for breast cancer risk in this study was validated in a large independent study cohort using targeted bisulphite sequencing, and was replicated in a meta-analysis including four different prospective study cohorts. This is one of the few DNA methylation signatures for breast cancer risk that has validated and replicated, although modestly, across several study cohorts.

The idea of using blood DNA methylation as a biomarker for breast cancer risk is becoming more accepted [187, 208, 334]. It is well established that alterations in DNA methylation is an early event in breast cancer [178, 179, 187, 188, 195, 267], and prospective studies have been able to identify changes in blood DNA methylation years before breast cancer diagnosis, supporting the use of a DNA methylation biomarker in blood [195, 197, 199, 204-208, 288, 289]. DNA methylation is modified by endogenous and exogenous exposures, and would in theory capture the exposure to different risk factors and thereby be able to quantitatively reflect the breast cancer risk. Also, DNA methylation is a relatively stable marker over time [207]. For example, it has been shown that DNA methylation signatures associated with smoking can remain up to 20-40 years after cessation [76, 221]. Furthermore, blood sampling is an accessible, less-invasive and relatively easy method to add into a population-screening programme, if a validated biomarker can be identified.

If a blood-based DNA methylation biomarker should be incorporated in population risk screening, it needs to be strongly and independently associated with breast cancer risk, validated and replicated across populations, and vary sufficiently between individuals to accurately stratify women into low and high-risk groups. Even though the association between the MI and breast cancer risk was both validated and significant in the meta-analysis in this project, the association was modest, which might not be sufficient to be used for risk screening, and further improvements of the model is needed. An increase in risk by 4% per one unit increase in the MI was observed in the meta-analysis, but a 45%

increase between the highest and lowest quartiles (**Figure 4.1 and 4.2**). This might indicate that the MI is better at estimating the risk for the women with lowest and highest risk, but will also not differentiate women in the middle-risk groups. Further studies are needed to show how efficiently this can be done and what thresholds should be used. Alone, the MI could classify women into lowest and highest quartile risk groups in the Generations Study with an estimated AUC of 0.61 (95% CI: 0.55 – 0.67, **Figure 3.7**), which is around the lower range reported for the Gail or Tyrer-Cuzick models [141]. Further studies are needed to determine if the MI can add valuable information to the risk models, in combination with other risk estimations such as polygenic risk scores.

In summary, a methylation signature in blood DNA has high potential to improve breast cancer risk assessment. A blood DNA methylation signature for breast cancer risk was identified in this project, but further optimisation and validation of the signature is needed before it can be assessed if it might add valuable information to breast cancer risk models, and improve risk-stratification for women.

6.1.3 Lifetime oestrogen exposure, blood DNA methylation, and breast cancer risk

It was hypothesised that the lifetime oestrogen exposure accumulates changes in DNA methylation, which can be detectable in blood and are associated with breast cancer risk. It is well established that oestrogens are involved in breast cancer development, and there is evidence that oestrogen exposure can affect DNA methylation [262-268]. A methylation signature for breast cancer risk was identified, however, the signature did not correlate with ELEE in other study cohorts examined, except in the EPIC-Italy HM450K data in which the signature was developed. This result was unexpected and is difficult to interpret. It is possible that the association between the MI and ELEE is a false positive and that the MI was over fitted in the EPIC-Italy data. Another suggestion might be that the MI captures the biological effects of lifetime oestrogen exposure more accurately than the ELEE model developed from questionnaire data itself, including any confounders that might affect ELEE. This could explain why an association between the MI and breast cancer risk was observed but no correlation between the MI and ELEE. Furthermore, differences between the study cohorts (or sub-cohorts for EPIC-Italy) that might affect both DNA methylation and lifetime oestrogen exposure cannot be excluded, including smoking, which is well established to affect DNA methylation, and other lifestyle and reproductive factors.

Furthermore, the combination of risk factors in the ELEE-model might be capturing something else other than lifetime oestrogen exposure or be confounded by other factors.

The mechanisms between oestrogen exposure and DNA methylation in blood were explored in this project. Two main mechanisms were hypothesised: either that DNA methylation is altered through a transcriptional, ER-dependent mechanism, or by DNA damage from different oestrogen metabolites. The expression levels of the ERs and DNA damaging response following oestrogen treatment were explored in LCLs, as a model system for blood DNA. No expression of either ER- α or ER- β was detected (**Figure 5.7**), however, some indication of DNA damage following increasing concentrations of oestrogen treatment was observed (**Figure 5.3**), but no change in DNA methylation levels was found in the five analysed assays (**Figure 5.4**). Interestingly, variation between individual samples and mutation-phenotypic DNA methylation profiles was observed for some of the methylation assays (**Figure 5.5D and 5.6**). Therefore, these studies in LCLs did not provide evidence to conclude which mechanism is most relevant to the methylation changes observed in blood DNA. It is acknowledged that different results for ER-expression or DNA damage response might have been observed in other blood cell types.

The EWAS of ELEE identified mainly an increase of DNA methylation levels in active promoter regions. This may be in line with the indications that higher oestrogen exposure increases DNA damage. Active promoters have low methylation and open chromatin for accessibility for the transcriptional machinery [177, 178], which makes them more prone for DNA damaging agents. It was hypothesised that DNA repair following DNA damage introduced by oestrogens would result in increased methylation at the damaged site, as a protective mechanism [261]. In *BRCA1* or *BRCA2*-mutation carriers the DNA repair-responses are impaired, which would result in less hypermethylation. This was, however, not observed in the five assays analysed in this project, but more evidence is needed to conclude this. It is more likely that there is interplay between different mechanisms for oestrogen-induced DNA methylation changes in different blood cell types, which can also vary between individuals. For example, alterations in the oestrogen metabolism with more production of 4-OHE2 might lead to more DNA damage than transcriptional response, which has been observed in breast cancer cases [251, 254].

From these experiments and previous studies, the conclusion would be that oestrogen causes changes in DNA methylation in blood through a number of different mechanisms, including change in

transcription and through DNA repair mechanisms following DNA damage. This has been suggested previously in breast cancer [398]. However, our results might indicate a stronger effect from DNA damage in blood that result in DNA methylation alterations, rather than transcriptional changes, due to the absence of ER expression.

6.2. STRENGTHS AND LIMITATIONS

All analyses in this study were conducted in prospective study cohorts, with blood samples taken years before diagnosis. This is important when exploring risk biomarkers, where retrospective data can be confounded by reverse causation, i.e. that the cancer itself is causing changes in DNA methylation. The results in this project did not find any signs of a reverse causation, but potentially the opposite: the association with breast cancer risk was reduced when date at diagnosis was closer to when the blood samples were taken, suggesting that the cancer progression might influence the methylation signature by increasing variability.

The initial EWAS of ELEE was limited by the small sample size in EPIC-Italy ($n=216$), which can influence the P -values and give biased results. To avoid focusing on potentially false positives, the rest of the analyses were focused on the CpG sites with largest changes in DNA methylation per year increase in ELEE, and with the most statistically significant P -values. The 694 CpG sites were filtered down to 212 CpG sites (**Appendix 2**). However, the P -values were still relatively large (maximum P -value was 9×10^{-5}), in comparison to a genome-wide significance of $P < 1 \times 10^{-7}$ that is often used. Only four CpG sites would pass this stringent threshold. Similar P -value thresholds have been observed in other EWAS: Xu *et al.* [288] reported a significance level of approximately $-\log_{10}(3.2) = 6 \times 10^{-4}$ for a sample size of 910, and Joo *et al.* [289] used a threshold of 5×10^{-5} with a sample size of 210. The low statistical power seems to be a common problem for EWAS, and therefore, future larger collaborations are recommended.

Importantly, the EWAS was limited by the HM450K array that only includes less than 2% of all CpG sites in the genome. Even though the array was designed to cover CpG islands, gene bodies, and regulatory elements, and it was assumed that CpG sites located close to each other were correlated, much important information is missing. However, for now, prior to the release of the Illumina HumanMethylation850K (HM850K) array, the HM450K array was the most cost-effective alternative for large DNA methylation studies.

A strength in this study was the use of newly recommended statistical methods in the EWAS. Recent publications have argued that linear regression on M-values might not be appropriate, however, there is yet no standardised method for EWAS. Statistical methods to use in EWAS are still under discussion and in process [307, 308, 318]. This includes also the pre-processing and normalisation

steps of the array data, which are also under development [399, 400]. Technical variation will be introduced from the laboratory and data analysis steps, including bisulphite conversion efficacy, variation in PCR amplification, batch effects, differences in normalisation of the data, WBC composition adjustments, and statistical methods in the association analysis. I believe future EWAS will benefit from having a more standardised pipeline to follow, to allow easier replication in future studies.

Another limitation was the inclusion of incident cases in the EWAS of ELEE. Since ELEE is associated with breast cancer risk in EPIC-Italy, the inclusion of cases, even though prospective data was used with blood samples collected prior to diagnosis, might affect the results. However, not including cases would have reduced the statistical power considerably. Sensitivity analyses of both the EWAS and development of MI were conducted. This showed that the results from the EWAS correlated well in controls-only EWAS (**Figure 2.5**), and that the MI developed in controls only showed similar association with breast cancer in the Generations Study (**Figure 3.8**). This suggests that the inclusion of cases in the initial EWAS did not affect the final results substantially.

Another strength of this work was the validation of the MI in a large independent study cohort using an independent method, and the meta-analysis including one additional sub-cohort from EPIC-Italy and two additional, independent study cohorts. For the validation step, the Fluidigm 48.48 Access Array was used, which let us include a total of 48 CpG sites in this step. This is still a small number of CpG sites compared with the 694 CpG hits, but many more than if bisulphite pyrosequencing would have been used. Five CpG sites to estimate WBC composition in the sequencing data were also included, which is important to consider when working in a mixture of cell types. Also, a test phase prior to sequencing in the Generations Study was performed, which meant I was familiar with the procedure and could analyse the quality for each assay, which probably resulted in the generation of high-quality data of 31 CpG sites that were included in the MI. However, the model had to be developed after the targeted sequencing, since it was dependent on the quality of the assays in the sequencing data.

Furthermore, this project also included exploratory analyses of the mechanisms between oestrogen exposure and DNA methylation. While these experiments were limited in scope, they gave indications that oestrogens might affect DNA methylation in a number of different ways, which needs to be taken into account in future studies.

Taking these limitations and strengths into account, how would the project have been conducted if it was started over today? Firstly, if sufficient data was available, the EWAS would have been conducted in healthy controls (not diagnosed with breast cancer) only, and the MI model would have been developed in controls only. If possible, replication steps of the ELEE model and association between the MI and ELEE would have been conducted before exploring the association with breast cancer. I would have considered including more variables in the ELEE model, if all information were available from the questionnaire data. After this, the association between the MI and breast cancer risk would have been assessed in case-control studies, including the validation step and meta-analysis.

6.3. FUTURE DIRECTIONS

The EWAS of ELEE needs further replication in larger study cohorts, to identify potential false positive CpG sites from this project, or new CpG sites that were not detected due to insufficient statistical power. This will provide further insights into where in the genome changes in DNA methylation levels occur in association with lifetime oestrogen exposure. The replication would preferably be conducted in controls only that have remained cancer-free several years after blood drawn, to exclude the possibility of any confounding factors from prospective breast cancer cases. Also, combinations of other hormonal risk factors not included in the ELEE model in this project might be explored, such as current or recent HRT use in postmenopausal women.

Large collaborations are needed for EWAS to gain sufficient statistical power, similar to the studies on DNA methylation and BMI by Wahl *et al.* [213] and alcohol consumption by Liu *et al.* [299] with 10,000 and 7,000 subjects respectively. During this project I had the opportunity to participate in a meta-analysis of EWAS for breast cancer risk (Bodelon *et al.*, manuscript under review) including approximately 3,500 samples in total. However, compared to large GWAS with several 100,000 samples [37, 401, 402], this is still a relatively small number of samples.

Further replication of the EWAS of ELEE might give indications of other CpG sites that could be included in the MI model. In addition to this, further replication and validation are needed to improve and calibrate the MI model in large population study cohorts to identify a strong breast cancer risk biomarker. This could be combined with longitudinal studies, to determine the stability of the DNA methylation changes for the CpG sites included in the MI. For example, it could be explored if the DNA methylation levels are stable in the transition from pre- to postmenopausal status. Also, it could be explored if different lifestyle changes, such as weight loss, reduction in alcohol drinking, increased physical activity or quitting smoking alter the MI and thereby result in lower risk estimation.

Furthermore, it needs to be explored how the MI can be incorporated into population screening for breast cancer risk, and how to combine the MI in addition to different existing risk models, polygenic risk scores, breast density measurements, and other reproductive and lifestyle factors for better breast cancer risk prediction. Finally, the thresholds to stratify women into low or high-risk groups need to be determined.

To further explore the changes in transcription following longer ELEE, the EWAS of ELEE could be combined with gene expression data, to explore which genes are affected with longer ELEE. Also, ChIP-seq data for the ERs in blood might add useful information to explore if the CpG sites associated with ELEE are enriched in ER-binding sites or ER-target genes. The expression levels of the ERs limit the generation of high-quality ChIP-seq data, so it might be recommended to first determine the levels of ER expression in different blood cell types. For this, another ER- β antibody than the one used in this project and to include the analysis of the expression of GPER1 could be considered.

The experiments to analyse the mechanisms between oestrogen exposure and DNA methylation in this project needs further optimisation. Since DNA methylation changes following oestrogen treatment varied between individuals, a larger number of samples for each phenotype might be needed to identify significant differences between wild-type, *BRC1*, and *BRC2* mutation carriers. The doses and time of the oestrogen treatment for the LCLs might need to be optimised, especially how long to treat the cells to first be able to see changes in DNA damage (which is expected to be short) and afterwards DNA methylation changes following that DNA repair mechanisms have been activated. The DNA damage quantification protocol using the γ H2AX-assay needs further optimisation to reduce the background noise. This might be a matter of the treatment time for the cells, but also other DNA damage assays (such as the comet assay [403]) might be more appropriate in LCLs. Additionally, with further optimisation of the MI, DNA methylation levels following DNA damage could be explored at other CpG sites in the pyrosequencing assays.

In this project the focus has been on lifetime oestrogen exposure as a risk factor for breast cancer, but similar analyses could be conducted for other risk factors. For example, methylation signatures that have been identified for alcohol consumption, BMI (potentially different signatures for pre- and postmenopausal women), epigenetic age acceleration, and smoking could be used [213-215, 299]. The risk factor methylation signatures could then be tested in case-control studies for prediction of breast cancer, and might add further information for risk assessment.

6.4. BREAST CANCER RISK AND PREVENTION

The incidence rate for breast cancer is increasing, and new approaches to reduce the economical and emotional burden of the disease need to be explored. For now, it is expected that 1 in 7 women will be diagnosed with breast cancer [164] (this was increased from 1 of 8 women at the start of this project). If women of high risk for breast cancer can be identified, they can be targeted with chemoprevention to efficiently reduce the number of breast cancer cases [156, 157, 159, 160], or lifestyle changes. However, as proven in the introduction, the risk models are for now not able to stratify women into low and high-risk groups efficiently, and will always miss some cases that will be grouped in the low-risk group. The recently updated BOADICEA model has the potential to improve risk stratification, but needs further validation [147]. Therefore, lifestyle changes, or guidelines, should be relevant for the entire population to additionally reduce the number of cases that cannot be identified in the risk models. Risk assessment and risk information may be efficiently included in screening occasions, including a blood sample for genetic and epigenetic testing. Risk assessment needs to be simple and economical. Nevertheless, the estimation of risk for an individual woman is complex. The interplay between family history, breast density, risk-SNPs and *BRCA1* and *BRCA2* mutations, reproductive and lifestyle factors, and other unknown environmental exposures will contribute to the risk picture. Future research including evidence for intervention strategies, biological mechanisms leading to breast cancer progression, and analysis of biomarkers are needed to improve risk models. To be able to prevent and reduce the breast cancer incidences in the population, research needs to be embedded in health guidelines and communicated to the public.

6.5. FINAL CONCLUSIONS

In this project, the association between lifetime oestrogen exposure, DNA methylation in blood and breast cancer risk has been assessed. Evidence that lifetime oestrogen exposure accumulates DNA methylation in blood has been generated, and potential mechanisms for this have been proposed. Importantly, a blood DNA methylation signature for breast cancer risk that validated and replicated across studies has been developed. This might add valuable information to breast cancer risk assessment, and thereby prevent the number of breast cancer cases. The results from this project will hopefully add new clues, perspectives, and data that can be used in future studies to further understand the biological mechanisms of breast cancer development.

REFERENCES

1. Cancer Research UK. *Breast cancer statistics*. [Accessed November 2018]; Available from: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer>.
2. Bray F., Ferlay J., Soerjomataram I., Siegel R.L., Torre L.A., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA Cancer J Clin, 2018.
3. Mandelblatt J.S., Stout N.K., Schechter C.B., van den Broek J.J., Miglioretti D.L., et al., *Collaborative Modeling of the Benefits and Harms Associated With Different U.S. Breast Cancer Screening Strategies*. Ann Intern Med, 2016. **164**(4): p. 215-25.
4. Johns L.E., Coleman D.A., Swerdlow A.J. and Moss S.M., *Effect of population breast screening on breast cancer mortality up to 2005 in England and Wales: an individual-level cohort study*. Br J Cancer, 2017. **116**(2): p. 246-252.
5. Eccles S.A., Aboagye E.O., Ali S., Anderson A.S., Armes J., et al., *Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer*. Breast Cancer Res, 2013. **15**(5): p. R92.
6. Perou C.M., Sorlie T., Eisen M.B., van de Rijn M., Jeffrey S.S., et al., *Molecular portraits of human breast tumours*. Nature, 2000. **406**(6797): p. 747-52.
7. Virnig B.A., Tuttle T.M., Shamlivan T. and Kane R.L., *Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes*. J Natl Cancer Inst, 2010. **102**(3): p. 170-8.
8. Sinn H.P. and Kreipe H., *A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition*. Breast Care (Basel), 2013. **8**(2): p. 149-54.
9. Guiu S., Wolfer A., Jacot W., Fumoleau P., Romieu G., et al., *Invasive lobular breast cancer and its variants: how special are they for systemic therapy decisions?* Crit Rev Oncol Hematol, 2014. **92**(3): p. 235-57.
10. McCart Reed A.E., Kutasovic J.R., Lakhani S.R. and Simpson P.T., *Invasive lobular carcinoma of the breast: morphology, biomarkers and 'omics*. Breast Cancer Res, 2015. **17**: p. 12.
11. Carroll J.S., *Mechanisms of oestrogen receptor (ER) gene regulation in breast cancer*. Eur J Endocrinol, 2016. **175**(1): p. R41-9.
12. Baird R.D. and Carroll J.S., *Understanding Oestrogen Receptor Function in Breast Cancer and its Interaction with the Progesterone Receptor. New Preclinical Findings and their Clinical Implications*. Clin Oncol (R Coll Radiol), 2016. **28**(1): p. 1-3.
13. Schnitt S.J., *Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy*. Mod Pathol, 2010. **23 Suppl 2**: p. S60-4.
14. Sabatier R., Finetti P., Guille A., Adelaide J., Chaffanet M., et al., *Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization*. Mol Cancer, 2014. **13**: p. 228.

15. Ali S., Buluwela L. and Coombes R.C., *Antiestrogens and their therapeutic applications in breast cancer and other diseases*. Annu Rev Med, 2011. **62**: p. 217-32.
16. Howell S.J., Johnston S.R. and Howell A., *The use of selective estrogen receptor modulators and selective estrogen receptor down-regulators in breast cancer*. Best Pract Res Clin Endocrinol Metab, 2004. **18**(1): p. 47-66.
17. Maximov P.Y., Lee T.M. and Jordan V.C., *The discovery and development of selective estrogen receptor modulators (SERMs) for clinical practice*. Curr Clin Pharmacol, 2013. **8**(2): p. 135-55.
18. Torrisi R., Rota S., Losurdo A., Zuradelli M., Masci G., et al., *Aromatase inhibitors in premenopause: Great expectations fulfilled?* Crit Rev Oncol Hematol, 2016. **107**: p. 82-89.
19. Smith I.E. and Dowsett M., *Aromatase inhibitors in breast cancer*. N Engl J Med, 2003. **348**(24): p. 2431-42.
20. Ma C.X., Reinert T., Chmielewska I. and Ellis M.J., *Mechanisms of aromatase inhibitor resistance*. Nat Rev Cancer, 2015. **15**(5): p. 261-75.
21. Chumsri S., Howes T., Bao T., Sabnis G. and Brodie A., *Aromatase, aromatase inhibitors, and breast cancer*. J Steroid Biochem Mol Biol, 2011. **125**(1-2): p. 13-22.
22. Ali S. and Coombes R.C., *Endocrine-responsive breast cancer and strategies for combating resistance*. Nat Rev Cancer, 2002. **2**(2): p. 101-12.
23. Martin L.A., Ribas R., Simigdala N., Schuster E., Pancholi S., et al., *Discovery of naturally occurring ESR1 mutations in breast cancer cell lines modelling endocrine resistance*. Nat Commun, 2017. **8**(1): p. 1865.
24. Nguyen V.T., Barozzi I., Faronato M., Lombardo Y., Steel J.H., et al., *Differential epigenetic reprogramming in response to specific endocrine therapies promotes cholesterol biosynthesis and cellular invasion*. Nat Commun, 2015. **6**: p. 10044.
25. Tutt A., Tovey H., Cheang M.C.U., Kernaghan S., Kilburn L., et al., *Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: the TNT Trial*. Nat Med, 2018. **24**(5): p. 628-637.
26. Cameron D., Piccart-Gebhart M.J., Gelber R.D., Procter M., Goldhirsch A., et al., *11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial*. Lancet, 2017. **389**(10075): p. 1195-1205.
27. Livraghi L. and Garber J.E., *PARP inhibitors in the management of breast cancer: current data and future prospects*. BMC Med, 2015. **13**: p. 188.
28. Collaborative Group on Hormonal Factors in Breast Cancer, *Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease*. Lancet, 2001. **358**(9291): p. 1389-99.
29. Lichtenstein P., Holm N.V., Verkasalo P.K., Iliadou A., Kaprio J., et al., *Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland*. N Engl J Med, 2000. **343**(2): p. 78-85.
30. Antoniou A., Pharoah P.D., Narod S., Risch H.A., Eyfjord J.E., et al., *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series*

- unselected for family history: a combined analysis of 22 studies.* Am J Hum Genet, 2003. **72**(5): p. 1117-30.
31. Yoshida K. and Miki Y., *Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage.* Cancer Sci, 2004. **95**(11): p. 866-71.
 32. Malone K.E., Daling J.R., Doody D.R., Hsu L., Bernstein L., et al., *Prevalence and predictors of BRCA1 and BRCA2 mutations in a population-based study of breast cancer in white and black American women ages 35 to 64 years.* Cancer Res, 2006. **66**(16): p. 8297-308.
 33. Anglian Breast Cancer Study Group, *Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases.* Anglian Breast Cancer Study Group. Br J Cancer, 2000. **83**(10): p. 1301-8.
 34. Easton D.F., Pharoah P.D., Antoniou A.C., Tischkowitz M., Tavtigian S.V., et al., *Gene-panel sequencing and the prediction of breast-cancer risk.* N Engl J Med, 2015. **372**(23): p. 2243-57.
 35. Turnbull C. and Rahman N., *Genetic predisposition to breast cancer: past, present, and future.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 321-45.
 36. Michailidou K., Hall P., Gonzalez-Neira A., Ghoussaini M., Dennis J., et al., *Large-scale genotyping identifies 41 new loci associated with breast cancer risk.* Nat Genet, 2013. **45**(4): p. 353-61, 361e1-2.
 37. Michailidou K., Lindstrom S., Dennis J., Beesley J., Hui S., et al., *Association analysis identifies 65 new breast cancer risk loci.* Nature, 2017. **551**(7678): p. 92-94.
 38. Cuzick J., Brentnall A.R., Segal C., Byers H., Reuter C., et al., *Impact of a Panel of 88 Single Nucleotide Polymorphisms on the Risk of Breast Cancer in High-Risk Women: Results From Two Randomized Tamoxifen Prevention Trials.* J Clin Oncol, 2017. **35**(7): p. 743-750.
 39. Mavaddat N., Pharoah P.D., Michailidou K., Tyrer J., Brook M.N., et al., *Prediction of breast cancer risk based on profiling with common genetic variants.* J Natl Cancer Inst, 2015. **107**(5).
 40. Garcia-Closas M., Gunsoy N.B. and Chatterjee N., *Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer.* J Natl Cancer Inst, 2014. **106**(11).
 41. Rudolph A., Song M., Brook M.N., Milne R.L., Mavaddat N., et al., *Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the Breast Cancer Association Consortium.* Int J Epidemiol, 2018. **47**(2): p. 526-536.
 42. Endogenous Hormones Breast Cancer Collaborative Group, *Body mass index, serum sex hormones, and breast cancer risk in postmenopausal women.* J Natl Cancer Inst, 2003. **95**(16): p. 1218-26.
 43. Renehan A.G., Tyson M., Egger M., Heller R.F. and Zwahlen M., *Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies.* Lancet, 2008. **371**(9612): p. 569-78.
 44. Bergstrom A., Pisani P., Tenet V., Wolk A. and Adami H.O., *Overweight as an avoidable cause of cancer in Europe.* Int J Cancer, 2001. **91**(3): p. 421-30.

45. Lahmann P.H., Hoffmann K., Allen N., van Gils C.H., Khaw K.T., et al., *Body size and breast cancer risk: findings from the European Prospective Investigation into Cancer And Nutrition (EPIC)*. Int J Cancer, 2004. **111**(5): p. 762-71.
46. van den Brandt P.A., Spiegelman D., Yaun S.S., Adami H.O., Beeson L., et al., *Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk*. Am J Epidemiol, 2000. **152**(6): p. 514-27.
47. Morimoto L.M., White E., Chen Z., Chlebowski R.T., Hays J., et al., *Obesity, body size, and risk of postmenopausal breast cancer: the Women's Health Initiative (United States)*. Cancer Causes Control, 2002. **13**(8): p. 741-51.
48. Calle E.E. and Kaaks R., *Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms*. Nat Rev Cancer, 2004. **4**(8): p. 579-91.
49. Huang Z., Hankinson S.E., Colditz G.A., Stampfer M.J., Hunter D.J., et al., *Dual effects of weight and weight gain on breast cancer risk*. Jama, 1997. **278**(17): p. 1407-11.
50. Endogenous Hormones Breast Cancer Collaborative Group, *Circulating sex hormones and breast cancer risk factors in postmenopausal women: reanalysis of 13 studies*. Br J Cancer, 2011. **105**(5): p. 709-22.
51. Wang X., Simpson E.R. and Brown K.A., *Aromatase overexpression in dysfunctional adipose tissue links obesity to postmenopausal breast cancer*. J Steroid Biochem Mol Biol, 2015. **153**: p. 35-44.
52. Munsell M.F., Sprague B.L., Berry D.A., Chisholm G. and Trentham-Dietz A., *Body mass index and breast cancer risk according to postmenopausal estrogen-progestin use and hormone receptor status*. Epidemiol Rev, 2014. **36**(1): p. 114-36.
53. Suzuki R., Orsini N., Saji S., Key T.J. and Wolk A., *Body weight and incidence of breast cancer defined by estrogen and progesterone receptor status--a meta-analysis*. Int J Cancer, 2009. **124**(3): p. 698-712.
54. Gunter M.J., Wang T., Cushman M., Xue X., Wassertheil-Smoller S., et al., *Circulating Adipokines and Inflammatory Markers and Postmenopausal Breast Cancer Risk*. J Natl Cancer Inst, 2015. **107**(9).
55. Endogenous Hormones Breast Cancer Collaborative Group, *Sex hormones and risk of breast cancer in premenopausal women: a collaborative reanalysis of individual participant data from seven prospective studies*. Lancet Oncol, 2013. **14**(10): p. 1009-19.
56. Schoemaker M.J., Nichols H.B., Wright L.B., Brook M.N., Jones M.E., et al., *Association of Body Mass Index and Age With Subsequent Breast Cancer Risk in Premenopausal Women*. JAMA Oncol, 2018: p. e181771.
57. Pierobon M. and Frankenfeld C.L., *Obesity as a risk factor for triple-negative breast cancers: a systematic review and meta-analysis*. Breast Cancer Res Treat, 2013. **137**(1): p. 307-14.
58. Frydenberg H., Flote V.G., Larsson I.M., Barrett E.S., Furberg A.S., et al., *Alcohol consumption, endogenous estrogen and mammographic density among premenopausal women*. Breast Cancer Res, 2015. **17**: p. 103.

59. Romieu I., Scoccianti C., Chajes V., de Batlle J., Biessy C., et al., *Alcohol intake and breast cancer in the European prospective investigation into cancer and nutrition*. Int J Cancer, 2015. **137**(8): p. 1921-30.
60. Seitz H.K., Pelucchi C., Bagnardi V. and La Vecchia C., *Epidemiology and pathophysiology of alcohol and breast cancer: Update 2012*. Alcohol Alcohol, 2012. **47**(3): p. 204-12.
61. Collaborative Group on Hormonal Factors in Breast Cancer, *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, 2002. **87**(11): p. 1234-45.
62. Key J., Hodgson S., Omar R.Z., Jensen T.K., Thompson S.G., et al., *Meta-analysis of studies of alcohol and breast cancer with consideration of the methodological issues*. Cancer Causes Control, 2006. **17**(6): p. 759-70.
63. Allen N.E., Beral V., Casabonne D., Kan S.W., Reeves G.K., et al., *Moderate alcohol intake and cancer incidence in women*. J Natl Cancer Inst, 2009. **101**(5): p. 296-305.
64. Chen W.Y., Rosner B., Hankinson S.E., Colditz G.A. and Willett W.C., *Moderate alcohol consumption during adult life, drinking patterns, and breast cancer risk*. Jama, 2011. **306**(17): p. 1884-90.
65. Dumitrescu R.G. and Shields P.G., *The etiology of alcohol-induced breast cancer*. Alcohol, 2005. **35**(3): p. 213-25.
66. Varela-Rey M., Woodhoo A., Martinez-Chantar M.L., Mato J.M. and Lu S.C., *Alcohol, DNA methylation, and cancer*. Alcohol Res, 2013. **35**(1): p. 25-35.
67. Onland-Moret N.C., Peeters P.H., van der Schouw Y.T., Grobbee D.E. and van Gils C.H., *Alcohol and endogenous sex steroid levels in postmenopausal women: a cross-sectional study*. J Clin Endocrinol Metab, 2005. **90**(3): p. 1414-9.
68. Etique N., Chardard D., Chesnel A., Merlin J.L., Flament S., et al., *Ethanol stimulates proliferation, ERalpha and aromatase expression in MCF-7 human breast cancer cells*. Int J Mol Med, 2004. **13**(1): p. 149-55.
69. Fan S., Meng Q., Gao B., Grossman J., Yadegari M., et al., *Alcohol stimulates estrogen receptor signaling in human breast cancer cell lines*. Cancer Res, 2000. **60**(20): p. 5635-9.
70. Sellers T.A., Kushi L.H., Cerhan J.R., Vierkant R.A., Gapstur S.M., et al., *Dietary folate intake, alcohol, and risk of breast cancer in a prospective study of postmenopausal women*. Epidemiology, 2001. **12**(4): p. 420-8.
71. Jung S., Wang M., Anderson K., Baglietto L., Bergkvist L., et al., *Alcohol consumption and breast cancer risk by estrogen receptor status: in a pooled analysis of 20 studies*. Int J Epidemiol, 2016. **45**(3): p. 916-28.
72. Coglian V.J., Baan R., Straif K., Grosse Y., Lauby-Secretan B., et al., *Preventable exposures associated with human cancers*. J Natl Cancer Inst, 2011. **103**(24): p. 1827-39.
73. Dossus L., Boutron-Ruault M.C., Kaaks R., Gram I.T., Vilier A., et al., *Active and passive cigarette smoking and breast cancer risk: results from the EPIC cohort*. Int J Cancer, 2014. **134**(8): p. 1871-88.

74. Johnson K.C., *Accumulating evidence on passive and active smoking and breast cancer risk*. Int J Cancer, 2005. **117**(4): p. 619-28.
75. Jones M.E., Schoemaker M.J., Wright L.B., Ashworth A. and Swerdlow A.J., *Smoking and risk of breast cancer in the Generations Study cohort*. Breast Cancer Res, 2017. **19**(1): p. 118.
76. Luo J., Margolis K.L., Wactawski-Wende J., Horn K., Messina C., et al., *Association of active and passive smoking with risk of breast cancer among postmenopausal women: a prospective cohort study*. Bmj, 2011. **342**: p. d1016.
77. Nyante S.J., Gierach G.L., Dallal C.M., Freedman N.D., Park Y., et al., *Cigarette smoking and postmenopausal breast cancer risk in a prospective cohort*. Br J Cancer, 2014. **110**(9): p. 2339-47.
78. Catsburg C., Miller A.B. and Rohan T.E., *Active cigarette smoking and risk of breast cancer*. Int J Cancer, 2015. **136**(9): p. 2204-9.
79. Magnusson C., Wedrén S. and Rosenberg L.U., *Cigarette smoking and breast cancer risk: a population-based study in Sweden*. Br J Cancer, 2007. **97**(9): p. 1287-90.
80. Gu F., Caporaso N.E., Schairer C., Fortner R.T., Xu X., et al., *Urinary concentrations of estrogens and estrogen metabolites and smoking in caucasian women*. Cancer Epidemiol Biomarkers Prev, 2013. **22**(1): p. 58-68.
81. Tanko L.B. and Christiansen C., *An update on the antiestrogenic effect of smoking: a literature review with implications for researchers and practitioners*. Menopause, 2004. **11**(1): p. 104-9.
82. Collaborative Group on Hormonal Factors in Breast Cancer, *Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies*. Lancet Oncol, 2012. **13**(11): p. 1141-51.
83. Kyu H.H., Bachman V.F., Alexander L.T., Mumford J.E., Afshin A., et al., *Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013*. Bmj, 2016. **354**: p. i3857.
84. Lahmann P.H., Friedenreich C., Schuit A.J., Salvini S., Allen N.E., et al., *Physical activity and breast cancer risk: the European Prospective Investigation into Cancer and Nutrition*. Cancer Epidemiol Biomarkers Prev, 2007. **16**(1): p. 36-42.
85. Pizot C., Boniol M., Mullie P., Koechlin A., Boniol M., et al., *Physical activity, hormone replacement therapy and breast cancer risk: A meta-analysis of prospective studies*. Eur J Cancer, 2016. **52**: p. 138-54.
86. Rinaldi S., Kaaks R., Friedenreich C.M., Key T.J., Travis R., et al., *Physical activity, sex steroid, and growth factor concentrations in pre- and post-menopausal women: a cross-sectional study within the EPIC cohort*. Cancer Causes Control, 2014. **25**(1): p. 111-24.
87. Steindorf K., Ritte R., Eomois P.P., Lukanova A., Tjønneland A., et al., *Physical activity and risk of breast cancer overall and by hormone receptor status: the European prospective investigation into cancer and nutrition*. Int J Cancer, 2013. **132**(7): p. 1667-78.

88. Matthews C.E., Fortner R.T., Xu X., Hankinson S.E., Eliassen A.H., et al., *Association between physical activity and urinary estrogens and estrogen metabolites in premenopausal women*. J Clin Endocrinol Metab, 2012. **97**(10): p. 3724-33.
89. Bodicoat D.H., Schoemaker M.J., Jones M.E., McFadden E., Griffin J., et al., *Timing of pubertal stages and breast cancer risk: the Breakthrough Generations Study*. Breast Cancer Res, 2014. **16**(1): p. R18.
90. Kelsey J.L., Gammon M.D. and John E.M., *Reproductive factors and breast cancer*. Epidemiol Rev, 1993. **15**(1): p. 36-47.
91. Clavel-Chapelon F. and the E3N-EPIC Group, *Differential effects of reproductive factors on the risk of pre- and postmenopausal breast cancer. Results from a large cohort of French women*. Br J Cancer, 2002. **86**(5): p. 723-7.
92. Ritte R., Lukanova A., Tjonneland A., Olsen A., Overvad K., et al., *Height, age at menarche and risk of hormone receptor-positive and -negative breast cancer: a cohort study*. Int J Cancer, 2013. **132**(11): p. 2619-29.
93. Chavez-MacGregor M., Elias S.G., Onland-Moret N.C., van der Schouw Y.T., Van Gils C.H., et al., *Postmenopausal breast cancer risk and cumulative number of menstrual cycles*. Cancer Epidemiol Biomarkers Prev, 2005. **14**(4): p. 799-804.
94. Clavel-Chapelon F. and the E3N-EPIC Group, *Cumulative number of menstrual cycles and breast cancer risk: results from the E3N cohort study of French women*. Cancer Causes Control, 2002. **13**(9): p. 831-8.
95. Collaborative Group on Hormonal Factors in Breast Cancer, *Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer*. Lancet, 1997. **350**(9084): p. 1047-59.
96. Horn J., Asvold B.O., Opdahl S., Tretli S. and Vatten L.J., *Reproductive factors and the risk of breast cancer in old age: a Norwegian cohort study*. Breast Cancer Res Treat, 2013. **139**(1): p. 237-43.
97. Colditz G.A. and Rosner B., *Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study*. Am J Epidemiol, 2000. **152**(10): p. 950-64.
98. Ewertz M., Duffy S.W., Adami H.O., Kvale G., Lund E., et al., *Age at first birth, parity and risk of breast cancer: a meta-analysis of 8 studies from the Nordic countries*. Int J Cancer, 1990. **46**(4): p. 597-603.
99. Lambe M., Hsieh C.C., Tsaih S.W., Ekbom A., Trichopoulos D., et al., *Parity, age at first birth and the risk of carcinoma in situ of the breast*. Int J Cancer, 1998. **77**(3): p. 330-2.
100. Collaborative Group on Hormonal Factors in Breast Cancer, *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, 2002. **360**(9328): p. 187-95.
101. MacMahon B., Cole P., Lin T.M., Lowe C.R., Mirra A.P., et al., *Age at first birth and breast cancer risk*. Bull World Health Organ, 1970. **43**(2): p. 209-21.

102. Albrektsen G., Heuch I., Hansen S. and Kvale G., *Breast cancer risk by age at birth, time since birth and time intervals between births: exploring interaction effects*. Br J Cancer, 2005. **92**(1): p. 167-75.
103. Russo J., Moral R., Balogh G.A., Mailo D. and Russo I.H., *The protective role of pregnancy in breast cancer*. Breast Cancer Res, 2005. **7**(3): p. 131-42.
104. Medina D., *Breast cancer: the protective effect of pregnancy*. Clin Cancer Res, 2004. **10**(1 Pt 2): p. 380s-4s.
105. Husby A., Wohlfahrt J., Oyen N. and Melbye M., *Pregnancy duration and breast cancer risk*. Nat Commun, 2018. **9**(1): p. 4255.
106. Lambertini M., Santoro L., Del Mastro L., Nguyen B., Livraghi L., et al., *Reproductive behaviors and risk of developing breast cancer according to tumor subtype: A systematic review and meta-analysis of epidemiological studies*. Cancer Treat Rev, 2016. **49**: p. 65-76.
107. Ma H., Bernstein L., Pike M.C. and Ursin G., *Reproductive factors and breast cancer risk according to joint estrogen and progesterone receptor status: a meta-analysis of epidemiological studies*. Breast Cancer Res, 2006. **8**(4): p. R43.
108. Yang X.R., Chang-Claude J., Goode E.L., Couch F.J., Nevanlinna H., et al., *Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies*. J Natl Cancer Inst, 2011. **103**(3): p. 250-63.
109. Ellingjord-Dale M., Vos L., Tretli S., Hofvind S., Dos-Santos-Silva I., et al., *Parity, hormones and breast cancer subtypes - results from a large nested case-control study in a national screening program*. Breast Cancer Res, 2017. **19**(1): p. 10.
110. Dorgan J.F., Reichman M.E., Judd J.T., Brown C., Longcope C., et al., *Relationships of age and reproductive characteristics with plasma estrogens and androgens in premenopausal women*. Cancer Epidemiol Biomarkers Prev, 1995. **4**(4): p. 381-6.
111. Bernstein L., Pike M.C., Ross R.K., Judd H.L., Brown J.B., et al., *Estrogen and sex hormone-binding globulin levels in nulliparous and parous women*. J Natl Cancer Inst, 1985. **74**(4): p. 741-5.
112. Azim H.A., Jr., Santoro L., Russell-Edu W., Pentheroudakis G., Pavlidis N., et al., *Prognosis of pregnancy-associated breast cancer: a meta-analysis of 30 studies*. Cancer Treat Rev, 2012. **38**(7): p. 834-42.
113. Abbassi-Ghanavati M., Greer L.G. and Cunningham F.G., *Pregnancy and laboratory studies: a reference table for clinicians*. Obstet Gynecol, 2009. **114**(6): p. 1326-31.
114. Figueroa J.D., Pfeiffer R.M., Patel D.A., Linville L., Brinton L.A., et al., *Terminal duct lobular unit involution of the normal breast: implications for breast cancer etiology*. J Natl Cancer Inst, 2014. **106**(10).
115. Bernier M.O., Plu-Bureau G., Bossard N., Ayzac L. and Thalabard J.C., *Breastfeeding and risk of breast cancer: a metaanalysis of published studies*. Hum Reprod Update, 2000. **6**(4): p. 374-86.

116. Chowdhury R., Sinha B., Sankar M.J., Taneja S., Bhandari N., et al., *Breastfeeding and maternal health outcomes: a systematic review and meta-analysis*. Acta Paediatr, 2015. **104**(467): p. 96-113.
117. Zhou Y., Chen J., Li Q., Huang W., Lan H., et al., *Association between breastfeeding and breast cancer risk: evidence from a meta-analysis*. Breastfeed Med, 2015. **10**(3): p. 175-82.
118. Newcomb P.A., Storer B.E., Longnecker M.P., Mittendorf R., Greenberg E.R., et al., *Lactation and a reduced risk of premenopausal breast cancer*. N Engl J Med, 1994. **330**(2): p. 81-7.
119. Michels K.B., Willett W.C., Rosner B.A., Manson J.E., Hunter D.J., et al., *Prospective assessment of breastfeeding and breast cancer incidence among 89,887 women*. Lancet, 1996. **347**(8999): p. 431-6.
120. Narod S.A., *Hormone replacement therapy and the risk of breast cancer*. Nat Rev Clin Oncol, 2011. **8**(11): p. 669-76.
121. Million Women Study Collaborators, *Breast cancer and hormone-replacement therapy in the Million Women Study*. Lancet, 2003. **362**(9382): p. 419-27.
122. Bakken K., Fournier A., Lund E., Waaseth M., Dumeaux V., et al., *Menopausal hormone therapy and breast cancer risk: impact of different treatments. The European Prospective Investigation into Cancer and Nutrition*. Int J Cancer, 2011. **128**(1): p. 144-56.
123. Fournier A., Berrino F., Riboli E., Avenel V. and Clavel-Chapelon F., *Breast cancer risk in relation to different types of hormone replacement therapy in the E3N-EPIC cohort*. Int J Cancer, 2005. **114**(3): p. 448-54.
124. Saxena T., Lee E., Henderson K.D., Clarke C.A., West D., et al., *Menopausal hormone therapy and subsequent risk of specific invasive breast cancer subtypes in the California Teachers Study*. Cancer Epidemiol Biomarkers Prev, 2010. **19**(9): p. 2366-78.
125. Writing Group for the Women's Health Initiative Investigators, *Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial*. JAMA, 2002. **288**(3): p. 321-33.
126. Chlebowski R.T., Rohan T.E., Manson J.E., Aragaki A.K., Kaunitz A., et al., *Breast Cancer After Use of Estrogen Plus Progestin and Estrogen Alone: Analyses of Data From 2 Women's Health Initiative Randomized Clinical Trials*. JAMA Oncol, 2015. **1**(3): p. 296-305.
127. Munsell M.F., Sprague B.L., Berry D.A., Chisholm G. and Trentham-Dietz A., *Body mass index and breast cancer risk according to postmenopausal estrogen-progestin use and hormone receptor status*. Epidemiol Rev, 2014. **36**: p. 114-36.
128. Prentice R.L., Chlebowski R.T., Stefanick M.L., Manson J.E., Pettinger M., et al., *Estrogen plus progestin therapy and breast cancer in recently postmenopausal women*. Am J Epidemiol, 2008. **167**(10): p. 1207-16.
129. Havrilesky L.J., Moorman P.G., Lowery W.J., Gierisch J.M., Coeytaux R.R., et al., *Oral contraceptive pills as primary prevention for ovarian cancer: a systematic review and meta-analysis*. Obstet Gynecol, 2013. **122**(1): p. 139-47.

130. Gierisch J.M., Coeytaux R.R., Urrutia R.P., Havrilesky L.J., Moorman P.G., et al., *Oral contraceptive use and risk of breast, cervical, colorectal, and endometrial cancers: a systematic review*. Cancer Epidemiol Biomarkers Prev, 2013. **22**(11): p. 1931-43.
131. Kahlenborn C., Modugno F., Potter D.M. and Severs W.B., *Oral contraceptive use as a risk factor for premenopausal breast cancer: a meta-analysis*. Mayo Clin Proc, 2006. **81**(10): p. 1290-302.
132. Collaborative Group on Hormonal Factors in Breast Cancer, *Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies*. Lancet, 1996. **347**(9017): p. 1713-27.
133. Morch L.S., Skovlund C.W., Hannaford P.C., Iversen L., Fielding S., et al., *Contemporary Hormonal Contraception and the Risk of Breast Cancer*. N Engl J Med, 2017. **377**(23): p. 2228-2239.
134. Charlton B.M., Rich-Edwards J.W., Colditz G.A., Missmer S.A., Rosner B.A., et al., *Oral contraceptive use and mortality after 36 years of follow-up in the Nurses' Health Study: prospective cohort study*. BMJ, 2014. **349**: p. g6356.
135. Marchbanks P.A., McDonald J.A., Wilson H.G., Folger S.G., Mandel M.G., et al., *Oral contraceptives and the risk of breast cancer*. N Engl J Med, 2002. **346**(26): p. 2025-32.
136. Beaber E.F., Buist D.S., Barlow W.E., Malone K.E., Reed S.D., et al., *Recent oral contraceptive use by formulation and breast cancer risk among women 20 to 49 years of age*. Cancer Res, 2014. **74**(15): p. 4078-89.
137. Clemons M. and Goss P., *Estrogen and the risk of breast cancer*. N Engl J Med, 2001. **344**(4): p. 276-85.
138. Zhu H., Lei X., Feng J. and Wang Y., *Oral contraceptive use and risk of breast cancer: a meta-analysis of prospective cohort studies*. Eur J Contracept Reprod Health Care, 2012. **17**(6): p. 402-14.
139. Hunter D.J., *Oral Contraceptives and the Small Increased Risk of Breast Cancer*. N Engl J Med, 2017. **377**(23): p. 2276-2277.
140. Moorman P.G., Havrilesky L.J., Gierisch J.M., Coeytaux R.R., Lowery W.J., et al., *Oral contraceptives and risk of ovarian cancer and breast cancer among high-risk women: a systematic review and meta-analysis*. J Clin Oncol, 2013. **31**(33): p. 4188-98.
141. Howell A., Anderson A.S., Clarke R.B., Duffy S.W., Evans D.G., et al., *Risk determination and prevention of breast cancer*. Breast Cancer Res, 2014. **16**(5): p. 446.
142. Gail M.H., Brinton L.A., Byar D.P., Corle D.K., Green S.B., et al., *Projecting individualized probabilities of developing breast cancer for white females who are being examined annually*. J Natl Cancer Inst, 1989. **81**(24): p. 1879-86.
143. Costantino J.P., Gail M.H., Pee D., Anderson S., Redmond C.K., et al., *Validation studies for models projecting the risk of invasive and total breast cancer incidence*. J Natl Cancer Inst, 1999. **91**(18): p. 1541-8.

144. Cuzick J. *IBIS Breast Cancer Risk Evaluation Tool*. [Accessed November 2018]; Available from: <http://www.ems-trials.org/riskevaluator/>.
145. Tyrer J., Duffy S.W. and Cuzick J., *A breast cancer prediction model incorporating familial and personal risk factors*. Stat Med, 2004. **23**(7): p. 1111-30.
146. Antoniou A.C., Cunningham A.P., Peto J., Evans D.G., Lalloo F., et al., *The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions*. Br J Cancer, 2008. **98**(8): p. 1457-66.
147. Lee A., Mavaddat N., Wilcox A.N., Cunningham A.P., Carver T., et al., *BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors*. Genet Med, 2019.
148. Amir E., Evans D.G., Shenton A., Lalloo F., Moran A., et al., *Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme*. J Med Genet, 2003. **40**(11): p. 807-14.
149. McCormack V.A. and dos Santos Silva I., *Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis*. Cancer Epidemiol Biomarkers Prev, 2006. **15**(6): p. 1159-69.
150. Brentnall A.R., Harkness E.F., Astley S.M., Donnelly L.S., Stavrinou P., et al., *Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort*. Breast Cancer Res, 2015. **17**(1): p. 147.
151. Evans D.G., Warwick J., Astley S.M., Stavrinou P., Sahin S., et al., *Assessing individual breast cancer risk within the U.K. National Health Service Breast Screening Program: a new paradigm for cancer prevention*. Cancer Prev Res (Phila), 2012. **5**(7): p. 943-51.
152. Santen R.J., Boyd N.F., Chlebowski R.T., Cummings S., Cuzick J., et al., *Critical assessment of new risk factors for breast cancer: considerations for development of an improved risk prediction model*. Endocr Relat Cancer, 2007. **14**(2): p. 169-87.
153. Figueroa J.D. and Brinton L.A., *Unraveling genes, hormones, and breast cancer*. J Natl Cancer Inst, 2012. **104**(9): p. 641-2.
154. Pike M.C., Krailo M.D., Henderson B.E., Casagrande J.T. and Hoel D.G., *'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer*. Nature, 1983. **303**(5920): p. 767-70.
155. Evans D.G. and Howell A., *Breast cancer risk-assessment models*. Breast Cancer Res, 2007. **9**(5): p. 213.
156. Cuzick J., Sestak I., Bonanni B., Costantino J.P., Cummings S., et al., *Selective oestrogen receptor modulators in prevention of breast cancer: an updated meta-analysis of individual participant data*. Lancet, 2013. **381**(9880): p. 1827-34.
157. Cuzick J., Sestak I., Cawthorn S., Hamed H., Holli K., et al., *Tamoxifen for prevention of breast cancer: extended long-term follow-up of the IBIS-I breast cancer prevention trial*. Lancet Oncol, 2015. **16**(1): p. 67-75.

158. Fisher B., Costantino J.P., Wickerham D.L., Cecchini R.S., Cronin W.M., et al., *Tamoxifen for the prevention of breast cancer: current status of the National Surgical Adjuvant Breast and Bowel Project P-1 study*. J Natl Cancer Inst, 2005. **97**(22): p. 1652-62.
159. Goss P.E., Ingle J.N., Ales-Martinez J.E., Cheung A.M., Chlebowski R.T., et al., *Exemestane for breast-cancer prevention in postmenopausal women*. N Engl J Med, 2011. **364**(25): p. 2381-91.
160. Cuzick J., Sestak I., Forbes J.F., Dowsett M., Knox J., et al., *Anastrozole for prevention of breast cancer in high-risk postmenopausal women (IBIS-II): an international, double-blind, randomised placebo-controlled trial*. Lancet, 2014. **383**(9922): p. 1041-8.
161. National Collaborating Centre for Cancer, *National Institute for Health and Clinical Excellence: Guidance*, in *Familial Breast Cancer: Classification and Care of People at Risk of Familial Breast Cancer and Management of Breast Cancer and Related Risks in People with a Family History of Breast Cancer*. 2013, National Collaborating Centre for Cancer (UK): Cardiff (UK).
162. Uray I.P. and Brown P.H., *Chemoprevention of hormone receptor-negative breast cancer: new approaches needed*. Recent Results Cancer Res, 2011. **188**: p. 147-62.
163. Litzenburger B.C. and Brown P.H., *Advances in Preventive Therapy for Estrogen-Receptor-Negative Breast Cancer*. Curr Breast Cancer Rep, 2014. **6**: p. 96-109.
164. Cancer Research UK. *Breast Cancer Risk Factors - Lifetime Risk of Breast Cancer*. [Accessed November 2018]; Available from: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/risk-factors>.
165. Harvie M., Howell A. and Evans D.G., *Can diet and lifestyle prevent breast cancer: what is the evidence?* Am Soc Clin Oncol Educ Book, 2015: p. e66-73.
166. Breast Cancer Now. *How can I reduce my risk of breast cancer?* [Accessed November 2018]; Available from: <https://breastcancernow.org/about-breast-cancer/want-to-know-about-breast-cancer/what-can-cause-breast-cancer/reduce-risk-of-breast-cancer>.
167. Kunzmann A.T., Coleman H.G., Huang W.Y. and Berndt S.I., *The association of lifetime alcohol use with mortality and cancer risk in older adults: A cohort study*. PLoS Med, 2018. **15**(6): p. e1002585.
168. Eliassen A.H., Colditz G.A., Rosner B., Willett W.C. and Hankinson S.E., *Adult weight change and risk of postmenopausal breast cancer*. Jama, 2006. **296**(2): p. 193-201.
169. Harvie M., Howell A., Vierkant R.A., Kumar N., Cerhan J.R., et al., *Association of gain and loss of weight before and after menopause with risk of postmenopausal breast cancer in the Iowa women's health study*. Cancer Epidemiol Biomarkers Prev, 2005. **14**(3): p. 656-61.
170. Aune D., Chan D.S., Greenwood D.C., Vieira A.R., Rosenblatt D.A., et al., *Dietary fiber and breast cancer risk: a systematic review and meta-analysis of prospective studies*. Ann Oncol, 2012. **23**(6): p. 1394-402.
171. Alexander D.D., Morimoto L.M., Mink P.J. and Cushing C.A., *A review and meta-analysis of red and processed meat consumption and breast cancer*. Nutr Res Rev, 2010. **23**(2): p. 349-65.

172. Toledo E., Salas-Salvado J., Donat-Vargas C., Buil-Cosiales P., Estruch R., et al., *Mediterranean Diet and Invasive Breast Cancer Risk Among Women at High Cardiovascular Risk in the PREDIMED Trial: A Randomized Clinical Trial*. JAMA Intern Med, 2015. **175**(11): p. 1752-1760.
173. van den Brandt P.A. and Schulpen M., *Mediterranean diet adherence and risk of postmenopausal breast cancer: results of a cohort study and meta-analysis*. Int J Cancer, 2017. **140**(10): p. 2220-2231.
174. Albuquerque R.C., Baltar V.T. and Marchioni D.M., *Breast cancer and dietary patterns: a systematic review*. Nutr Rev, 2014. **72**(1): p. 1-17.
175. Holliday R., *Epigenetics: a historical overview*. Epigenetics, 2006. **1**(2): p. 76-80.
176. Bird A., *Perceptions of epigenetics*. Nature, 2007. **447**(7143): p. 396-8.
177. Goldberg A.D., Allis C.D. and Bernstein E., *Epigenetics: a landscape takes shape*. Cell, 2007. **128**(4): p. 635-8.
178. Jones P.A. and Baylin S.B., *The epigenomics of cancer*. Cell, 2007. **128**(4): p. 683-92.
179. Esteller M., *Epigenetics in cancer*. N Engl J Med, 2008. **358**(11): p. 1148-59.
180. Zhang T., Cooper S. and Brockdorff N., *The interplay of histone modifications - writers that read*. EMBO Rep, 2015. **16**(11): p. 1467-81.
181. Bestor T.H., *The DNA methyltransferases of mammals*. Hum Mol Genet, 2000. **9**(16): p. 2395-402.
182. *Illumina - Introduction to Methylation Array Analysis*. [Accessed December 2018]; Available from: <https://emea.illumina.com/techniques/microarrays/methylation-arrays.html>.
183. Delaney C., Garg S.K. and Yung R., *Analysis of DNA Methylation by Pyrosequencing*. Methods Mol Biol, 2015. **1343**: p. 249-64.
184. Li N., Ye M., Li Y., Yan Z., Butcher L.M., et al., *Whole genome DNA methylation analysis based on high throughput sequencing technology*. Methods, 2010. **52**(3): p. 203-12.
185. Karimi M., Johansson S., Stach D., Corcoran M., Grander D., et al., *LUMA (LUMinometric Methylation Assay)--a high throughput method to the analysis of genomic DNA methylation*. Exp Cell Res, 2006. **312**(11): p. 1989-95.
186. Terry M.B., Ferris J.S., Pilsner R., Flom J.D., Tehranifar P., et al., *Genomic DNA methylation among women in a multiethnic New York City birth cohort*. Cancer Epidemiol Biomarkers Prev, 2008. **17**(9): p. 2306-10.
187. Widschwendter M. and Jones P.A., *DNA methylation and breast carcinogenesis*. Oncogene, 2002. **21**(35): p. 5462-82.
188. Jovanovic J., Ronneberg J.A., Tost J. and Kristensen V., *The epigenetics of breast cancer*. Mol Oncol, 2010. **4**(3): p. 242-54.
189. Esteller M., Corn P.G., Baylin S.B. and Herman J.G., *A gene hypermethylation profile of human cancer*. Cancer Res, 2001. **61**(8): p. 3225-9.
190. Stefansson O.A., Jonasson J.G., Olafsdottir K., Hilmarisdottir H., Olafsdottir G., et al., *CpG island hypermethylation of BRCA1 and loss of pRb as co-occurring events in basal/triple-negative breast cancer*. Epigenetics, 2011. **6**(5): p. 638-49.

191. Holm K., Hegardt C., Staaf J., Vallon-Christersson J., Jonsson G., et al., *Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns*. Breast Cancer Res, 2010. **12**(3): p. R36.
192. Li L., Lee K.M., Han W., Choi J.Y., Lee J.Y., et al., *Estrogen and progesterone receptor status affect genome-wide DNA methylation profile in breast cancer*. Hum Mol Genet, 2010. **19**(21): p. 4273-7.
193. Fackler M.J., Umbricht C.B., Williams D., Argani P., Cruz L.A., et al., *Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence*. Cancer Res, 2011. **71**(19): p. 6195-207.
194. Ottaviano Y.L., Issa J.P., Parl F.F., Smith H.S., Baylin S.B., et al., *Methylation of the estrogen receptor gene CpG island marks loss of estrogen receptor expression in human breast cancer cells*. Cancer Res, 1994. **54**(10): p. 2552-5.
195. Brennan K. and Flanagan J.M., *Is there a link between genome-wide hypomethylation in blood and cancer risk?* Cancer Prev Res (Phila), 2012. **5**(12): p. 1345-57.
196. Choi J.Y., James S.R., Link P.A., McCann S.E., Hong C.C., et al., *Association between global DNA hypomethylation in leukocytes and risk of breast cancer*. Carcinogenesis, 2009. **30**(11): p. 1889-97.
197. van Veldhoven K., Polidoro S., Baglietto L., Severi G., Sacerdote C., et al., *Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis*. Clin Epigenetics, 2015. **7**: p. 67.
198. Shenker N.S., Flower K.J., Wilhelm-Benartzi C.S., Dai W., Bell E., et al., *Transcriptional implications of intragenic DNA methylation in the oestrogen receptor alpha gene in breast cancer cells and tissues*. BMC Cancer, 2015. **15**: p. 337.
199. Anjum S., Fourkala E.O., Zikan M., Wong A., Gentry-Maharaj A., et al., *A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival*. Genome Med, 2014. **6**(6): p. 47.
200. Iwamoto T., Yamamoto N., Taguchi T., Tamaki Y. and Noguchi S., *BRCA1 promoter methylation in peripheral blood cells is associated with increased risk of breast cancer with BRCA1 promoter methylation*. Breast Cancer Res Treat, 2011. **129**(1): p. 69-77.
201. Brennan K., Garcia-Closas M., Orr N., Fletcher O., Jones M., et al., *Intragenic ATM methylation in peripheral blood DNA as a biomarker of breast cancer risk*. Cancer Res, 2012. **72**(9): p. 2304-13.
202. Flanagan J.M., Munoz-Alegre M., Henderson S., Tang T., Sun P., et al., *Gene-body hypermethylation of ATM in peripheral blood DNA of bilateral breast cancer patients*. Hum Mol Genet, 2009. **18**(7): p. 1332-42.
203. Shenker N.S., Polidoro S., van Veldhoven K., Sacerdote C., Ricceri F., et al., *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. **22**(5): p. 843-51.

204. Severi G., Southey M.C., English D.R., Jung C.H., Lonie A., et al., *Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer*. Breast Cancer Res Treat, 2014. **148**(3): p. 665-73.
205. Tang Q., Cheng J., Cao X., Surowy H. and Burwinkel B., *Blood-based DNA methylation as biomarker for breast cancer: a systematic review*. Clin Epigenetics, 2016. **8**: p. 115.
206. Ambatipudi S., Horvath S., Perrier F., Cuenin C., Hernandez-Vargas H., et al., *DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility*. Eur J Cancer, 2017. **75**: p. 299-307.
207. Flanagan J.M., Brook M.N., Orr N., Tomczyk K., Coulson P., et al., *Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study*. Cancer Epidemiol Biomarkers Prev, 2015. **24**(1): p. 221-9.
208. Johansson A. and Flanagan J.M., *Epigenome-wide association studies for breast cancer risk and risk factors*. Trends Cancer Res, 2017. **12**: p. 19-28.
209. Vineis P., Chatziioannou A., Cunliffe V.T., Flanagan J.M., Hanson M., et al., *Epigenetic memory in response to environmental stressors*. Faseb j, 2017. **31**(6): p. 2241-2251.
210. Gao X., Jia M., Zhang Y., Breitling L.P. and Brenner H., *DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies*. Clin Epigenetics, 2015. **7**: p. 113.
211. Zhang Y., Schottker B., Ordonez-Mena J., Holleccek B., Yang R., et al., *F2RL3 methylation, lung cancer incidence and mortality*. Int J Cancer, 2015. **137**(7): p. 1739-48.
212. Dick K.J., Nelson C.P., Tsaprouni L., Sandling J.K., Aissi D., et al., *DNA methylation and body-mass index: a genome-wide analysis*. Lancet, 2014. **383**(9933): p. 1990-8.
213. Wahl S., Drong A., Lehne B., Loh M., Scott W.R., et al., *Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity*. Nature, 2017. **541**(7635): p. 81-86.
214. Hannum G., Guinney J., Zhao L., Zhang L., Hughes G., et al., *Genome-wide methylation profiles reveal quantitative views of human aging rates*. Mol Cell, 2013. **49**(2): p. 359-67.
215. Horvath S., *DNA methylation age of human tissues and cell types*. Genome Biol, 2013. **14**(10): p. R115.
216. Weidner C.I., Lin Q., Koch C.M., Eisele L., Beier F., et al., *Aging of blood can be tracked by DNA methylation changes at just three CpG sites*. Genome Biol, 2014. **15**(2): p. R24.
217. Levine M.E., Lu A.T., Quach A., Chen B.H., Assimes T.L., et al., *An epigenetic biomarker of aging for lifespan and healthspan*. Aging (Albany NY), 2018. **10**(4): p. 573-591.
218. Perna L., Zhang Y., Mons U., Holleccek B., Saum K.U., et al., *Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort*. Clin Epigenetics, 2016. **8**: p. 64.
219. Breitling L.P., Yang R., Korn B., Burwinkel B. and Brenner H., *Tobacco-smoking-related differential DNA methylation: 27K discovery and replication*. Am J Hum Genet, 2011. **88**(4): p. 450-7.

220. Fasanelli F., Baglietto L., Ponzi E., Guida F., Campanella G., et al., *Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts*. Nat Commun, 2015. **6**: p. 10192.
221. Guida F., Sandanger T.M., Castagne R., Campanella G., Polidoro S., et al., *Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation*. Hum Mol Genet, 2015. **24**(8): p. 2349-59.
222. Philibert R., Hollenbeck N., Andersen E., Osborn T., Gerrard M., et al., *A quantitative epigenetic approach for the assessment of cigarette consumption*. Front Psychol, 2015. **6**: p. 656.
223. Shenker N.S., Ueland P.M., Polidoro S., van Veldhoven K., Ricceri F., et al., *DNA methylation as a long-term biomarker of exposure to tobacco smoke*. Epidemiology, 2013. **24**(5): p. 712-6.
224. Zhang R., Miao Q., Wang C., Zhao R., Li W., et al., *Genome-wide DNA methylation analysis in alcohol dependence*. Addict Biol, 2013. **18**(2): p. 392-403.
225. Zhao R., Zhang R., Li W., Liao Y., Tang J., et al., *Genome-wide DNA methylation patterns in discordant sib pairs with alcohol dependence*. Asia Pac Psychiatry, 2013. **5**(1): p. 39-50.
226. Weng J.T., Wu L.S., Lee C.S., Hsu P.W. and Cheng A.T., *Integrative epigenetic profiling analysis identifies DNA methylation changes associated with chronic alcohol consumption*. Comput Biol Med, 2015. **64**: p. 299-306.
227. Bruckmann C., Islam S.A., MacIsaac J.L., Morin A.M., Karle K.N., et al., *DNA methylation signatures of chronic alcohol dependence in purified CD3(+) T-cells of patients undergoing alcohol treatment*. Sci Rep, 2017. **7**(1): p. 6605.
228. Bruckmann C., Di Santo A., Karle K.N., Batra A. and Nieratschker V., *Validation of differential GAD65 DNA methylation in alcohol dependence and its potential function as a biomarker for disease severity and therapy outcome*. Epigenetics, 2016. **11**(6): p. 456-63.
229. Aslibekyan S., Demerath E.W., Mendelson M., Zhi D., Guan W., et al., *Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference*. Obesity (Silver Spring), 2015. **23**(7): p. 1493-501.
230. Demerath E.W., Guan W., Grove M.L., Aslibekyan S., Mendelson M., et al., *Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci*. Hum Mol Genet, 2015. **24**(15): p. 4464-79.
231. Sayols-Baixeras S., Subirana I., Fernandez-Sanles A., Senti M., Lluís-Ganella C., et al., *DNA methylation and obesity traits: An epigenome-wide association study. The REGICOR study*. Epigenetics, 2017. **12**(10): p. 909-916.
232. Wilson L.E., Harlid S., Xu Z., Sandler D.P. and Taylor J.A., *An epigenome-wide study of body mass index and DNA methylation in blood using participants from the Sister Study cohort*. Int J Obes (Lond), 2017. **41**(1): p. 194-199.
233. Key T., Appleby P., Barnes I. and Reeves G., *Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies*. J Natl Cancer Inst, 2002. **94**(8): p. 606-16.
234. Nelson L.R. and Bulun S.E., *Estrogen production and action*. J Am Acad Dermatol, 2001. **45**(3 Suppl): p. S116-24.

235. Fox M., Berzuini C. and Knapp L.A., *Cumulative estrogen exposure, number of menstrual cycles, and Alzheimer's risk in a cohort of British women*. Psychoneuroendocrinology, 2013. **38**(12): p. 2973-82.
236. Shaw N.D., Histed S.N., Srouji S.S., Yang J., Lee H., et al., *Estrogen negative feedback on gonadotropin secretion: evidence for a direct pituitary effect in women*. J Clin Endocrinol Metab, 2010. **95**(4): p. 1955-61.
237. Dowsett M. and Folkard E., *Reduced progesterone levels explain the reduced risk of breast cancer in obese premenopausal women: a new hypothesis*. Breast Cancer Res Treat, 2015. **149**(1): p. 1-4.
238. Dorgan J.F., Stanczyk F.Z., Kahle L.L. and Brinton L.A., *Prospective case-control study of premenopausal serum estradiol and testosterone levels and breast cancer risk*. Breast Cancer Res, 2010. **12**(6): p. R98.
239. Kaaks R., Tikk K., Sookthai D., Schock H., Johnson T., et al., *Premenopausal serum sex hormone levels in relation to breast cancer risk, overall and by hormone receptor status - results from the EPIC cohort*. Int J Cancer, 2014. **134**(8): p. 1947-57.
240. Eliassen A.H., Missmer S.A., Tworoger S.S., Spiegelman D., Barbieri R.L., et al., *Endogenous steroid hormone concentrations and risk of breast cancer among premenopausal women*. J Natl Cancer Inst, 2006. **98**(19): p. 1406-15.
241. Fortner R.T., Eliassen A.H., Spiegelman D., Willett W.C., Barbieri R.L., et al., *Premenopausal endogenous steroid hormones and breast cancer risk: results from the Nurses' Health Study II*. Breast Cancer Res, 2013. **15**(2): p. R19.
242. Filardo E.J. and Thomas P., *Minireview: G protein-coupled estrogen receptor-1, GPER-1: its mechanism of action and role in female reproductive cancer, renal and vascular physiology*. Endocrinology, 2012. **153**(7): p. 2953-62.
243. Nilsson S., Makela S., Treuter E., Tujague M., Thomsen J., et al., *Mechanisms of estrogen action*. Physiol Rev, 2001. **81**(4): p. 1535-65.
244. Liang J. and Shang Y., *Estrogen and cancer*. Annu Rev Physiol, 2013. **75**: p. 225-40.
245. Santen R.J., Yue W. and Wang J.P., *Estrogen metabolites and breast cancer*. Steroids, 2015. **99**(Pt A): p. 61-6.
246. Huang B., Omoto Y., Iwase H., Yamashita H., Toyama T., et al., *Differential expression of estrogen receptor alpha, beta1, and beta2 in lobular and ductal breast cancer*. Proc Natl Acad Sci U S A, 2014. **111**(5): p. 1933-8.
247. Hurtado A., Holmes K.A., Ross-Innes C.S., Schmidt D. and Carroll J.S., *FOXAI is a key determinant of estrogen receptor function and endocrine response*. Nat Genet, 2011. **43**(1): p. 27-33.
248. Clarke R.B., Howell A., Potten C.S. and Anderson E., *Dissociation between steroid receptor expression and cell proliferation in the human breast*. Cancer Res, 1997. **57**(22): p. 4987-91.
249. Savage K.I., Matchett K.B., Barros E.M., Cooper K.M., Irwin G.W., et al., *BRCA1 deficiency exacerbates estrogen-induced DNA damage and genomic instability*. Cancer Res, 2014. **74**(10): p. 2773-2784.

250. Cavalieri E., Chakravarti D., Guttenplan J., Hart E., Ingle J., et al., *Catechol estrogen quinones as initiators of breast and other human cancers: implications for biomarkers of susceptibility and cancer prevention*. Biochim Biophys Acta, 2006. **1766**(1): p. 63-78.
251. Cavalieri E.L., Stack D.E., Devanesan P.D., Todorovic R., Dwivedy I., et al., *Molecular origin of cancer: catechol estrogen-3,4-quinones as endogenous tumor initiators*. Proc Natl Acad Sci U S A, 1997. **94**(20): p. 10937-42.
252. Mailander P.C., Meza J.L., Higginbotham S. and Chakravarti D., *Induction of A.T to G.C mutations by erroneous repair of depurinated DNA following estrogen treatment of the mammary gland of ACI rats*. J Steroid Biochem Mol Biol, 2006. **101**(4-5): p. 204-15.
253. Yager J.D., *Mechanisms of estrogen carcinogenesis: The role of E2/E1-quinone metabolites suggests new approaches to preventive intervention--A review*. Steroids, 2015. **99**(Pt A): p. 56-60.
254. Rogan E.G., Badawi A.F., Devanesan P.D., Meza J.L., Edney J.A., et al., *Relative imbalances in estrogen metabolism and conjugation in breast tissue of women with carcinoma: potential biomarkers of susceptibility to cancer*. Carcinogenesis, 2003. **24**(4): p. 697-702.
255. Cavalieri E. and Rogan E., *The molecular etiology and prevention of estrogen-initiated cancers: Ockham's Razor: Pluralitas non est ponenda sine necessitate. Plurality should not be posited without necessity*. Mol Aspects Med, 2014. **36**: p. 1-55.
256. Sancar A., Lindsey-Boltz L.A., Unsal-Kacmaz K. and Linn S., *Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints*. Annu Rev Biochem, 2004. **73**: p. 39-85.
257. Cuozzo C., Porcellini A., Angrisano T., Morano A., Lee B., et al., *DNA damage, homology-directed repair, and DNA methylation*. PLoS Genet, 2007. **3**(7): p. e110.
258. O'Hagan H.M., Mohammad H.P. and Baylin S.B., *Double strand breaks can initiate gene silencing and SIRT1-dependent onset of DNA methylation in an exogenous promoter CpG island*. PLoS Genet, 2008. **4**(8): p. e1000155.
259. O'Hagan H.M., Wang W., Sen S., Destefano Shields C., Lee S.S., et al., *Oxidative damage targets complexes containing DNA methyltransferases, SIRT1, and polycomb members to promoter CpG Islands*. Cancer Cell, 2011. **20**(5): p. 606-19.
260. Jin B. and Robertson K.D., *DNA methyltransferases, DNA damage repair, and cancer*. Adv Exp Med Biol, 2013. **754**: p. 3-29.
261. Flanagan J.M., Wilson A., Koo C., Masrour N., Gallon J., et al., *Platinum-Based Chemotherapy Induces Methylation Changes in Blood DNA Associated with Overall Survival in Patients with Ovarian Cancer*. Clin Cancer Res, 2017. **23**(9): p. 2213-2222.
262. Rodriguez B.A., Weng Y.I., Liu T.M., Zuo T., Hsu P.Y., et al., *Estrogen-mediated epigenetic repression of the imprinted gene cyclin-dependent kinase inhibitor 1C in breast cancer cells*. Carcinogenesis, 2011. **32**(6): p. 812-21.
263. Ramaswamy B., Majumder S., Roy S., Ghoshal K., Kutay H., et al., *Estrogen-mediated suppression of the gene encoding protein tyrosine phosphatase PTPRO in human breast cancer: mechanism and role in tamoxifen sensitivity*. Mol Endocrinol, 2009. **23**(2): p. 176-87.

264. Fernandez S.V. and Russo J., *Estrogen and xenoestrogens in breast cancer*. Toxicol Pathol, 2010. **38**(1): p. 110-22.
265. Hsu P.Y., Hsu H.K., Singer G.A., Yan P.S., Rodriguez B.A., et al., *Estrogen-mediated epigenetic repression of large chromosomal regions through DNA looping*. Genome Res, 2010. **20**(6): p. 733-44.
266. Jadhav R.R., Ye Z., Huang R.L., Liu J., Hsu P.Y., et al., *Genome-wide DNA methylation analysis reveals estrogen-mediated epigenetic repression of metallothionein-1 gene cluster in breast cancer*. Clin Epigenetics, 2015. **7**: p. 13.
267. Kovalchuk O., Tryndyak V.P., Montgomery B., Boyko A., Kutanzi K., et al., *Estrogen-induced rat breast carcinogenesis is characterized by alterations in DNA methylation, histone modifications and aberrant microRNA expression*. Cell Cycle, 2007. **6**(16): p. 2010-8.
268. Kutanzi K.R., Koturbash I. and Kovalchuk O., *Reversibility of pre-malignant estrogen-induced epigenetic changes*. Cell Cycle, 2010. **9**(15): p. 3078-84.
269. Russo J., Santucci-Pereira J., de Cicco R.L., Sheriff F., Russo P.A., et al., *Pregnancy-induced chromatin remodeling in the breast of postmenopausal women*. Int J Cancer, 2012. **131**(5): p. 1059-70.
270. Izadi P., Noruzinia M., Fereidooni F., Mostakhdemine Hosseini Z. and Kamali F., *Epigenetic marks in estrogen receptor alpha CpG island correlate with some reproductive risk factors in breast cancer*. Mol Biol Rep, 2014. **41**(11): p. 7607-12.
271. Ghosh S., Gu F., Wang C.M., Lin C.L., Liu J., et al., *Genome-wide DNA methylation profiling reveals parity-associated hypermethylation of FOXA1*. Breast Cancer Res Treat, 2014. **147**(3): p. 653-9.
272. Tao M.H., Marian C., Shields P.G., Potischman N., Nie J., et al., *Exposures in early life: associations with DNA promoter methylation in breast tumors*. J Dev Orig Health Dis, 2013. **4**(2): p. 182-90.
273. Demetriou C.A., Chen J., Polidoro S., van Veldhoven K., Cuenin C., et al., *Methylome analysis and epigenetic changes associated with menarcheal age*. PLoS One, 2013. **8**(11): p. e79391.
274. Lu S., Niu Z., Chen Y., Tu Q., Zhang Y., et al., *Repetitive Element DNA Methylation is Associated with Menopausal Age*. Aging Dis, 2018. **9**(3): p. 435-443.
275. El-Maarri O., Walier M., Behne F., van Uum J., Singer H., et al., *Methylation at global LINE-1 repeats in human blood are affected by gender but not by age or natural hormone cycles*. PLoS One, 2011. **6**(1): p. e16252.
276. Levine M.E., Lu A.T., Chen B.H., Hernandez D.G., Singleton A.B., et al., *Menopause accelerates biological aging*. Proc Natl Acad Sci U S A, 2016. **113**(33): p. 9327-32.
277. Bahl A., Pollanen E., Ismail K., Sipila S., Mikkola T.M., et al., *Hormone Replacement Therapy Associated White Blood Cell DNA Methylation and Gene Expression are Associated With Within-Pair Differences of Body Adiposity and Bone Mass*. Twin Res Hum Genet, 2015. **18**(6): p. 647-61.
278. Campesi I., Sanna M., Zinellu A., Carru C., Rubattu L., et al., *Oral contraceptives modify DNA methylation and monocyte-derived macrophage function*. Biol Sex Differ, 2012. **3**: p. 4.

279. Guthikonda K., Zhang H., Nolan V.G., Soto-Ramirez N., Ziyab A.H., et al., *Oral contraceptives modify the effect of GATA3 polymorphisms on the risk of asthma at the age of 18 years via DNA methylation*. Clin Epigenetics, 2014. **6**(1): p. 17.
280. Hair B.Y., Troester M.A., Edmiston S.N., Parrish E.A., Robinson W.R., et al., *Body mass index is associated with gene methylation in estrogen receptor-positive breast tumors*. Cancer Epidemiol Biomarkers Prev, 2015. **24**(3): p. 580-6.
281. Lewis J.R., McNab T.J., Liew L.J., Tan J., Hudson P., et al., *DNA methylation within the I.4 promoter region correlates with CYP119A1 gene expression in human ex vivo mature omental and subcutaneous adipocytes*. BMC Med Genet, 2013. **14**: p. 87.
282. Tao M.H., Marian C., Shields P.G., Nie J., McCann S.E., et al., *Alcohol consumption in relation to aberrant DNA methylation in breast tumors*. Alcohol, 2011. **45**(7): p. 689-99.
283. Llanos A.A., Dumitrescu R.G., Brasky T.M., Liu Z., Mason J.B., et al., *Relationships among folate, alcohol consumption, gene variants in one-carbon metabolism and p16INK4a methylation and expression in healthy breast tissues*. Carcinogenesis, 2015. **36**(1): p. 60-7.
284. White A.J., Chen J., Teitelbaum S.L., McCullough L.E., Xu X., et al., *Sources of polycyclic aromatic hydrocarbons are associated with gene-specific promoter methylation in women with breast cancer*. Environ Res, 2016. **145**: p. 93-100.
285. White A.J., Sandler D.P., Bolick S.C., Xu Z., Taylor J.A., et al., *Recreational and household physical activity at different time points and DNA global methylation*. Eur J Cancer, 2013. **49**(9): p. 2199-206.
286. Zhang F.F., Cardarelli R., Carroll J., Zhang S., Fulda K.G., et al., *Physical activity and global genomic DNA methylation in a cancer-free population*. Epigenetics, 2011. **6**(3): p. 293-9.
287. Coyle Y.M., Xie X.J., Lewis C.M., Bu D., Milchgrub S., et al., *Role of physical activity in modulating breast cancer risk as defined by APC and RASSF1A promoter hypermethylation in nonmalignant breast tissue*. Cancer Epidemiol Biomarkers Prev, 2007. **16**(2): p. 192-6.
288. Xu Z., Bolick S.C., DeRoo L.A., Weinberg C.R., Sandler D.P., et al., *Epigenome-wide association study of breast cancer using prospectively collected sister study samples*. J Natl Cancer Inst, 2013. **105**(10): p. 694-700.
289. Joo J.E., Dowty J.G., Milne R.L., Wong E.M., Dugue P.A., et al., *Heritable DNA methylation marks associated with susceptibility to breast cancer*. Nat Commun, 2018. **9**(1): p. 867.
290. Deroo L.A., Bolick S.C., Xu Z., Umbach D.M., Shore D., et al., *Global DNA methylation and one-carbon metabolism gene polymorphisms and the risk of breast cancer in the Sister Study*. Carcinogenesis, 2014. **35**(2): p. 333-8.
291. Xu X., Gammon M.D., Hernandez-Vargas H., Herceg Z., Wetmur J.G., et al., *DNA methylation in peripheral blood measured by LUMA is associated with breast cancer in a population-based study*. Faseb j, 2012. **26**(6): p. 2657-66.
292. Christensen B.C., Houseman E.A., Marsit C.J., Zheng S., Wrensch M.R., et al., *Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context*. PLoS Genet, 2009. **5**(8): p. e1000602.

293. Rakyan V.K., Down T.A., Maslau S., Andrew T., Yang T.P., et al., *Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains*. Genome Res, 2010. **20**(4): p. 434-9.
294. Gao X., Zhang Y., Breitling L.P. and Brenner H., *Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration*. Oncotarget, 2016. **7**(30): p. 46878-46889.
295. Nevalainen T., Kananen L., Marttila S., Jylhava J., Mononen N., et al., *Obesity accelerates epigenetic aging in middle-aged but not in elderly individuals*. Clin Epigenetics, 2017. **9**: p. 20.
296. Quach A., Levine M.E., Tanaka T., Lu A.T., Chen B.H., et al., *Epigenetic clock analysis of diet, exercise, education, and lifestyle factors*. Aging (Albany NY), 2017. **9**(2): p. 419-37.
297. Philibert R.A., Plume J.M., Gibbons F.X., Brody G.H. and Beach S.R., *The impact of recent alcohol use on genome wide DNA methylation signatures*. Front Genet, 2012. **3**: p. 54.
298. Clark S.L., Aberg K.A., Nerella S., Kumar G., McClay J.L., et al., *Combined Whole Methylome and Genomewide Association Study Implicates CNTN4 in Alcohol Use*. Alcohol Clin Exp Res, 2015. **39**(8): p. 1396-405.
299. Liu C., Marioni R.E., Hedman A.K., Pfeiffer L., Tsai P.C., et al., *A DNA methylation biomarker of alcohol consumption*. Mol Psychiatry, 2016.
300. Smith C.A., McCleary C.A., Murdock G.A., Wilshire T.W., Buckwalter D.K., et al., *Lifelong estrogen exposure and cognitive performance in elderly women*. Brain Cogn, 1999. **39**(3): p. 203-18.
301. Hesson J., *Cumulative estrogen exposure and prospective memory in older women*. Brain Cogn, 2012. **80**(1): p. 89-95.
302. Atsma F., van der Schouw Y.T., Grobbee D.E., Kors J.A. and Bartelink M.L., *Lifetime endogenous estrogen exposure and electrocardiographic frontal T axis changes in postmenopausal women*. Maturitas, 2009. **63**(4): p. 347-51.
303. Alonso de Lecinana M., Egido J.A., Fernandez C., Martinez-Vila E., Santos S., et al., *Risk of ischemic stroke and lifetime estrogen exposure*. Neurology, 2007. **68**(1): p. 33-8.
304. Jansen S.C., Temme E.H. and Schouten E.G., *Lifetime estrogen exposure versus age at menopause as mortality predictor*. Maturitas, 2002. **43**(2): p. 105-12.
305. Green L.E., Dinh T.A. and Smith R.A., *An estrogen model: the relationship between body mass index, menopausal status, estrogen replacement therapy, and breast cancer risk*. Comput Math Methods Med, 2012. **2012**: p. 792375.
306. Du P., Zhang X., Huang C.C., Jafari N., Kibbe W.A., et al., *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis*. BMC Bioinformatics, 2010. **11**: p. 587.
307. Saadati M. and Benner A., *Statistical challenges of high-dimensional methylation data*. Stat Med, 2014. **33**(30): p. 5347-57.

308. Wahl S., Fenske N., Zeilinger S., Suhre K., Gieger C., et al., *On the potential of models for location and scale for genome-wide DNA methylation data*. BMC Bioinformatics, 2014. **15**: p. 232.
309. Ferrari S. and Cribari-Neto F., *Beta Regression for Modelling Rates and Proportions*. Journal of Applied Statistics, 2004. **31**(7): p. 799-815.
310. Yee T.W. *Package 'VGAM'*. [Accessed October 2018]; Available from: <https://cran.r-project.org/web/packages/VGAM/VGAM.pdf>.
311. Zeileis A., Cribari-Neto F., Gruen B., Kosmidis I., Simas A.B., et al. *Package 'betareg'*. [Accessed October 2018]; Available from: <https://cran.r-project.org/web/packages/betareg/betareg.pdf>.
312. Campanella G., Polidoro S., Di Gaetano C., Fiorito G., Guarrera S., et al., *Epigenetic signatures of internal migration in Italy*. Int J Epidemiol, 2015. **44**(4): p. 1442-9.
313. Kananen L., Marttila S., Nevalainen T., Jylhava J., Mononen N., et al., *Aging-associated DNA methylation changes in middle-aged individuals: the Young Finns study*. BMC Genomics, 2016. **17**: p. 103.
314. Nevalainen T., Kananen L., Marttila S., Jylha M., Hervonen A., et al., *Transcriptomic and epigenetic analyses reveal a gender difference in aging-associated inflammation: the Vitality 90+ study*. Age (Dordr), 2015. **37**(4): p. 9814.
315. Plusquin M., Guida F., Polidoro S., Vermeulen R., Raaschou-Nielsen O., et al., *DNA methylation and exposure to ambient air pollution in two prospective cohorts*. Environ Int, 2017. **108**: p. 127-136.
316. Tang Q., Holland-Letz T., Slynko A., Cuk K., Marme F., et al., *DNA methylation array analysis identifies breast cancer associated RPTOR, MGRN1 and RAPSN hypomethylation in peripheral blood DNA*. Oncotarget, 2016. **7**(39): p. 64191-64202.
317. Triche T.J., Laird P.W. and D. S.K., *Beta regression improves the detection of differential DNA methylation for epigenetic epidemiology*. bioRxiv.
318. van Iterson M., van Zwet E.W. and Heijmans B.T., *Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution*. Genome Biol, 2017. **18**(1): p. 19.
319. International Agency for Research on Cancer. *EPIC Study*. [Accessed September 2018]; Available from: <http://epic.iarc.fr>.
320. Riboli E., Hunt K.J., Slimani N., Ferrari P., Norat T., et al., *European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection*. Public Health Nutr, 2002. **5**(6b): p. 1113-24.
321. Palli D., Berrino F., Vineis P., Tumino R., Panico S., et al., *A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants*. Tumori, 2003. **89**(6): p. 586-93.
322. Staley J.R., Jones E., Kaptoge S., Butterworth A.S., Sweeting M.J., et al., *A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design*. Eur J Hum Genet, 2017. **25**(7): p. 854-62.

323. Vandenbroucke J.P. and Pearce N., *Case-control studies: basic concepts*. Int J Epidemiol, 2012. **41**(5): p. 1480-9.
324. Dedeurwaerder S., Defrance M., Calonne E., Denis H., Sotiriou C., et al., *Evaluation of the Infinium Methylation 450K technology*. Epigenomics, 2011. **3**(6): p. 771-84.
325. Teschendorff A.E., Marabita F., Lechner M., Bartlett T., Tegner J., et al., *A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data*. Bioinformatics, 2013. **29**(2): p. 189-96.
326. Johnson W.E., Li C. and Rabinovic A., *Adjusting batch effects in microarray expression data using empirical Bayes methods*. Biostatistics, 2007. **8**(1): p. 118-27.
327. Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., et al., *Missing value estimation methods for DNA microarrays*. Bioinformatics, 2001. **17**(6): p. 520-5.
328. Chen Y.A., Lemire M., Choufani S., Butcher D.T., Grafodatskaya D., et al., *Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray*. Epigenetics, 2013. **8**(2): p. 203-9.
329. Houseman E.A., Accomando W.P., Koestler D.C., Christensen B.C., Marsit C.J., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution*. BMC Bioinformatics, 2012. **13**: p. 86.
330. Amadou A., Ferrari P., Muwonge R., Moskal A., Biessy C., et al., *Overweight, obesity and risk of premenopausal breast cancer according to ethnicity: a systematic review and dose-response meta-analysis*. Obes Rev, 2013. **14**(8): p. 665-78.
331. Rosenberg L., Boggs D.A., Bethea T.N., Wise L.A., Adams-Campbell L.L., et al., *A prospective study of smoking and breast cancer risk among African-American women*. Cancer Causes Control, 2013. **24**(12): p. 2207-15.
332. Carroll Lab, Cambridge Research Institute, Data. [Accessed December 2018]; Available from: <http://www.carroll-lab.org.uk/data>.
333. Rakyan V.K., Down T.A., Balding D.J. and Beck S., *Epigenome-wide association studies for common human diseases*. Nat Rev Genet, 2011. **12**(8): p. 529-41.
334. Flanagan J.M., *Epigenome-wide association studies (EWAS): past, present, and future*. Methods Mol Biol, 2015. **1238**: p. 51-63.
335. Michels K.B., Binder A.M., Dedeurwaerder S., Epstein C.B., Greally J.M., et al., *Recommendations for the design and analysis of epigenome-wide association studies*. Nat Methods, 2013. **10**(10): p. 949-55.
336. Teschendorff A.E., Menon U., Gentry-Maharaj A., Ramus S.J., Gayther S.A., et al., *An epigenetic signature in peripheral blood predicts active ovarian cancer*. PLoS One, 2009. **4**(12): p. e8274.
337. Teschendorff A.E., Yang Z., Wong A., Pipinikas C.P., Jiao Y., et al., *Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer*. JAMA Oncol, 2015. **1**(4): p. 476-85.

338. Ek W.E., Tobi E.W., Ahsan M., Lampa E., Ponzi E., et al., *Tea and coffee consumption in relation to DNA methylation in four European cohorts*. Hum Mol Genet, 2017. **26**(16): p. 3221-3231.
339. Pharoah P.D., Antoniou A.C., Easton D.F. and Ponder B.A., *Polygenes, risk prediction, and targeted prevention of breast cancer*. N Engl J Med, 2008. **358**(26): p. 2796-803.
340. Tibshirani R., *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological), 1996. **58**(1): p. 267-288.
341. Hastie T., Tibshirani R. and Friedman J., *The elements of statistical learning*. 2008. p. 43-100.
342. Friedman J., Hastie T. and Tibshirani R., *Regularization Paths for Generalized Linear Models via Coordinate Descent*. J Stat Softw, 2010. **33**(1): p. 1-22.
343. Zou H. and Hastie T., *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society. Series B (Methodological), 2005. **67**(2): p. 301-320.
344. Swerdlow A.J., Jones M.E., Schoemaker M.J., Hemming J., Thomas D., et al., *The Breakthrough Generations Study: design of a long-term UK cohort study to investigate breast cancer aetiology*. Br J Cancer, 2011. **105**(7): p. 911-7.
345. Johnson M.D., Mueller M., Adamowicz-Brice M., Collins M.J., Gellert P., et al., *Genetic analysis of the cardiac methylome at single nucleotide resolution in a model of human cardiovascular disease*. PLoS Genet, 2014. **10**(12): p. e1004813.
346. Li Z., Guo X., Wu Y., Li S., Yan J., et al., *Methylation profiling of 48 candidate genes in tumor and matched normal tissues from breast cancer patients*. Breast Cancer Res Treat, 2015. **149**(3): p. 767-79.
347. Paliwal A., Temkin A.M., Kerkel K., Yale A., Yotova I., et al., *Comparative anatomy of chromosomal domains with imprinted and non-imprinted allele-specific DNA methylation*. PLoS Genet, 2013. **9**(8): p. e1003622.
348. Bundo M., Sunaga F., Ueda J., Kasai K., Kato T., et al., *A systematic evaluation of whole genome amplification of bisulfite-modified DNA*. Clin Epigenetics, 2012. **4**(1): p. 22.
349. Adamowicz M., Maratou K. and Aitman T.J., *Multiplexed DNA Methylation Analysis of Target Regions Using Microfluidics (Fluidigm)*. Methods Mol Biol, 2018. **1708**: p. 349-363.
350. Fluidigm Corporation, *Access Array System for Illumina Sequencing System - User Guide - PN 100-3770 J1* 2016.
351. Babraham Bioinformatics. *Trim Galore!* [Accessed March 2018]; Available from: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
352. Bolger A.M., Lohse M. and Usadel B., *Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics, 2014. **30**(15): p. 2114-20.
353. Krueger F. and Andrews S.R., *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications*. Bioinformatics, 2011. **27**(11): p. 1571-2.
354. Hunn J. and Rodriguez G.C., *Ovarian cancer: etiology, risk factors, and epidemiology*. Clin Obstet Gynecol, 2012. **55**(1): p. 3-23.

355. Salehi F., Dunfield L., Phillips K.P., Krewski D. and Vanderhyden B.C., *Risk factors for ovarian cancer: an overview with emphasis on hormonal factors*. J Toxicol Environ Health B Crit Rev, 2008. **11**(3-4): p. 301-21.
356. Adalsteinsson B.T., Gudnason H., Aspelund T., Harris T.B., Launer L.J., et al., *Heterogeneity in white blood cells has potential to confound DNA methylation measurements*. PLoS One, 2012. **7**(10): p. e46705.
357. Jones M.J., Islam S.A., Edgar R.D. and Kobor M.S., *Adjusting for Cell Type Composition in DNA Methylation Data Using a Regression-Based Approach*. Methods Mol Biol, 2017. **1589**: p. 99-106.
358. Kile M.L., Houseman E.A., Baccarelli A.A., Quamruzzaman Q., Rahman M., et al., *Effect of prenatal arsenic exposure on DNA methylation and leukocyte subpopulations in cord blood*. Epigenetics, 2014. **9**(5): p. 774-82.
359. Stroup D.F., Berlin J.A., Morton S.C., Olkin I., Williamson G.D., et al., *Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group*. Jama, 2000. **283**(15): p. 2008-12.
360. Riley R.D., Higgins J.P. and Deeks J.J., *Interpretation of random effects meta-analyses*. Bmj, 2011. **342**: p. d549.
361. Brockwell S.E. and Gordon I.R., *A comparison of statistical methods for meta-analysis*. Stat Med, 2001. **20**(6): p. 825-40.
362. Kelley G.A. and Kelley K.S., *Statistical models for meta-analysis: A brief tutorial*. World J Methodol, 2012. **2**(4): p. 27-32.
363. Higgins J.P. and Thompson S.G., *Quantifying heterogeneity in a meta-analysis*. Stat Med, 2002. **21**(11): p. 1539-58.
364. DerSimonian R. and Laird N., *Meta-analysis in clinical trials*. Control Clin Trials, 1986. **7**(3): p. 177-88.
365. Viechtbauer W., *Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model*. Journal of Educational and Behavioral Statistics, 2005. **30**(3): p. 261-293.
366. Thompson S.G. and Sharp S.J., *Explaining heterogeneity in meta-analysis: a comparison of methods*. Stat Med, 1999. **18**(20): p. 2693-708.
367. Aryee M.J., Jaffe A.E., Corrada-Bravo H., Ladd-Acosta C., Feinberg A.P., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays*. Bioinformatics, 2014. **30**(10): p. 1363-9.
368. Maksimovic J., Gordon L. and Oshlack A., *SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips*. Genome Biol, 2012. **13**(6): p. R44.
369. Suenaga R., Evans M.J., Mitamura K., Rider V. and Abdou N.I., *Peripheral blood T cells and monocytes and B cell lines derived from patients with lupus express estrogen receptor transcripts similar to those of normal cells*. J Rheumatol, 1998. **25**(7): p. 1305-12.
370. Rider V., Li X., Peterson G., Dawson J., Kimler B.F., et al., *Differential expression of estrogen receptors in women with systemic lupus erythematosus*. J Rheumatol, 2006. **33**(6): p. 1093-101.

371. Zhang J., Chen X., Zhang S., Zhou G., Xia X., et al., *Effects of transdermal estrogen therapy on expressions of estrogen receptors and T-lymphocyte apoptosis in surgically menopausal women*. Cell Mol Immunol, 2009. **6**(4): p. 277-83.
372. Shim G.J., Gherman D., Kim H.J., Omoto Y., Iwase H., et al., *Differential expression of oestrogen receptors in human secondary lymphoid tissues*. J Pathol, 2006. **208**(3): p. 408-14.
373. Phiel K.L., Henderson R.A., Adelman S.J. and Elloso M.M., *Differential estrogen receptor gene expression in human peripheral blood mononuclear cell populations*. Immunol Lett, 2005. **97**(1): p. 107-13.
374. Kassi E.N., Vlachoyiannopoulos P.G., Moutsopoulos H.M., Sekeris C.E. and Moutsatsou P., *Molecular analysis of estrogen receptor alpha and beta in lupus patients*. Eur J Clin Invest, 2001. **31**(1): p. 86-93.
375. Nancy P. and Berrih-Aknin S., *Differential estrogen receptor expression in autoimmune myasthenia gravis*. Endocrinology, 2005. **146**(5): p. 2345-53.
376. Pierdominici M., Maselli A., Colasanti T., Giammarioli A.M., Delunardo F., et al., *Estrogen receptor profiles in human peripheral blood lymphocytes*. Immunol Lett, 2010. **132**(1-2): p. 79-85.
377. Scariano J.K., Emery-Cohen A.J., Pickett G.G., Morgan M., Simons P.C., et al., *Estrogen receptors alpha (ESR1) and beta (ESR2) are expressed in circulating human lymphocytes*. J Recept Signal Transduct Res, 2008. **28**(3): p. 285-93.
378. Pelekanou V., Kampa M., Kiagiadaki F., Deli A., Theodoropoulos P., et al., *Estrogen anti-inflammatory activity on human monocytes is mediated through cross-talk between estrogen receptor ERalpha36 and GPR30/GPER1*. J Leukoc Biol, 2016. **99**(2): p. 333-47.
379. Tosato G. and Cohen J.I., *Generation of Epstein-Barr Virus (EBV)-immortalized B cell lines*. Curr Protoc Immunol, 2007. **Chapter 7**: p. Unit 7.22.
380. Hui-Yuen J., McAllister S., Koganti S., Hill E. and Bhaduri-McIntosh S., *Establishment of Epstein-Barr virus growth-transformed lymphoblastoid cell lines*. J Vis Exp, 2011(57).
381. Hussain T. and Mulherkar R., *Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair*. Int J Mol Cell Med, 2012. **1**(2): p. 75-87.
382. Rizza P., Barone I., Zito D., Giordano F., Lanzino M., et al., *Estrogen receptor beta as a novel target of androgen receptor action in breast cancer cell lines*. Breast Cancer Res, 2014. **16**(1): p. R21.
383. Al-Bader M., Ford C., Al-Ayadhy B. and Francis I., *Analysis of estrogen receptor isoforms and variants in breast cancer cell lines*. Exp Ther Med, 2011. **2**(3): p. 537-544.
384. Ford C.H., Al-Bader M., Al-Ayadhi B. and Francis I., *Reassessment of estrogen receptor expression in human breast cancer cell lines*. Anticancer Res, 2011. **31**(2): p. 521-7.
385. Monroe D.G., Getz B.J., Johnsen S.A., Riggs B.L., Khosla S., et al., *Estrogen receptor isoform-specific regulation of endogenous gene expression in human osteoblastic cell lines expressing either ERalpha or ERbeta*. J Cell Biochem, 2003. **90**(2): p. 315-26.
386. Barnes D. and Sato G., *Serum-free cell culture: a unifying approach*. Cell, 1980. **22**(3): p. 649-55.

387. Berthois Y., Katzenellenbogen J.A. and Katzenellenbogen B.S., *Phenol red in tissue culture media is a weak estrogen: implications concerning the study of estrogen-responsive cells in culture*. Proc Natl Acad Sci U S A, 1986. **83**(8): p. 2496-500.
388. Rogakou E.P., Pilch D.R., Orr A.H., Ivanova V.S. and Bonner W.M., *DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139*. J Biol Chem, 1998. **273**(10): p. 5858-68.
389. Krauer K.G., Buck M., Belzer D.K., Flanagan J., Chojnowski G.M., et al., *The Epstein-Barr virus nuclear antigen-6 protein co-localizes with EBNA-3 and survival of motor neurons protein*. Virology, 2004. **318**(1): p. 280-94.
390. Nelson A.W., Groen A.J., Miller J.L., Warren A.Y., Holmes K.A., et al., *Comprehensive assessment of estrogen receptor beta antibodies in cancer cell line models and tissue reveals critical limitations in reagent specificity*. Mol Cell Endocrinol, 2017. **440**: p. 138-150.
391. Hafez A.Y. and Luftig M.A., *Characterization of the EBV-Induced Persistent DNA Damage Response*. Viruses, 2017. **9**(12).
392. Zijno A., Porcedda P., Saini F., Allione A., Garofalo B., et al., *Unsuitability of lymphoblastoid cell lines as surrogate of cryopreserved isolated lymphocytes for the analysis of DNA double-strand break repair activity*. Mutat Res, 2010. **684**(1-2): p. 98-105.
393. Lovelock P.K., Wong E.M., Sprung C.N., Marsh A., Hobson K., et al., *Prediction of BRCA1 and BRCA2 mutation status using post-irradiation assays of lymphoblastoid cell lines is compromised by inter-cell-line phenotypic variability*. Breast Cancer Res Treat, 2007. **104**(3): p. 257-66.
394. Sugawara H., Iwamoto K., Bundo M., Ueda J., Ishigooka J., et al., *Comprehensive DNA methylation analysis of human peripheral blood leukocytes and lymphoblastoid cell lines*. Epigenetics, 2011. **6**(4): p. 508-15.
395. Taniguchi I., Iwaya C., Ohnaka K., Shibata H. and Yamamoto K., *Genome-wide DNA methylation analysis reveals hypomethylation in the low-CpG promoter regions in lymphoblastoid cell lines*. Hum Genomics, 2017. **11**(1): p. 8.
396. Aberg K., Khachane A.N., Rudolf G., Nerella S., Fugman D.A., et al., *Methylome-wide comparison of human genomic DNA extracted from whole blood and from EBV-transformed lymphocyte cell lines*. Eur J Hum Genet, 2012. **20**(9): p. 953-5.
397. Thompson T.M., Sharfi D., Lee M., Yrigollen C.M., Naumova O.Y., et al., *Comparison of whole-genome DNA methylation patterns in whole blood, saliva, and lymphoblastoid cell lines*. Behav Genet, 2013. **43**(2): p. 168-76.
398. Yue W., Santen R.J., Wang J.P., Li Y., Verderame M.F., et al., *Genotoxic metabolites of estradiol in breast: potential mechanism of estradiol induced carcinogenesis*. J Steroid Biochem Mol Biol, 2003. **86**(3-5): p. 477-86.
399. Fortin J.P., Labbe A., Lemire M., Zanke B.W., Hudson T.J., et al., *Functional normalization of 450k methylation array data improves replication in large cancer studies*. Genome Biol, 2014. **15**(12): p. 503.

400. Min J.L., Hemani G., Davey Smith G., Relton C. and Suderman M., *Meffil: efficient normalization and analysis of very large DNA methylation datasets*. Bioinformatics, 2018.
401. Hunter D.J., Kraft P., Jacobs K.B., Cox D.G., Yeager M., et al., *A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer*. Nat Genet, 2007. **39**(7): p. 870-4.
402. Michailidou K., Beesley J., Lindstrom S., Canisius S., Dennis J., et al., *Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer*. Nat Genet, 2015. **47**(4): p. 373-80.
403. Olive P.L. and Banath J.P., *The comet assay: a method to measure DNA damage in individual cells*. Nat Protoc, 2006. **1**(1): p. 23-9.

APPENDICES

- Appendix 1: Hormonal factors and breast cancer risk in EPIC-Italy
- Appendix 2: The 212 top hits from the EWAS of ELEE
- Appendix 3. List of primer sequences for targeted bisulphite sequencing
- Appendix 4 All CpG sites included in the target regions and breast cancer risk in the Generations Study
- Appendix 5. Script for the meta-analysis
- Appendix 6 Primer sequences for the bisulphite pyrosequencing assays

Appendix 1: Hormonal factors and breast cancer risk in EPIC-Italy

	All women				Premenopausal women				Postmenopausal women			
	Controls	Cases	HR (95% CI)	P value	Controls	Cases	HR (95% CI)	P value	Controls	Cases	HR (95% CI)	P value
Age at menarche	30263 408	1180 13	0.94 (0.91 - 0.98)	0.003	15084 2	477 0	0.97 (0.91 - 1.03)	0.281	15161 2	703 0	0.94 (0.89 - 0.98)	0.008
≤ 12	15166	609	Ref	Ref	8267	260	Ref	Ref	6892	349	Ref	Ref
≥ 13	15097	571	0.85 (0.76 - 0.96)	0.007	6817	217	0.92 (0.77 - 1.10)	0.348	8269	354	0.82 (0.71 - 0.96)	0.011
Age at menarche	15166	609	Ref	Ref	8267	260	Ref	Ref	6892	349	Ref	Ref
≤ 12	7412	267	0.85 (0.74 - 0.99)	0.031	3712	113	0.92 (0.74 - 1.15)	0.459	3695	154	0.81 (0.67 - 0.98)	0.034
13	5188	225	0.96 (0.82 - 1.11)	0.561	2248	89	1.09 (0.86 - 1.39)	0.465	2938	136	0.89 (0.73 - 1.08)	0.235
14	2497	79	0.64 (0.51 - 0.82)	2.67E-04	857	15	0.46 (0.27 - 0.77)	0.003	1636	64	0.73 (0.56 - 0.96)	0.025
≥ 15	408	13			2	0			2	0		
Age at menopause									15084	699	1.02 (1.01 - 1.04)	0.006
≤ 49									79	4		
≥ 50									7638	311	Ref	Ref
Age at menopause									7446	388	1.25 (1.06 - 1.46)	0.006
≤ 45									79	4		
46 - 49									3513	131	Ref	Ref
50 - 51									4125	180	1.16 (0.92 - 1.45)	0.207
≥ 52									3561	169	1.25 (0.99 - 1.58)	0.064
Parity	27125	1060	1.03 (0.85 - 1.24)	0.769	13575	434	0.99 (0.72 - 1.35)	0.928	13531	626	0.99 (0.78 - 1.25)	0.928
Parous	3146	120	Ref	Ref	1511	43	Ref	Ref	1632	77	Ref	Ref
Nulliparous	400	13			0	0			0	0		
Number of pregnancies	29367	1130	0.92 (0.87 - 0.97)	0.003	14556	451	0.86 (0.78 - 0.94)	0.002	14791	679	0.94 (0.88 - 1.01)	0.070
≤ 2	23110	910	Ref	Ref	530	26	Ref	Ref	372	24		
3	4618	164	0.85 (0.72 - 1.00)	0.053	11939	383	Ref	Ref	11155	527	Ref	Ref
≥ 4	1639	56	0.74 (0.56 - 0.97)	0.028	2077	51	0.68 (0.50 - 0.91)	0.009	2538	113	0.93 (0.76 - 1.14)	0.503
Age at first pregnancy	26209	1010	1.02 (1.00 - 1.03)	0.008	540	17	0.78 (0.48 - 1.27)	0.322	1098	39	0.73 (0.53 - 1.02)	0.062
≤ 23	8246	258	Ref	Ref	330	26			372	24		
24 - 25	5528	217	1.17 (0.98 - 1.40)	0.087	13040	408	1.03 (1.01 - 1.05)	0.006	13152	602	1.01 (0.99 - 1.03)	0.181
26 - 28	6295	273	1.28 (1.08 - 1.51)	0.005	2046	69	Ref	Ref	2011	101	Ref	Ref
≥ 29	6140	262	1.25 (1.06 - 1.49)	0.010	4617	103	1.58 (1.20 - 2.08)	0.001	3626	155	0.91 (0.71 - 1.16)	0.429
Age at first pregnancy					2648	104	1.35 (1.02 - 1.78)	0.037	2880	113	1.19 (0.96 - 1.48)	0.107
≤ 23					2871	95	1.56 (1.19 - 2.05)	0.001	3418	178	1.10 (0.88 - 1.37)	0.419
24 - 25					2904	106			3228	156		
26 - 28					2046	69			2011	101		
≥ 29												

Breastfeeding ever	Yes No Missing	21950 4403 4318	836 178 179	0.92 (0.78 - 1.08) Ref Ref	0.314 Ref Ref	10871 2257 1958	340 71 66	1.01 (0.78 - 1.30) Ref Ref	0.942 Ref Ref	11065 2143 1955	496 107 100	0.88 (0.72 - 1.09) Ref Ref	0.245 Ref Ref
Breastfeeding duration	Missing	27525 3146	1073 120	0.88 (0.80 - 0.97) 0.008	0.008	13575 1511	434 43	0.86 (0.72 - 1.02) Ref	0.086 Ref	13531 1632	626 77	0.91 (0.81 - 1.02) Ref	0.116 Ref
Breastfeeding duration	< 0.25 0.25 - 0.75 ≥ 0.75	11282 8760 7483	449 330 294	Ref 0.93 (0.80 - 1.07) 0.88 (0.76 - 1.03)	Ref 0.289 0.102	5853 4533 3189	193 149 92	Ref 0.98 (0.79 - 1.21) 0.84 (0.65 - 1.08)	Ref 0.855 0.169	5020 4221 4290	243 181 202	0.88 (0.73 - 1.07) 0.88 (0.73 - 1.07) 0.94 (0.77 - 1.13)	Ref 0.195 0.487
BMI	Missing	30403 268	1186 7	1.01 (1.00 - 1.03) 0.035	0.035	14956 130	475 2	1.00 (0.98 - 1.02) Ref	0.733 Ref	15085 78	699 4	1.02 (1.01 - 1.04) Ref	0.010 Ref
HRT ever	Yes No Missing	4739 25386 546	222 955 16	1.10 (0.95 - 1.28) Ref Ref	0.198 Ref Ref	1313 13699 74	64 411 2	0.93 (0.71 - 1.23) Ref Ref	0.628 Ref Ref	3424 11670 69	158 544 1	1.02 (0.85 - 1.21) Ref Ref	0.870 Ref Ref
HRT duration	Missing	30271 400	1180 13	1.00 (0.96 - 1.05) 0.892	0.892	15086 0	477 0	1.00 (0.92 - 1.08) Ref	0.925 Ref	15163 0	703 0	0.97 (0.92 - 1.03) Ref	0.303 Ref
HRT duration (yrs)	< 1 1 - 3 ≥ 4 Missing	27534 1909 828 400	1044 103 33 13	Ref 1.25 (1.02 - 1.53) 0.90 (0.64 - 1.28)	Ref 0.033 0.564	14389 469 228 0	436 30 11 0	Ref 1.16 (0.79 - 1.70) 0.71 (0.38 - 1.31)	Ref 0.444 0.273	13123 1440 600 0	608 73 22 0	Ref 1.12 (0.88 - 1.43) 0.81 (0.53 - 1.24)	Ref 0.371 0.339 Ref
OC ever	Yes No Missing	12434 17774 463	437 741 15	1.07 (0.94 - 1.21) Ref Ref	0.296 Ref Ref	8303 6753 30	243 233 1	1.03 (0.86 - 1.23) Ref Ref	0.774 Ref Ref	4123 11009 31	194 508 1	1.07 (0.90 - 1.27) Ref Ref	0.416 Ref Ref
OC duration	Missing	30271 400	1180 13	1.01 (0.99 - 1.03) 0.281	0.281	15086 0	477 0	1.01 (0.99 - 1.04) Ref	0.345 Ref	15163 0	703 0	1.01 (0.98 - 1.03) Ref	0.697 Ref
OC duration (yrs)	0 - 2 3 - 5 ≥ 6 Missing	24590 2732 2949 400	977 90 113 13	Ref 1.02 (0.82 - 1.27) 1.18 (0.96 - 1.43)	Ref 0.856 0.109	11135 1866 2085 0	357 53 67 0	Ref 1.04 (0.78 - 1.39) 1.16 (0.89 - 1.51)	Ref 0.785 0.271	13434 865 864 0	620 37 46 0	Ref 0.98 (0.70 - 1.37) 1.19 (0.88 - 1.61)	Ref 0.905 0.247 Ref
Drink alcohol	Yes Previous Never Missing	18200 6496 581 5394	737 282 21 153	1.20 (0.78 - 1.85) 1.24 (0.80 - 1.93) Ref Ref	0.409 0.344 Ref Ref	9311 2910 209 2656	302 98 7 70	1.10 (0.52 - 2.32) 1.16 (0.54 - 2.51) Ref Ref	0.806 0.697 Ref Ref	8885 3585 372 2321	435 184 14 70	1.29 (0.76 - 2.19) 1.34 (0.78 - 2.31) Ref Ref	0.353 0.287 Ref Ref
Smoking status	Smoker Former Never Missing	7955 6100 16216 400	285 223 672 13	0.97 (0.84 - 1.12) 0.97 (0.83 - 1.13) Ref Ref	0.665 0.679 Ref Ref	4445 3404 7237 0	146 107 224 0	1.23 (0.99 - 1.51) 1.12 (0.89 - 1.41) Ref Ref	0.057 0.351 Ref Ref	3503 2694 8966 0	139 116 448 0	0.81 (0.67 - 0.99) 0.88 (0.72 - 1.08) Ref Ref	0.036 0.222 Ref Ref
Smoking duration (yrs)	Missing	30220 451	1178 15	1.00 (1.00 - 1.00) 0.753	0.753	15060 26	477 0	1.01 (1.00 - 1.01) Ref	0.059 Ref	15139 24	701 2	1.00 (0.99 - 1.00) Ref	0.070 Ref
Smoking intensity	Missing	27333 3338	1098 95	1.02 (0.99 - 1.05) 0.142	0.142	13357 1729	428 49	1.08 (1.03 - 1.13) Ref	8.09E-04 Ref	13962 1201	670 33	0.98 (0.95 - 1.03) Ref	0.441 Ref

Appendix 2: The 212 top hits from the EWAS of ELEE

HM450K probe	Chromosome	Position	Nearest gene	Distance to gene	UCSC RefGene Group	Relation to CpG Island	EWAS of ELEE				St.dev beta-values
							Estimate	SE	P-value	Q-value	
cg23051559	chr11	119015827	ABCG4	3921	Intergenic	N_Shelf	-0.15	0.04	2.96E-05	0.036	3.65
cg26252045	chr11	34256703	ABTB2	0	Body	OpenSea	-0.12	0.04	1.36E-06	0.011	3.08
cg08335389	chr12	109554302	ACACB	96	Intergenic	OpenSea	-0.24	0.07	4.73E-05	0.042	4.26
cg03785646	chr2	236774164	AGAP1	0	Body	S_Shore	-0.52	0.12	1.62E-07	0.005	7.74
cg15981851	chr2	236773924	AGAP1	0	Body	Island	-0.31	0.09	7.18E-05	0.048	5.89
cg19761110	chr2	236774071	AGAP1	0	Body	Island	-0.29	0.10	4.98E-05	0.043	5.94
cg00811930	chr14	96861856	AK7	0	Body	S_Shelf	-0.27	0.07	1.17E-05	0.027	4.89
cg00643897	chr2	30142677	ALK	0	Body	N_Shore	-0.28	0.09	6.74E-05	0.047	5.55
cg01154254	chr2	71205159	ANKRD53	414	TSS1500	N_Shore	-0.23	0.05	4.66E-07	0.006	3.47
cg18994049	chr11	69813825	ANO1	110581	Intergenic	S_Shore	-0.13	0.04	1.02E-05	0.025	2.50
cg00740803	chr4	113143664	AP1AR	9229	Intergenic	OpenSea	0.21	0.07	1.76E-05	0.029	4.26
cg03768916	chr10	49813307	ARHGAP22	0	TSS200	Island	-0.13	0.04	6.32E-05	0.046	2.92
cg24549608	chrX	101854372	ARMCX5	0	5'UTR	Island	-0.20	0.06	2.19E-05	0.031	3.55
cg17396404	chr6	32086878	ATF6B	0	Body	OpenSea	-0.16	0.05	7.35E-05	0.048	3.23
cg09705798	chr5	160048394	ATP10B	0	Body	OpenSea	-0.21	0.06	1.25E-05	0.027	4.19
cg13941987	chrX	129118022	BCORL1	21140	5'UTR	Island	-0.15	0.04	1.89E-05	0.030	2.49
cg25517187	chr10	88626981	BMPRI1A	0	5'UTR	OpenSea	-0.17	0.05	6.69E-06	0.021	3.35
cg08254089	chr20	36933189	BPI	0	Body	OpenSea	-0.28	0.08	7.04E-05	0.048	5.19
cg25063710	chrX	80064750	BRWD3	0	Body	N_Shore	-0.39	0.10	3.92E-05	0.039	6.56
cg21590238	chr12	121454837	C12orf43	536	TSS1500	S_Shore	-0.29	0.08	2.67E-06	0.014	5.01
cg26143719	chr22	37584227	C1QTNF6	0	1stExon	OpenSea	-0.23	0.06	1.07E-05	0.025	4.18
cg09562331	chr3	50596116	C3orf18	0	3'UTR	OpenSea	-0.15	0.04	3.25E-06	0.015	3.27
cg06537801	chr1	57315465	C8A	4976	Intergenic	OpenSea	-0.17	0.04	3.35E-06	0.015	2.82
cg08601628	chr8	10555466	C8orf74	0	Body	OpenSea	-0.23	0.07	4.51E-07	0.006	4.09
cg24658867	chr7	134626923	CALD1	0	Body	OpenSea	-0.18	0.05	5.55E-05	0.044	3.05
cg19964454	chr11	69458595	CCND1	0	Body	Island	-0.19	0.04	6.90E-06	0.021	2.92
cg27317813	chr12	4398508	CCND2	0	Body	OpenSea	0.15	0.05	5.64E-06	0.021	6.71
cg08969925	chr3	33015641	CCR4	19237	Intergenic	OpenSea	-0.16	0.06	4.95E-05	0.043	4.30
cg16075139	chr11	2406923	CD81	0	Body	Island	-0.14	0.03	2.75E-06	0.014	3.60
cg09313587	chr19	3295021	CELF5	0	3'UTR	S_Shore	-0.17	0.06	5.07E-05	0.043	3.50
cg21153102	chr15	41252147	CHAC1	3429	Intergenic	OpenSea	-0.43	0.14	6.16E-05	0.046	9.42
cg05308645	chr3	184107534	CHRD	0	3'UTR	S_Shelf	-0.18	0.04	1.76E-05	0.029	2.86
cg23100671	chr10	125733991	CHST15	33189	Intergenic	S_Shore	-0.14	0.05	8.39E-05	0.050	2.86
cg16922039	chr18	12287385	CIDEA	9790	Intergenic	Island	-0.18	0.04	1.04E-05	0.025	3.59
cg01895740	chr19	7852739	CLEC4GP1	0	Body	N_Shore	-0.29	0.08	1.70E-05	0.029	4.93
cg23501228	chr6	75917468	COL12A1	1844	Intergenic	S_Shore	-0.25	0.07	3.63E-05	0.038	4.57
cg07630274	chr10	71583235	COL13A1	0	Body	OpenSea	-0.15	0.05	2.22E-07	0.005	3.77
cg22232704	chr13	110961606	COL4A2	0	Body	S_Shore	-0.40	0.12	6.90E-05	0.048	7.51
cg08772752	chr3	48632672	COL7A1	78	TSS200	Island	-0.13	0.05	4.23E-05	0.040	2.84
cg03340215	chr15	83315615	CPEB1	0	Body	Island	-0.31	0.08	4.78E-05	0.042	4.90
cg08686879	chr5	131409634	CSF2	0	1stExon	OpenSea	-0.12	0.03	6.97E-05	0.048	3.12

cg06968859	chr2	80724209	CTNNA2	0	Body	OpenSea	-0.26	0.05	2.52E-09	0.001	4.70
cg14503537	chr15	75044351	CYP1A2	0	Body	OpenSea	-0.16	0.03	1.82E-05	0.029	2.71
cg07904073	chr10	74017122	DDIT4	16553	Intergenic	N_Shelf	0.14	0.06	5.53E-05	0.044	5.94
cg05566582	chr7	157257942	DNAJB6	47808	Intergenic	N_Shelf	-0.42	0.10	1.53E-05	0.028	6.60
cg05771997	chr1	205175985	DSTYK	0	Body	N_Shelf	-0.17	0.05	5.01E-05	0.043	3.28
cg16571825	chr8	49648297	EFCAB1	426	TSS1500	S_Shore	-0.22	0.06	6.93E-05	0.048	4.26
cg16076328	chr2	47597118	EPCAM	0	Body	Island	-0.26	0.07	7.97E-05	0.049	4.93
cg01337514	chr1	6511026	ESPN	0	Body	N_Shore	-0.18	0.05	3.36E-05	0.037	3.22
cg13408086	chr9	140221397	EXD3	0	Body	S_Shelf	0.21	0.07	3.48E-05	0.038	6.02
cg07462256	chrX	68725088	FAM155B	0	1stExon	Island	-0.26	0.08	6.51E-05	0.047	4.89
cg09438069	chr4	187581093	FAT1	0	Body	OpenSea	-1.60	0.44	4.76E-05	0.042	27.00
cg22243439	chr4	187580694	FAT1	0	Body	OpenSea	-0.16	0.04	1.07E-05	0.025	2.49
cg10290724	chr19	50312544	FUZ	0	Body	Island	-0.20	0.06	1.80E-05	0.029	4.41
cg25085950	chr21	27108830	GABPA	0	TSS1500	S_Shore	0.12	0.04	3.19E-05	0.037	3.06
cg19005707	chr15	27111541	GABRB3	0	TSS1500	N_Shore	0.24	0.08	6.50E-05	0.047	5.38
cg03143347	chr12	132893664	GALNT9	0	Body	N_Shelf	-0.18	0.04	6.36E-07	0.007	2.85
cg03975235	chr22	38220611	GALR3	0	Body	Island	-0.14	0.04	6.23E-05	0.046	2.90
cg27211198	chr2	121776456	GLI2	26226	Intergenic	OpenSea	-0.31	0.08	2.06E-05	0.031	4.98
cg02780849	chr1	235814163	GNG4	108	TSS1500	N_Shore	-0.20	0.05	1.75E-05	0.029	4.16
cg17345569	chr6	42928274	GNMT	224	TSS1500	Island	0.16	0.04	2.62E-05	0.034	3.30
cg16840364	chr4	84539569	GPAT3	12541	Intergenic	OpenSea	-0.14	0.06	4.77E-05	0.042	3.50
cg07799395	chr5	125577565	GRAMD3	118221	Intergenic	OpenSea	-0.17	0.05	2.10E-06	0.013	3.64
cg08058213	chr4	156715823	GUCY1B3	0	Body	OpenSea	-0.15	0.04	1.82E-05	0.029	2.99
cg00231966	chr17	80377299	HEXDC	0	TSS1500	S_Shore	0.29	0.10	7.04E-05	0.048	6.78
cg24838316	chr6	29895260	HLA-J	0	Intergenic	S_Shore	-1.51	0.34	7.10E-06	0.022	20.87
cg23681866	chr6	29895175	HLA-J	0	Intergenic	S_Shore	-1.33	0.31	2.17E-05	0.031	19.24
cg03495030	chr12	121437256	HNF1A	0	Body	OpenSea	-0.41	0.11	4.36E-05	0.041	7.03
cg22968966	chr16	22959875	HS3ST2	32215	Intergenic	Island	-0.40	0.12	7.95E-05	0.049	7.88
cg03210314	chr16	22959594	HS3ST2	31934	Intergenic	Island	-0.26	0.08	6.15E-05	0.046	5.29
cg09740468	chr6	3025325	HTATSF1P2	319	Intergenic	OpenSea	-0.18	0.04	2.70E-05	0.034	5.21
cg26367719	chr19	55875605	IL11	143	TSS1500	S_Shore	-0.20	0.05	1.68E-05	0.029	3.94
cg00239353	chr16	3115133	IL32	178	TSS1500	OpenSea	-0.16	0.05	1.44E-05	0.028	5.00
cg21950518	chr5	55290746	IL6ST	0	1stExon	Island	-0.20	0.06	1.50E-06	0.012	3.65
cg11264584	chr11	15135663	INSC	0	Body	N_Shore	-0.22	0.07	1.35E-05	0.028	4.20
cg15127563	chr2	231729487	ITM2C	132	TSS200	N_Shore	0.36	0.09	1.77E-05	0.029	5.97
cg20020161	chr2	231732669	ITM2C	0	Body	S_Shore	-0.14	0.04	4.49E-05	0.041	2.65
cg23597162	chr7	28102341	JAZF1	0	Body	OpenSea	-0.18	0.05	7.34E-05	0.048	5.21
cg16773814	chr16	87644704	JPH3	0	Body	N_Shelf	-0.19	0.05	8.22E-05	0.050	3.80
cg24394097	chr3	155860792	KCNAB1	0	1stExon	OpenSea	-0.16	0.05	6.50E-05	0.047	3.07
cg22097768	chr17	61615913	KCNH6	0	Body	Island	-0.20	0.05	4.88E-05	0.042	3.26
cg19937056	chr3	178275698	KCNMB2	0	TSS1500	OpenSea	-0.15	0.06	2.62E-05	0.034	4.37
cg18587984	chr8	136473366	KHDRBS3	0	Body	S_Shelf	-0.16	0.05	5.21E-05	0.043	3.97
cg15509065	chr19	36346314	KIRREL2	1494	TSS1500	N_Shore	-0.18	0.05	8.20E-06	0.023	3.30
cg27066295	chr16	87778903	KLHDC4	0	Body	OpenSea	-0.15	0.04	3.18E-05	0.037	2.69
cg17969123	chr19	18745971	KLHL26	1865	Intergenic	OpenSea	-0.15	0.04	3.18E-05	0.037	2.64

cg18190861	chr12	52689098	KRT86	0	Intergenic	S_Shelf	-0.32	0.09	4.74E-05	0.042	6.33
cg13136556	chr21	46424880	LINC00162	237	TSS1500	S_Shelf	-0.30	0.08	1.21E-05	0.027	5.03
cg27511206	chr13	107314819	LINC00443	0	Intergenic	OpenSea	-0.15	0.05	1.98E-05	0.030	3.23
cg22306038	chr17	48845758	LINC00483	881	TSS1500	OpenSea	-0.12	0.03	3.94E-05	0.039	3.38
cg20187049	chr3	28982939	LINC00693	183110	Intergenic	OpenSea	-0.23	0.07	4.07E-05	0.040	5.44
cg07195851	chr5	3517522	LINC01019	0	Intergenic	OpenSea	-0.38	0.12	6.54E-05	0.047	7.69
cg14018513	chr10	134789814	LINC01168	0	Intergenic	N_Shore	-0.18	0.06	1.06E-05	0.025	3.70
cg25104872	chr6	28828977	LINC01623	0	Intergenic	N_Shelf	0.26	0.07	6.61E-05	0.047	4.41
cg04635069	chr7	156636277	LMBR1	0	Body	OpenSea	-0.20	0.07	3.32E-05	0.037	4.03
cg17269733	chr5	43041146	LOC153684	529	TSS1500	Island	0.16	0.05	8.57E-05	0.050	3.15
cg06682039	chr6	170536047	LOC154449	27373	Intergenic	OpenSea	-0.27	0.08	7.52E-05	0.048	7.91
cg01802043	chr5	172903876	LOC285593	102768	Intergenic	OpenSea	-0.14	0.05	8.23E-05	0.050	3.02
cg00471141	chr8	19796299	LPL	281	TSS1500	N_Shore	-0.34	0.10	8.48E-05	0.050	6.50
cg06667968	chr7	2107380	MAD1L1	0	Body	Island	-0.14	0.04	2.05E-05	0.031	2.89
cg05422360	chrX	75648455	MAGEE1	0	1stExon	N_Shore	-0.44	0.12	1.72E-07	0.005	7.29
cg14448104	chr20	43933556	MATN4	0	Body	N_Shore	-0.27	0.09	6.02E-05	0.045	5.47
cg01768446	chr16	89982419	MC1R	1866	Intergenic	N_Shelf	-0.13	0.03	4.57E-05	0.041	2.47
cg13645651	chr18	58040213	MC4R	211	TSS1500	OpenSea	-0.48	0.11	1.41E-07	0.005	6.86
cg18233497	chr18	58039968	MC4R	0	1stExon	OpenSea	-0.44	0.11	2.07E-05	0.031	6.66
cg10097150	chr18	58040021	MC4R	19	TSS200	OpenSea	-0.32	0.07	3.06E-05	0.036	4.56
cg17443388	chr13	113632624	MCF2L	0	Body	N_Shore	-0.15	0.03	1.68E-06	0.013	2.40
cg27170816	chr11	12352225	MICALCL	0	Body	OpenSea	-0.17	0.05	2.96E-06	0.015	2.94
cg12673377	chr7	1388271	MICALL2	85722	Intergenic	OpenSea	-0.23	0.07	7.68E-05	0.049	4.40
cg13098800	chr14	101509144	MIR1185-1	168	TSS1500	OpenSea	-0.22	0.05	1.96E-05	0.030	3.35
cg25372296	chr1	98510328	MIR137HG	0	Intergenic	N_Shore	0.33	0.09	3.83E-06	0.017	5.51
cg19112977	chrX	133678716	MIR503HG	0	Body	Island	0.27	0.10	5.42E-06	0.020	6.58
cg11820605	chr11	68519066	MTL5	77	TSS200	S_Shore	0.16	0.06	6.66E-05	0.047	3.58
cg18810376	chr13	29913935	MTUS2	0	Body	Island	-0.15	0.04	6.35E-05	0.046	2.59
cg04519403	chr5	79298951	MTX3	11862	Intergenic	OpenSea	-0.24	0.05	7.51E-06	0.022	3.50
cg22109675	chr1	1297567	MXRA8	409	Intergenic	OpenSea	-0.29	0.08	3.08E-05	0.036	5.62
cg22885000	chr2	192223280	MYO1B	0	Body	OpenSea	-0.14	0.05	3.29E-05	0.037	2.84
cg10660253	chr12	7940546	NANOG	1447	TSS1500	OpenSea	-0.13	0.05	5.84E-05	0.045	4.28
cg18596139	chr1	145389393	NBPF20	0	Intergenic	S_Shelf	-0.30	0.09	5.46E-05	0.044	5.44
cg12091786	chr20	61877942	NKAIN4	0	Body	Island	-0.37	0.10	2.07E-05	0.031	6.04
cg01798261	chr20	21378495	NKX2-4	447	TSS1500	S_Shore	-0.24	0.07	3.17E-05	0.037	4.68
cg00426855	chr4	56496836	NMU	0	Body	OpenSea	-0.19	0.06	4.74E-05	0.042	4.75
cg07997737	chr19	5822630	NRTN	1186	TSS1500	N_Shore	-0.24	0.07	5.62E-05	0.044	4.70
cg06609051	chr2	51231649	NRXN1	0	Body	OpenSea	-0.25	0.07	5.22E-05	0.043	4.54
cg06604589	chr17	800512	NXN	0	Body	OpenSea	0.19	0.07	5.24E-05	0.043	4.54
cg07037019	chr1	159506621	OR10J5	823	TSS1500	OpenSea	-0.16	0.05	7.73E-05	0.049	2.94
cg25279613	chr7	24956523	OSBPL3	0	5'UTR	OpenSea	0.20	0.06	4.95E-06	0.020	4.34
cg25235958	chr17	79797058	P4HB	3974	Intergenic	N_Shelf	-0.14	0.06	1.85E-05	0.029	3.76
cg25670061	chr6	163612896	PACRG	0	Body	OpenSea	-0.14	0.05	7.94E-05	0.049	2.98
cg14524955	chr1	17531661	PADI1	0	5'UTR	OpenSea	-0.15	0.04	1.47E-05	0.028	3.86
cg24536703	chr11	77183438	PAK1	0	5'UTR	N_Shore	-0.32	0.09	2.71E-05	0.034	5.73

cg24036523	chr14	73712256	PAPLN	0	Body	N_Shore	-0.45	0.11	3.05E-05	0.036	7.25
cg15593298	chr3	142681996	PAQR9	0	1stExon	Island	0.16	0.06	5.70E-05	0.044	4.41
cg12436019	chr13	101174364	PCCA	0	Body	OpenSea	-0.33	0.11	5.83E-05	0.045	6.86
cg08985968	chr13	101174096	PCCA	0	Body	OpenSea	-0.24	0.08	1.99E-05	0.030	4.78
cg14009504	chr14	71555837	PCNX1	0	Body	OpenSea	-0.33	0.09	2.10E-05	0.031	5.64
cg21688288	chr21	44181258	PDE9A	0	Body	OpenSea	-0.17	0.04	1.97E-05	0.030	3.08
cg22139555	chr8	17434123	PDGFRL	0	TSS1500	N_Shore	0.16	0.04	1.11E-05	0.026	2.80
cg16720405	chr3	122790178	PDIA5	0	Body	S_Shelf	-0.17	0.05	5.17E-05	0.043	3.08
cg24384195	chr7	106505370	PIK3CG	351	TSS1500	N_Shelf	0.12	0.04	4.64E-05	0.042	2.46
cg02898019	chr4	128800215	PLK4	1799	Intergenic	N_Shore	-0.37	0.09	4.04E-05	0.040	6.01
cg25697727	chr19	291986	PLPP2	550	TSS1500	Island	-0.37	0.12	8.33E-05	0.050	7.67
cg00443797	chr10	122349232	PLPP4	0	3'UTR	OpenSea	-0.27	0.07	2.47E-07	0.005	4.81
cg06196145	chr11	818752	PNPLA2	147	TSS200	N_Shore	-0.24	0.06	1.42E-05	0.028	5.07
cg06430061	chr5	75013592	POC5	278	TSS1500	S_Shore	-0.39	0.09	1.88E-06	0.013	6.54
cg25320496	chr2	130876007	POTEF	0	Body	N_Shelf	-0.24	0.07	1.99E-05	0.030	4.16
cg23350744	chr5	122422098	PRDM6	2741	Intergenic	N_Shelf	-0.13	0.04	1.95E-05	0.030	2.47
cg18642720	chr7	142493865	PRSS3P2	11465	Intergenic	N_Shore	-0.17	0.05	3.16E-05	0.037	5.28
cg14774290	chr8	141934252	PTK2	0	5'UTR	OpenSea	-0.20	0.06	4.39E-05	0.041	4.80
cg13551137	chr14	88946380	PTPN21	0	Body	Island	-0.19	0.06	8.57E-05	0.050	3.85
cg11688093	chr13	111178359	RAB20	0	Body	OpenSea	-0.14	0.06	5.56E-05	0.044	4.81
cg01430588	chr17	56769767	RAD51C	194	TSS1500	N_Shore	-0.32	0.09	6.04E-05	0.045	5.75
cg22273487	chr20	32580931	RALY	525	TSS1500	Island	0.19	0.05	1.60E-05	0.028	3.13
cg05740325	chr12	113537933	RASAL1	0	Body	N_Shelf	-0.14	0.04	4.26E-05	0.040	3.06
cg09152136	chr1	178464723	RASAL2	16074	Intergenic	OpenSea	-0.12	0.04	2.40E-05	0.033	3.94
cg20936321	chr1	8408752	RERE	3710	Intergenic	OpenSea	-0.20	0.05	1.23E-05	0.027	3.23
cg02105458	chr15	93653048	RGMA	20604	Intergenic	Island	-0.36	0.09	2.03E-05	0.031	5.45
cg22343083	chr8	54786401	RGS20	0	Body	N_Shelf	-0.30	0.07	1.92E-06	0.013	5.15
cg24186774	chr14	63757917	RHOJ	0	3'UTR	OpenSea	-0.17	0.05	6.90E-05	0.048	3.30
cg05488279	chr17	1819770	RPA1	16921	Intergenic	OpenSea	-0.17	0.05	1.54E-05	0.028	3.20
cg22758104	chr17	50465	RPH3AL	11713	Intergenic	OpenSea	-0.23	0.06	4.62E-07	0.006	3.65
cg02623471	chr8	27850313	SCARA5	0	TSS200	OpenSea	0.13	0.04	6.59E-05	0.047	2.75
cg19904058	chr10	135279010	SCART1	0	Body	Island	-0.52	0.13	1.45E-05	0.028	8.69
cg25965776	chr4	53823843	SCFD2	0	Body	OpenSea	-0.14	0.04	1.79E-06	0.013	3.38
cg16733643	chr1	41575522	SCMH1	0	Body	OpenSea	-0.42	0.11	5.00E-05	0.043	7.76
cg01474424	chr5	155754231	SGCD	0	1stExon	OpenSea	-0.38	0.12	4.26E-05	0.040	7.67
cg17588491	chr22	25198892	SGSM1	3242	Intergenic	N_Shelf	-0.17	0.04	4.41E-05	0.041	2.86
cg25799433	chr20	1600623	SIRPB1	0	5'UTR	OpenSea	0.13	0.04	7.15E-06	0.022	3.71
cg13971030	chr11	35366721	SLC1A2	0	Body	OpenSea	-0.49	0.14	9.19E-06	0.024	9.26
cg02043000	chr5	1228081	SLC6A18	0	Body	S_Shore	-0.32	0.12	6.41E-05	0.046	7.47
cg03723715	chr19	47937352	SLC8A2	0	Body	S_Shore	-0.18	0.04	1.35E-07	0.005	2.71
cg17567562	chr3	47687980	SMARCC1	0	Body	OpenSea	-0.48	0.11	1.18E-05	0.027	10.62
cg12442902	chr4	7427446	SORCS2	0	Body	OpenSea	-0.32	0.11	5.53E-05	0.044	6.60
cg06481800	chr13	112763963	SOX1	37942	Intergenic	S_Shelf	0.27	0.08	4.33E-05	0.041	5.05
cg10298859	chr13	112883993	SPACA7	146656	Intergenic	OpenSea	-0.20	0.04	2.26E-07	0.005	2.85
cg15472797	chr10	135243109	SPRN	4987	Intergenic	N_Shelf	-0.40	0.12	9.72E-06	0.025	7.56

cg08814772	chrX	153046246	SRPK3	0	TSS1500	N_Shore	-0.16	0.05	6.76E-05	0.047	3.29
cg19701531	chr7	76026560	SSC4D	0	Body	N_Shore	-0.16	0.04	1.04E-05	0.025	2.95
cg21224380	chr12	109240533	SSH1	0	Body	OpenSea	-0.23	0.06	1.81E-07	0.005	7.66
cg25031061	chr1	36809495	STK40	0	Body	S_Shore	-0.17	0.04	4.24E-07	0.006	3.09
cg18640660	chr15	74495657	STRA6	0	5'UTR	OpenSea	-0.14	0.04	3.36E-06	0.015	2.48
cg20663219	chr7	73130521	STX1A	0	Body	N_Shelf	-0.18	0.05	5.62E-06	0.021	4.39
cg26888941	chr6	33398522	SYNGAP1	0	Body	S_Shelf	0.14	0.04	5.77E-08	0.005	2.71
cg00987370	chr4	141680259	TBC1D9	2787	Intergenic	S_Shelf	-0.17	0.05	8.47E-05	0.050	3.27
cg05402891	chr14	96152661	TCL1B	0	TSS200	OpenSea	-0.47	0.12	2.19E-05	0.031	7.58
cg19216791	chr19	5568216	TINCR	210	TSS1500	S_Shore	-0.25	0.07	4.13E-05	0.040	4.18
cg18180056	chr12	130344375	TMEM132D	0	Body	S_Shore	-0.18	0.05	9.13E-06	0.024	3.09
cg05825596	chr2	62756405	TMEM17	0	Intergenic	OpenSea	-0.16	0.05	3.87E-05	0.039	4.09
cg26528484	chr2	103387474	TMEM182	0	Body	OpenSea	-0.15	0.05	2.23E-05	0.031	3.29
cg25936380	chr2	120981591	TMEM185B	606	TSS1500	S_Shore	-0.29	0.07	2.22E-07	0.005	4.38
cg04078118	chr22	40439969	TNRC6B	850	TSS1500	N_Shore	-0.16	0.04	3.11E-07	0.006	2.53
cg26648818	chr16	52581719	TOX3	4	TSS200	S_Shore	-0.18	0.06	7.68E-05	0.049	4.15
cg01824466	chr8	95959531	TP53INP1	0	5'UTR	N_Shore	-0.23	0.08	8.08E-05	0.050	5.68
cg16953966	chrX	106962264	TSC22D3	0	Body	S_Shelf	0.21	0.13	3.71E-05	0.039	7.54
cg03712798	chr1	156548584	TTC24	933	TSS1500	OpenSea	-0.14	0.04	1.88E-05	0.030	2.50
cg12608727	chr1	64711033	UBE2U	1005	Intergenic	OpenSea	-0.21	0.05	1.69E-07	0.005	3.35
cg17295666	chr20	3140861	UBOX5	304	TSS1500	S_Shore	-0.35	0.09	2.68E-06	0.014	5.80
cg08551047	chr15	91473569	UNC45A	0	5'UTR	Island	-0.58	0.14	2.87E-06	0.015	8.64
cg03291024	chr15	91473291	UNC45A	117	TSS200	N_Shore	-0.31	0.08	2.45E-05	0.033	5.22
cg18472881	chr15	91473103	UNC45A	305	TSS1500	N_Shore	-0.30	0.07	2.93E-06	0.015	4.53
cg08267442	chr15	91473386	UNC45A	22	TSS200	Island	-0.26	0.08	3.92E-05	0.039	4.84
cg08939371	chr15	91473059	UNC45A	349	TSS1500	N_Shore	-0.25	0.07	4.87E-05	0.042	4.22
cg13860281	chr7	55638724	VOPP1	0	Body	N_Shore	-0.17	0.05	1.01E-05	0.025	6.11
cg24053992	chr1	119554564	WARS2	19273	Intergenic	S_Shelf	-0.15	0.05	3.77E-05	0.039	3.18
cg20394620	chrX	48541924	WAS	260	TSS1500	OpenSea	-0.20	0.05	3.02E-06	0.015	3.40
cg20858107	chr8	10823238	XKR6	0	Body	OpenSea	-0.45	0.14	3.17E-05	0.037	8.92
cg08674227	chrX	100186995	XKRX	3096	Intergenic	S_Shelf	-0.27	0.07	3.07E-05	0.036	4.50
cg08527226	chr19	36191758	ZBTB32	12070	Intergenic	N_Shore	-0.13	0.04	7.09E-06	0.022	2.65
cg00219626	chr6	28983835	ZNF311	10741	Intergenic	N_Shore	-0.38	0.11	3.60E-05	0.038	6.92
cg18074165	chr3	48315011	ZNF589	0	Intergenic	OpenSea	-0.32	0.08	4.17E-05	0.040	5.06
cg06849504	chr1	91301484	ZNF644	79371	Intergenic	Island	-0.12	0.04	6.54E-05	0.047	3.53
cg20693190	chr19	56741019	ZSCAN5A	0	TSS1500	S_Shore	-0.13	0.04	5.77E-05	0.044	2.72

Appendix 3: List of primer sequences for targeted bisulphite sequencing

In the targeted bisulphite sequencing 42 target CpG sites and five CpG sites for WBC estimation were included. The difference between amplicon and product length is the added common sequence at the 5' end and barcode sequence for the Fluidigm 48.48 Access Array.

HM450K probe	Nearest gene	Forward primer	Reverse primer	Strand	Amplicon length	Product length
cg01893629	ALG10	TAAATAGTTAGGTAGGT TTGAAAAATGAGA	CCCTACCTCTTTAAACA ACCCC	F	186	289
cg08254089	BPI	TTGAGGAGGTTGAGTA GGGTTAGA	ACACAAATTAAACAA ATCTTTCTCTCC	F	183	286
cg21590238	C12orf43	TTTGATAGGTGGAATTT AGGAATTTG	CCTTATAATCTACCTAC CTCAACCCC	F	193	296
cg21153102	CHAC1	GAAGGGTAATYGGAGT TTAGGGG	AAAACAACCTCCTACCT CCTCCC	F	164	267
cg03340215	CPEB1	GGTGGGATATTTTAAGG TGAGGAT	AAACCAATACCCAAA CTCAAAAAC	F	174	277
cg06968859	CTNNA2	TTAAAGTAATATGTGAT TTAGGAGAAGTATGAA	ACACATACTCAAATATC TTTTCTTCAACAA	F	164	267
cg12105860	DENND5B	GTGTATYGTGTGTAGT GTTTGGTT	AATCAATTATTTCTCCC CGACCTCC	R	189	292
cg16840364	GPAT3	RTTGTTTTAGGTTTGGAG GATATAATAATGA	CCAAACTTACATAAAA AATATTCAAAAACA	F	163	266
cg08835688	GRB10	ATAATATTGTTTTATGG TTGGYGGTAAY	GACCTACRCAAACTAC AAAAACCC	F	154	257
cg08349826	GRIN2A	GTGAGAGAATTGGAT AGGAAGGGT	CTAATAATTCCACGCA CCAAATCC	R	195	298
cg23681866	HLA-J	GTAGATTAGGGTTTATT TGGAGGGTA	CCTTCTCTCTTATAAAA AAACCAACC	F	158	261
cg22968966	HS3ST2	TTAGGAGAGGTTATGTT AGGATTGG	CCCTCACAAAACCTACTA TTAATACAAAAATC	F	199	302
cg15127563	ITM2C	GGATGTTGGTTTTGGAT AGGGAATG	AAACCCRCAATAACCC CTACC	F	202	305
cg20020161	ITM2C	GGATTATTTTTATATT TTGTTTTGGGT	AAAATAAAAAACACCTA CCACCATACC	R	191	294
cg22097768	KCNH6	GAGATGAAGTAGAGGG TGGAGAGTA	CTAACATCTACCTACAC CTACACYG	R	172	275
cg17969123	KLHL26	GTTGATTTAGGGTAGGT AAGGGTGT	CTCTCAATAACCCCTAT ACAACTAACAAA	F	172	275
cg05422360	MAGEE1	GTTTGAGGTTTATAGAGA TGGTGG	AAATACAAACCCCTAA TACCCCC	R	188	291
cg01768446	MC1R	AAAGTTGAAGAGAGGT TTTTGTTTATGT	TCCCAATTCACCTCTC TAACATAAAAT	F	197	300
cg25372296	MIR137HG	GTAGGGAGGAAATTTG GTTTTAGTG	CACCTTACCTTTCCTTC TTATAAAACAT	F	151	254
cg04519403	MTX3	TTATATTATGATTTGTT ATAGGTGAGTGGG	TAAAAARTAACTCCAA AAACATTAAYTCCT	F	185	288
cg12091786	NKAIN4	TGGTTTGATTTTTTGTA GGATTAGAGTAG	CAYACCAATAAACCT AAAACCTCC	F	179	282
cg25279613	OSBPL3	GGTGTAGTTTTTTTTAT TTTTGGGAAT	TCCTTCCCTCTCTCTAC CCATATAC	F	199	302
cg24536703	PAK1	GGGTTAATGTTTTTAAA ATTTTTTTTGA	ACTAACCTCCATCCTAT CTCTCCICT	F	190	293
cg24036523	PAPLN	GTTTTAAATTTAAATTA TAGGGTGGGTAGG	TTCTAAAACCTCTTACA AACCCCAAC	F	178	281
cg16720405	PDIA5	TAGGGAGGGAAGGGAT TTATTTAA	CCCCTATCCTCCAAAA TCTAAACT	F	171	274
cg13674411	PLEKHA6	TGGGGAAGGGTTATGG AATAGTTATAA	CTAACACACTCCCTCCA CCAC	F	164	267
cg20684174	PPFIBP2	TGTGGTTGTTTTTATG TATAGAGGG	TTTTAAACCTCAACCTA CCTACACC	F	195	298
cg01430588	RAD51C	GTTTTATTTAATAGGTT AYGGTGGAGGT	ACCATTCTGTTTAAAT CCCCAAAA	R	161	264
cg22273487	RALY	TGGYGTTTAAATTTGTA GGTTGTGG	CTCCRCCCAATCCCAAA TTAAAC	R	151	254
cg22343083	RGS20	TTGTAGAGGAAGAAGA GATTTTAAAGGTAG	AATTACAAATATAAAC CACTACACCCATC	F	150	253
cg22758104	RPH3AL	GGAGGGGTTGAGAAT AGTAATTAG	TACCTAAAACCAACCCC TACCTATA	F	155	258
cg16733643	SCMH1	GGATTTTAGAGATATAG AAGAGAGGAAAAATAGT	AACAATTCTCCAAAACC AACAAAAC	F	168	271
cg17588491	SGSM1	ATTAYRTTTGTGTTGGT TAGGTTGG	TCTCCAACAATATTTCT TCAACRCC	F	198	301
cg13971030	SLC1A2	GGTTGGGTTTGTA AAAAG TAGGAGTAAG	CCCAACAATCCAATCC ATAATAC	R	188	291
cg17567562	SMARCC1	TTTGATTTATGTTTGTA GYGTAAAAGAGG	CTTCTCCAATATACCTC TCCCAATTATTTTC	F	187	290
cg10298859	SPACA7	AGGAAGGATTTTAGAG AGGGTTAGG	TCATAATTCTTCTCAA CAATAAAACC	F	183	286
cg19216791	TINCR	TTTAGAGGYGGAGTGG TTATTTAGT	CACTAYGCTCCAACCTA TCCTTAA	F	159	262

cg25936380	TMEM185B	ATGAGGATTAGATATTT GAAGGTAGGAGT	ACCTATCCCTCTCACTC CAAAATAA	F	176	279
cg01824466	TP53INP1	TGATTTTTTGAAATTGT GAAGGAGT	TTTAAACTCCCAATCT CAACTATAAATCT	R	182	285
cg26657235	ULBP3	TTATTTTGGAAGGGGTT TTAGTTTAG	TCCCCCTCCTACTTACT TTAAAAATACTA	R	169	272
cg08551047	UNC45A	GTTTGAGGTATTGGGGA ATYGG	CCACTACACTCCAACCT AAATAACACA	F	176	279
cg20394620	WAS	GGTTTATGATTATTTT TGTATAGGGAGT	AAACCCACCTTTTAAAT CAACTTAAAT	F	159	262
cg04642300 (WBC)	ARMC2	GTGGTTTTTTTTGTTATT TAAGTTTGG	CCTTCAAATACACTAC CCCTACTAA	F	150	253
cg13917614 (WBC)	CNP	AGTAATTGTAGTTGTGG TCGAGTGA	CCTTCTCCTACCGCAA ATCTCTAA	F	156	259
cg26692003 (WBC)	IQSEC1	TTAAATTTGGGTAGTTT GGTGTGTTG	CCCCACAACATCTCTAA ACCTATTAC	F	159	262
cg11429292 (WBC)	LAG3	TGGAAGATTTGATAAG GTAGTTTTTTG	CTTCAACACTTATCCTT AACCCTCC	R	161	264
cg26872907 (WBC)	PARK2	GGTTTGATGTTATTTT TTTGAATTTATG	TCCTAACACTAACTCAA CTTTTCACAAC	F	153	256

Appendix 4: All CpG sites included in the target regions and breast cancer risk in the Generations Study

The association between all CpG sites included in the target regions (n=274) and breast cancer risk in matched case-control pairs in the Generations Study targeted sequencing data. The association between methylation levels and breast cancer risk was explored using conditional logistic regression adjusted for age, BMI, alcohol consumption, smoking duration and WBC composition. Methylation levels were converted to z-scores for each CpG sites, giving OR per 1 standard deviation (ORperSD) The *P*-values were corrected for using FDR. Pairs with coverage > 30 reads were used in the analysis.

Nearest gene	chr	position	Target	n pairs	Meth cases	Meth ctrls	ORperSD (95 % CI)	<i>P</i> -value	<i>Q</i> -value
ALG10	chr12	34494825	TARGET	401	94.22	93.44	1.06 (0.89 - 1.27)	0.490	0.837
ALG10	chr12	34494852	NA	401	94.95	94.85	1.05 (0.88 - 1.24)	0.610	0.871
ALG10	chr12	34494855	NA	401	76.18	75.55	1.06 (0.89 - 1.27)	0.503	0.837
ALG10	chr12	34494893	NA	401	95.90	95.56	1.09 (0.94 - 1.28)	0.264	0.805
ALG10	chr12	34494904	NA	401	92.58	93.13	0.88 (0.74 - 1.03)	0.107	0.775
ALG10	chr12	34494915	NA	401	81.72	81.43	1.09 (0.92 - 1.30)	0.323	0.828
ALG10	chr12	34494923	NA	401	70.79	69.14	1.06 (0.89 - 1.26)	0.539	0.842
ALG10	chr12	34494938	NA	401	77.21	77.25	0.99 (0.83 - 1.17)	0.876	0.958
ALG10	chr12	34494942	NA	401	53.64	53.21	1.09 (0.92 - 1.29)	0.338	0.831
ALG10	chr12	34494955	NA	401	83.08	83.11	1.01 (0.85 - 1.18)	0.941	0.972
BPI	chr20	36933163	NA	328	20.11	18.04	1.07 (0.90 - 1.27)	0.447	0.831
BPI	chr20	36933189	TARGET	328	41.91	38.08	1.10 (0.92 - 1.32)	0.296	0.826
C12orf43	chr12	121454808	NA	411	33.14	35.85	0.77 (0.64 - 0.93)	0.007	0.513
C12orf43	chr12	121454837	TARGET	411	81.19	80.01	0.93 (0.77 - 1.12)	0.425	0.831
C12orf43	chr12	121454847	NA	411	88.70	87.93	1.09 (0.90 - 1.31)	0.380	0.831
CHAC1	chr15	41252097	NA	418	55.81	55.69	1.00 (0.84 - 1.19)	0.991	0.991
CHAC1	chr15	41252147	TARGET	418	39.42	39.90	1.01 (0.85 - 1.20)	0.906	0.959
CHAC1	chr15	41252179	NA	418	22.74	22.90	0.98 (0.82 - 1.18)	0.863	0.957
CHAC1	chr15	41252199	NA	418	10.81	10.53	1.05 (0.88 - 1.24)	0.596	0.871
CHAC1	chr15	41252202	SNP	418	5.39	5.39	1.02 (0.87 - 1.20)	0.793	0.932
CHAC1	chr15	41252213	NA	418	13.90	13.38	1.11 (0.93 - 1.32)	0.261	0.803
CHAC1	chr15	41252223	NA	418	11.36	11.24	0.99 (0.84 - 1.18)	0.948	0.972
CPEB1	chr15	83315615	TARGET	321	25.87	24.77	0.99 (0.83 - 1.19)	0.954	0.973
CPEB1	chr15	83315627	NA	321	17.75	18.55	0.96 (0.81 - 1.14)	0.665	0.897
CPEB1	chr15	83315635	NA	321	20.38	22.32	0.92 (0.76 - 1.10)	0.359	0.831
CPEB1	chr15	83315639	NA	321	17.10	17.02	1.01 (0.85 - 1.20)	0.916	0.959
CPEB1	chr15	83315647	NA	321	31.22	31.05	0.94 (0.77 - 1.14)	0.533	0.842
CPEB1	chr15	83315654	NA	320	38.49	41.64	0.88 (0.74 - 1.05)	0.147	0.775
CPEB1	chr15	83315659	NA	321	21.68	28.24	0.78 (0.66 - 0.94)	0.008	0.513
CPEB1	chr15	83315675	NA	321	46.25	46.86	0.98 (0.82 - 1.16)	0.791	0.932
CPEB1	chr15	83315683	NA	321	76.78	79.51	0.90 (0.76 - 1.08)	0.255	0.803
CPEB1	chr15	83315698	NA	321	66.65	66.94	0.94 (0.79 - 1.12)	0.495	0.837
CTNNA2	chr2	80724117	NA	360	90.90	91.68	0.96 (0.81 - 1.13)	0.606	0.871
CTNNA2	chr2	80724122	NA	360	91.77	91.59	1.02 (0.86 - 1.20)	0.852	0.957
CTNNA2	chr2	80724209	TARGET	360	74.53	78.90	0.80 (0.67 - 0.95)	0.012	0.562
DENND5B	chr12	31742762	NA	151	98.23	98.18	1.14 (0.82 - 1.59)	0.436	0.831
DENND5B	chr12	31742773	NA	151	1.69	3.95	0.85 (0.64 - 1.12)	0.257	0.803

DENND5B	chr12	31742781	NA	151	6.10	4.73	1.11 (0.86 - 1.42)	0.422	0.831
DENND5B	chr12	31742788	NA	151	4.83	3.59	1.09 (0.82 - 1.44)	0.559	0.854
DENND5B	chr12	31742802	TARGET	151	1.84	1.22	1.07 (0.83 - 1.36)	0.611	0.871
DENND5B	chr12	31742809	NA	151	1.18	1.88	0.93 (0.71 - 1.23)	0.607	0.871
DENND5B	chr12	31742819	NA	151	0.60	3.10	0.59 (0.30 - 1.15)	0.122	0.775
DENND5B	chr12	31742832	NA	151	0.51	1.13	0.87 (0.55 - 1.38)	0.566	0.857
DENND5B	chr12	31742835	NA	151	0.57	1.09	0.92 (0.67 - 1.28)	0.632	0.883
DENND5B	chr12	31742838	NA	151	0.34	0.28	1.09 (0.84 - 1.42)	0.521	0.842
DENND5B	chr12	31742857	NA	151	0.99	1.01	1.05 (0.83 - 1.34)	0.685	0.906
DENND5B	chr12	31742864	NA	151	0.92	0.40	1.73 (0.89 - 3.36)	0.107	0.775
DENND5B	chr12	31742883	NA	150	3.67	1.32	1.37 (0.99 - 1.90)	0.060	0.775
DENND5B	chr12	31742892	NA	151	0.44	0.43	1.01 (0.80 - 1.27)	0.950	0.972
DENND5B	chr12	31742903	NA	150	0.21	0.18	1.08 (0.82 - 1.41)	0.588	0.871
DENND5B	chr12	31742926	NA	151	14.01	12.82	1.16 (0.90 - 1.49)	0.258	0.803
GPAT3	chr4	84539569	TARGET	85	84.89	90.27	0.81 (0.55 - 1.20)	0.297	0.826
GRB10	chr7	50849910	NA	409	72.01	71.18	1.17 (0.94 - 1.45)	0.163	0.775
GRB10	chr7	50849916	NA	409	69.42	68.66	1.18 (0.96 - 1.46)	0.112	0.775
GRB10	chr7	50849923	NA	409	79.01	77.53	1.19 (0.98 - 1.44)	0.081	0.775
GRB10	chr7	50849931	TARGET	409	80.43	78.07	1.23 (1.02 - 1.49)	0.033	0.764
GRB10	chr7	50849941	NA	409	82.66	80.83	1.20 (0.98 - 1.47)	0.082	0.775
GRB10	chr7	50849947	NA	409	82.96	82.06	1.07 (0.87 - 1.30)	0.528	0.842
GRB10	chr7	50849949	NA	409	83.29	81.82	1.14 (0.93 - 1.39)	0.218	0.782
GRB10	chr7	50849956	NA	409	82.62	80.86	1.17 (0.95 - 1.44)	0.150	0.775
GRB10	chr7	50849968	NA	409	83.12	80.72	1.24 (1.02 - 1.53)	0.035	0.764
GRB10	chr7	50849970	NA	409	82.92	81.91	1.08 (0.88 - 1.33)	0.451	0.831
GRB10	chr7	50849972	NA	409	82.65	82.05	1.03 (0.85 - 1.26)	0.742	0.927
GRB10	chr7	50849986	NA	409	81.50	79.66	1.18 (0.97 - 1.44)	0.098	0.775
GRB10	chr7	50849997	NA	409	79.99	79.08	1.17 (0.96 - 1.43)	0.120	0.775
GRB10	chr7	50850008	NA	409	82.25	81.03	1.10 (0.90 - 1.35)	0.366	0.831
GRB10	chr7	50850011	NA	409	83.06	81.25	1.19 (0.97 - 1.47)	0.099	0.775
GRB10	chr7	50850013	NA	409	82.71	81.02	1.19 (0.97 - 1.46)	0.095	0.775
GRB10	chr7	50850017	NA	409	83.16	81.07	1.25 (1.02 - 1.53)	0.033	0.764
GRB10	chr7	50850035	NA	409	80.01	79.33	1.13 (0.94 - 1.35)	0.201	0.775
GRB10	chr7	50850042	NA	409	99.89	99.87	1.07 (0.91 - 1.25)	0.437	0.831
GRIN2A	chr16	10346379	NA	413	99.76	99.78	0.91 (0.75 - 1.09)	0.291	0.826
GRIN2A	chr16	10346404	TARGET	413	92.74	92.99	0.90 (0.77 - 1.05)	0.190	0.775
GRIN2A	chr16	10346414	NA	413	92.44	92.19	1.08 (0.90 - 1.30)	0.418	0.831
GRIN2A	chr16	10346423	NA	413	93.72	93.60	1.03 (0.88 - 1.20)	0.723	0.913
GRIN2A	chr16	10346433	NA	413	69.89	71.07	0.90 (0.77 - 1.05)	0.176	0.775
GRIN2A	chr16	10346473	NA	413	96.34	96.38	0.99 (0.83 - 1.19)	0.909	0.959
GRIN2A	chr16	10346491	SNP	413	74.54	77.08	0.88 (0.75 - 1.03)	0.109	0.775
GRIN2A	chr16	10346511	NA	413	79.45	80.30	0.97 (0.81 - 1.16)	0.702	0.908
HLA-J	chr6	29895171	NA	402	19.49	17.75	1.09 (0.91 - 1.29)	0.355	0.831
HLA-J	chr6	29895175	TARGET	402	33.29	31.26	1.07 (0.91 - 1.27)	0.414	0.831
HLA-J	chr6	29895187	NA	402	26.04	25.48	1.02 (0.86 - 1.20)	0.850	0.957

HLA-J	chr6	29895204	NA	402	19.98	19.04	1.05 (0.89 - 1.25)	0.560	0.854
HLA-J	chr6	29895215	SNP	402	4.31	4.16	1.12 (0.94 - 1.34)	0.221	0.782
HLA-J	chr6	29895227	SNP	402	14.21	13.21	1.08 (0.9 - 1.28)	0.409	0.831
HLA-J	chr6	29895248	NA	402	14.04	15.04	1.01 (0.85 - 1.20)	0.914	0.959
HLA-J	chr6	29895260	NA	402	30.23	30.21	1.02 (0.86 - 1.20)	0.861	0.957
HS3ST2	chr16	22959799	NA	406	9.43	8.47	1.10 (0.91 - 1.31)	0.322	0.828
HS3ST2	chr16	22959808	NA	406	13.45	13.68	1.01 (0.85 - 1.20)	0.876	0.958
HS3ST2	chr16	22959819	NA	405	6.75	5.51	1.07 (0.91 - 1.26)	0.408	0.831
HS3ST2	chr16	22959821	NA	405	8.03	8.64	0.91 (0.77 - 1.07)	0.257	0.803
HS3ST2	chr16	22959823	NA	406	13.23	13.18	1.03 (0.87 - 1.22)	0.726	0.913
HS3ST2	chr16	22959833	NA	406	19.83	19.86	1.04 (0.89 - 1.22)	0.632	0.883
HS3ST2	chr16	22959852	NA	406	16.30	16.42	1.04 (0.87 - 1.25)	0.654	0.894
HS3ST2	chr16	22959868	NA	406	19.71	19.67	1.09 (0.92 - 1.30)	0.312	0.828
HS3ST2	chr16	22959875	TARGET	406	15.85	16.34	0.99 (0.84 - 1.18)	0.927	0.964
HS3ST2	chr16	22959925	NA	406	19.58	21.38	0.98 (0.83 - 1.15)	0.774	0.932
HS3ST2	chr16	22959941	NA	406	26.30	27.44	1.01 (0.85 - 1.19)	0.913	0.959
ITM2C_1	chr2	231729362	NA	370	2.42	2.68	0.92 (0.77 - 1.09)	0.328	0.828
ITM2C_1	chr2	231729378	NA	370	2.73	3.16	0.92 (0.77 - 1.09)	0.324	0.828
ITM2C_1	chr2	231729385	NA	370	1.21	1.02	1.06 (0.90 - 1.24)	0.474	0.837
ITM2C_1	chr2	231729391	NA	370	1.08	0.76	1.10 (0.93 - 1.30)	0.273	0.817
ITM2C_1	chr2	231729395	NA	370	1.33	0.81	1.09 (0.88 - 1.36)	0.428	0.831
ITM2C_1	chr2	231729400	NA	370	1.24	0.93	1.03 (0.89 - 1.20)	0.680	0.905
ITM2C_1	chr2	231729420	NA	370	22.14	20.99	0.96 (0.81 - 1.14)	0.668	0.897
ITM2C_1	chr2	231729444	NA	370	45.19	39.71	1.17 (0.95 - 1.45)	0.140	0.775
ITM2C_1	chr2	231729448	SNP	370	19.67	19.74	0.94 (0.79 - 1.12)	0.504	0.837
ITM2C_1	chr2	231729464	NA	370	7.59	8.16	0.94 (0.80 - 1.12)	0.511	0.842
ITM2C_1	chr2	231729468	NA	370	8.11	9.30	0.91 (0.76 - 1.09)	0.317	0.828
ITM2C_1	chr2	231729472	NA	370	7.51	8.38	0.94 (0.79 - 1.12)	0.501	0.837
ITM2C_1	chr2	231729487	TARGET	370	0.28	0.29	0.95 (0.81 - 1.11)	0.526	0.842
ITM2C_1	chr2	231729492	NA	370	0.34	0.35	1.03 (0.88 - 1.20)	0.725	0.913
ITM2C_1	chr2	231729503	NA	370	0.51	0.33	1.32 (0.96 - 1.82)	0.090	0.775
ITM2C_1	chr2	231729508	NA	370	0.34	0.36	0.87 (0.71 - 1.07)	0.194	0.775
ITM2C_1	chr2	231729511	NA	370	0.37	0.44	0.90 (0.72 - 1.13)	0.369	0.831
ITM2C_1	chr2	231729513	NA	370	0.37	0.35	1.05 (0.90 - 1.23)	0.528	0.842
ITM2C_1	chr2	231729517	NA	370	0.47	0.51	1.02 (0.87 - 1.19)	0.782	0.932
ITM2C_1	chr2	231729532	NA	370	35.46	35.16	1.16 (0.97 - 1.40)	0.105	0.775
ITM2C_2	chr2	231732535	NA	327	74.29	75.42	0.91 (0.77 - 1.08)	0.305	0.827
ITM2C_2	chr2	231732670	TARGET	327	88.58	87.43	1.04 (0.87 - 1.24)	0.671	0.897
KCNH6	chr17	61615826	NA	407	96.73	96.20	1.15 (0.97 - 1.38)	0.116	0.775
KCNH6	chr17	61615828	NA	407	7.06	7.76	0.91 (0.76 - 1.09)	0.304	0.827
KCNH6	chr17	61615858	NA	407	8.84	8.98	0.97 (0.82 - 1.14)	0.688	0.906
KCNH6	chr17	61615861	NA	407	5.62	6.08	0.84 (0.70 - 1.01)	0.059	0.775
KCNH6	chr17	61615864	NA	407	8.40	8.52	0.91 (0.76 - 1.09)	0.300	0.827
KCNH6	chr17	61615880	NA	407	15.34	16.60	0.82 (0.69 - 0.98)	0.030	0.764
KCNH6	chr17	61615882	NA	407	11.94	12.19	0.92 (0.78 - 1.09)	0.325	0.828

KCNH6	chr17	61615884	NA	407	14.73	16.83	0.85 (0.71 - 1.01)	0.060	0.775
KCNH6	chr17	61615891	NA	407	9.37	9.03	0.88 (0.75 - 1.05)	0.152	0.775
KCNH6	chr17	61615912	NA	407	2.68	2.65	0.99 (0.84 - 1.16)	0.856	0.957
KCNH6	chr17	61615914	TARGET	407	7.03	7.73	0.97 (0.81 - 1.17)	0.781	0.932
KCNH6	chr17	61615917	NA	407	7.92	7.06	1.06 (0.90 - 1.26)	0.487	0.837
KCNH6	chr17	61615929	NA	407	2.53	2.23	1.15 (0.94 - 1.39)	0.173	0.775
KCNH6	chr17	61615942	NA	407	2.37	1.76	1.09 (0.92 - 1.28)	0.314	0.828
KCNH6	chr17	61615945	NA	407	4.21	3.58	1.04 (0.88 - 1.24)	0.648	0.891
KCNH6	chr17	61615948	NA	407	1.42	1.54	0.94 (0.80 - 1.10)	0.429	0.831
KLHL26	chr19	18745966	NA	405	70.81	72.10	0.92 (0.77 - 1.11)	0.389	0.831
KLHL26	chr19	18745971	TARGET	405	44.51	43.69	0.98 (0.82 - 1.18)	0.867	0.958
KLHL26	chr19	18745988	NA	405	69.90	69.70	0.97 (0.82 - 1.16)	0.756	0.932
KLHL26	chr19	18745990	NA	405	70.56	71.62	0.90 (0.75 - 1.07)	0.238	0.803
MAGEE1	chrX	75648426	NA	416	47.99	45.48	1.45 (1.18 - 1.78)	4.13E-04	0.131
MAGEE1	chrX	75648435	NA	416	27.09	26.64	1.01 (0.84 - 1.20)	0.944	0.972
MAGEE1	chrX	75648456	TARGET	416	26.08	26.46	1.00 (0.84 - 1.19)	0.978	0.984
MAGEE1	chrX	75648474	NA	416	36.49	34.44	1.22 (1.02 - 1.46)	0.031	0.764
MAGEE1	chrX	75648501	NA	416	39.13	37.85	1.19 (0.99 - 1.42)	0.058	0.775
MAGEE1	chrX	75648525	NA	416	31.12	29.97	1.21 (1.01 - 1.45)	0.042	0.775
MAGEE1	chrX	75648540	NA	416	34.22	33.77	0.99 (0.82 - 1.18)	0.888	0.959
MAGEE1	chrX	75648566	NA	416	42.98	41.63	1.16 (0.96 - 1.40)	0.122	0.775
MC1R	chr16	89982419	TARGET	385	91.61	92.87	0.90 (0.77 - 1.07)	0.238	0.803
MC1R	chr16	89982437	SNP	385	90.56	90.49	1.02 (0.88 - 1.19)	0.800	0.932
MC1R	chr16	89982456	NA	385	90.88	91.27	0.97 (0.82 - 1.14)	0.707	0.911
MC1R	chr16	89982462	NA	385	94.39	94.35	1.01 (0.86 - 1.20)	0.884	0.959
MC1R	chr16	89982491	NA	385	88.74	85.61	1.13 (0.96 - 1.34)	0.150	0.775
MC1R	chr16	89982504	NA	385	93.03	93.58	1.01 (0.85 - 1.19)	0.951	0.972
MC1R	chr16	89982507	NA	385	93.99	93.48	1.03 (0.88 - 1.20)	0.711	0.913
MC1R	chr16	89982518	NA	385	90.55	89.39	1.09 (0.93 - 1.28)	0.289	0.826
MC1R	chr16	89982532	NA	385	90.68	92.09	0.96 (0.82 - 1.13)	0.615	0.871
MC1R	chr16	89982545	SNP	385	92.00	92.93	0.93 (0.78 - 1.11)	0.421	0.831
MC1R	chr16	89982554	NA	385	95.25	95.20	0.98 (0.84 - 1.16)	0.831	0.954
MIR137HG	chr1	98510328	TARGET	330	19.52	18.30	1.07 (0.91 - 1.26)	0.390	0.831
MIR137HG	chr1	98510330	NA	330	16.76	15.36	1.08 (0.91 - 1.28)	0.398	0.831
MIR137HG	chr1	98510350	NA	330	19.34	18.76	1.02 (0.85 - 1.21)	0.850	0.957
MIR137HG	chr1	98510359	NA	330	16.42	14.41	1.11 (0.94 - 1.31)	0.211	0.776
MIR137HG	chr1	98510398	NA	330	21.04	21.13	0.97 (0.81 - 1.16)	0.723	0.913
MIR137HG	chr1	98510411	NA	330	11.05	10.18	1.07 (0.91 - 1.26)	0.396	0.831
MTX3	chr5	79298951	TARGET	226	84.16	88.38	0.84 (0.68 - 1.05)	0.128	0.775
NKAIN4	chr20	61877855	NA	275	72.13	67.23	1.12 (0.93 - 1.34)	0.253	0.803
NKAIN4	chr20	61877865	NA	275	57.91	54.14	1.12 (0.93 - 1.36)	0.232	0.803
NKAIN4	chr20	61877875	SNP	275	57.37	54.09	1.13 (0.93 - 1.36)	0.222	0.782
NKAIN4	chr20	61877880	NA	275	76.39	75.02	1.08 (0.90 - 1.29)	0.396	0.831
NKAIN4	chr20	61877885	NA	275	56.31	58.25	0.93 (0.77 - 1.13)	0.479	0.837
NKAIN4	chr20	61877900	NA	275	66.83	70.07	0.88 (0.73 - 1.07)	0.206	0.775

NKAIN4	chr20	61877932	NA	275	89.62	89.31	1.06 (0.88 - 1.28)	0.538	0.842
NKAIN4	chr20	61877939	SNP	275	65.39	60.83	1.13 (0.94 - 1.37)	0.200	0.775
NKAIN4	chr20	61877942	TARGET	275	79.62	83.38	0.93 (0.77 - 1.11)	0.422	0.831
NKAIN4	chr20	61877956	SNP	275	60.26	56.43	1.16 (0.95 - 1.41)	0.140	0.775
OSBPL3	chr7	24956523	TARGET	62	73.94	70.70	1.08 (0.65 - 1.80)	0.763	0.932
PAK1	chr11	77183438	TARGET	417	77.23	78.58	0.90 (0.76 - 1.05)	0.187	0.775
PAPLN	chr14	73712152	NA	398	56.53	56.28	0.98 (0.83 - 1.16)	0.823	0.948
PAPLN	chr14	73712178	NA	398	74.74	73.51	1.07 (0.90 - 1.26)	0.446	0.831
PAPLN	chr14	73712203	NA	399	56.33	54.80	1.13 (0.94 - 1.35)	0.208	0.775
PAPLN	chr14	73712214	NA	399	38.74	38.50	0.94 (0.80 - 1.11)	0.491	0.837
PAPLN	chr14	73712256	TARGET	399	51.71	51.52	0.96 (0.81 - 1.14)	0.665	0.897
PDIA5	chr3	122790152	NA	352	67.28	67.58	0.96 (0.82 - 1.13)	0.649	0.891
PDIA5	chr3	122790178	TARGET	352	80.91	82.46	0.88 (0.74 - 1.04)	0.121	0.775
PDIA5	chr3	122790209	NA	352	59.37	60.78	0.91 (0.77 - 1.08)	0.292	0.826
PLEKHA6	chr1	204232640	NA	400	77.58	77.25	1.03 (0.86 - 1.23)	0.768	0.932
PLEKHA6	chr1	204232677	TARGET	400	89.27	88.06	1.14 (0.96 - 1.37)	0.135	0.775
PLEKHA6	chr1	204232697	NA	400	80.11	79.46	1.07 (0.89 - 1.28)	0.472	0.837
PPFIBP2	chr11	7541255	TARGET	395	92.39	93.08	0.91 (0.76 - 1.08)	0.260	0.803
PPFIBP2	chr11	7541270	SNP	395	79.98	81.92	0.93 (0.79 - 1.09)	0.370	0.831
PPFIBP2	chr11	7541288	NA	395	94.06	93.88	0.93 (0.78 - 1.10)	0.390	0.831
PPFIBP2	chr11	7541332	NA	395	94.86	95.58	0.91 (0.74 - 1.11)	0.352	0.831
PPFIBP2	chr11	7541349	NA	395	77.34	78.03	0.94 (0.80 - 1.10)	0.422	0.831
PPFIBP2	chr11	7541354	NA	395	96.47	96.67	0.90 (0.76 - 1.06)	0.204	0.775
PPFIBP2	chr11	7541362	NA	395	95.99	96.83	0.83 (0.67 - 1.01)	0.064	0.775
PPFIBP2	chr11	7541384	NA	395	91.88	91.44	1.05 (0.88 - 1.24)	0.594	0.871
RAD51C	chr17	56769696	NA	90	63.55	65.65	0.86 (0.62 - 1.20)	0.379	0.831
RAD51C	chr17	56769718	NA	90	96.47	94.04	1.15 (0.84 - 1.56)	0.393	0.831
RAD51C	chr17	56769746	NA	89	89.63	83.64	1.39 (0.94 - 2.06)	0.101	0.775
RAD51C	chr17	56769760	NA	89	83.43	84.21	0.97 (0.71 - 1.33)	0.860	0.957
RAD51C	chr17	56769767	TARGET	89	69.42	65.70	1.11 (0.81 - 1.51)	0.525	0.842
RAD51C	chr17	56769830	NA	90	97.84	98.40	0.97 (0.71 - 1.32)	0.855	0.957
RALY	chr20	32580828	NA	385	40.04	40.88	0.90 (0.75 - 1.08)	0.250	0.803
RALY	chr20	32580850	NA	385	0.41	0.32	1.14 (0.95 - 1.35)	0.151	0.775
RALY	chr20	32580857	NA	384	0.32	0.31	1.04 (0.90 - 1.21)	0.608	0.871
RALY	chr20	32580882	NA	384	0.32	0.31	1.04 (0.85 - 1.28)	0.701	0.908
RALY	chr20	32580895	NA	384	0.36	0.32	0.88 (0.63 - 1.24)	0.472	0.837
RALY	chr20	32580923	NA	384	0.30	0.25	0.84 (0.58 - 1.20)	0.335	0.831
RALY	chr20	32580932	TARGET	384	0.26	0.26	0.98 (0.82 - 1.16)	0.787	0.932
RALY	chr20	32580948	NA	384	0.70	0.58	1.06 (0.91 - 1.25)	0.440	0.831
RALY	chr20	32580971	NA	385	19.56	19.88	0.94 (0.77 - 1.13)	0.491	0.837
RGS20	chr8	54786401	TARGET	417	76.85	77.77	0.86 (0.72 - 1.02)	0.082	0.775
RGS20	chr8	54786475	SNP	417	83.39	83.39	0.97 (0.83 - 1.14)	0.716	0.913
RPH3AL	chr17	50438	NA	417	83.14	83.79	0.95 (0.80 - 1.13)	0.559	0.854
RPH3AL	chr17	50465	TARGET	417	85.66	86.65	0.83 (0.69 - 0.99)	0.036	0.764
SCMH1	chr1	41575522	TARGET	309	54.02	54.01	1.01 (0.85 - 1.20)	0.875	0.958

SGSM1	chr22	25198798	SNP	394	66.16	65.96	0.93 (0.78 - 1.11)	0.410	0.831
SGSM1	chr22	25198822	SNP	394	95.30	95.48	1.01 (0.85 - 1.21)	0.894	0.959
SGSM1	chr22	25198840	NA	394	95.51	95.91	0.96 (0.78 - 1.17)	0.668	0.897
SGSM1	chr22	25198849	NA	394	94.28	94.70	0.88 (0.74 - 1.05)	0.165	0.775
SGSM1	chr22	25198854	NA	394	86.62	85.72	1.15 (0.97 - 1.37)	0.106	0.775
SGSM1	chr22	25198892	TARGET	394	83.44	82.77	1.00 (0.85 - 1.17)	0.970	0.980
SGSM1	chr22	25198894	NA	394	66.40	65.51	1.05 (0.89 - 1.24)	0.539	0.842
SGSM1	chr22	25198898	NA	394	84.04	83.14	1.11 (0.94 - 1.30)	0.213	0.776
SGSM1	chr22	25198934	NA	394	93.22	92.54	1.11 (0.95 - 1.30)	0.193	0.775
SGSM1	chr22	25198969	NA	394	90.20	90.11	0.98 (0.82 - 1.18)	0.854	0.957
SLC1A2	chr11	35366722	TARGET	374	76.61	75.62	1.07 (0.92 - 1.25)	0.383	0.831
SLC1A2	chr11	35366734	SNP	374	32.98	33.86	1.00 (0.86 - 1.17)	0.965	0.978
SLC1A2	chr11	35366819	NA	374	19.69	19.23	0.98 (0.83 - 1.15)	0.790	0.932
SMARCC1	chr3	47687921	NA	260	81.66	81.97	0.96 (0.80 - 1.16)	0.694	0.907
SMARCC1	chr3	47687939	SNP	257	84.38	83.72	1.12 (0.93 - 1.35)	0.247	0.803
SMARCC1	chr3	47687980	TARGET	260	47.00	51.87	0.84 (0.70 - 1.01)	0.068	0.775
SPACA7	chr13	112883974	NA	355	78.62	76.69	1.10 (0.92 - 1.32)	0.279	0.826
SPACA7	chr13	112883990	NA	355	84.63	84.72	0.99 (0.84 - 1.16)	0.908	0.959
SPACA7	chr13	112883993	TARGET	355	86.73	86.48	0.97 (0.83 - 1.15)	0.751	0.929
SPACA7	chr13	112884003	NA	355	80.40	79.50	1.07 (0.89 - 1.27)	0.482	0.837
SPACA7	chr13	112884032	NA	355	71.23	69.72	1.04 (0.88 - 1.24)	0.611	0.871
SPACA7	chr13	112884039	NA	355	79.13	79.67	1.03 (0.86 - 1.22)	0.750	0.929
SPACA7	chr13	112884075	NA	355	76.10	74.34	1.13 (0.95 - 1.35)	0.161	0.775
SPACA7	chr13	112884080	NA	355	62.78	61.47	1.12 (0.94 - 1.32)	0.203	0.775
SPACA7	chr13	112884102	NA	354	82.14	78.12	1.25 (1.05 - 1.49)	0.012	0.562
TINCR	chr19	5568163	NA	413	41.26	39.71	1.01 (0.85 - 1.20)	0.881	0.959
TINCR	chr19	5568193	NA	413	25.72	25.94	0.94 (0.80 - 1.11)	0.457	0.833
TINCR	chr19	5568216	TARGET	413	47.64	50.08	0.77 (0.64 - 0.93)	0.006	0.513
TINCR	chr19	5568218	NA	413	26.46	28.59	0.75 (0.63 - 0.90)	0.002	0.370
TINCR	chr19	5568220	NA	413	59.11	59.31	0.94 (0.80 - 1.10)	0.426	0.831
TINCR	chr19	5568307	NA	413	99.78	99.76	1.08 (0.92 - 1.28)	0.349	0.831
TMEM185B	chr2	120981515	NA	405	1.27	1.73	0.98 (0.83 - 1.15)	0.792	0.932
TMEM185B	chr2	120981571	NA	407	44.68	44.85	0.89 (0.75 - 1.05)	0.163	0.775
TMEM185B	chr2	120981590	NA	407	68.52	68.73	0.94 (0.79 - 1.11)	0.453	0.831
TMEM185B	chr2	120981592	TARGET	407	52.96	52.67	0.99 (0.85 - 1.16)	0.927	0.964
TMEM185B	chr2	120981595	NA	407	66.86	65.07	1.06 (0.90 - 1.26)	0.494	0.837
TP53INP1	chr8	95959449	NA	260	30.17	31.33	0.95 (0.79 - 1.13)	0.552	0.853
TP53INP1	chr8	95959497	NA	259	19.56	21.98	0.92 (0.76 - 1.12)	0.431	0.831
TP53INP1	chr8	95959532	TARGET	261	27.38	29.55	0.89 (0.74 - 1.08)	0.249	0.803
ULBP3	chr6	150378973	TARGET	370	92.34	92.28	1.02 (0.88 - 1.18)	0.806	0.936
ULBP3	chr6	150378977	NA	370	87.14	88.97	0.90 (0.77 - 1.05)	0.188	0.775
ULBP3	chr6	150379004	NA	370	84.89	84.46	1.05 (0.89 - 1.22)	0.573	0.861
ULBP3	chr6	150379009	NA	370	86.21	84.80	1.13 (0.95 - 1.34)	0.157	0.775
ULBP3	chr6	150379028	NA	369	79.10	79.03	1.05 (0.89 - 1.23)	0.568	0.857
ULBP3	chr6	150379045	NA	369	93.50	94.13	0.93 (0.80 - 1.09)	0.388	0.831

ULBP3	chr6	150379048	NA	369	83.30	85.56	0.94 (0.80 - 1.10)	0.438	0.831
ULBP3	chr6	150379077	NA	370	93.74	94.36	0.96 (0.80 - 1.14)	0.637	0.886
UNC45A	chr15	91473498	NA	358	22.32	22.46	0.98 (0.82 - 1.17)	0.800	0.932
UNC45A	chr15	91473511	NA	358	4.92	4.83	1.02 (0.86 - 1.22)	0.777	0.932
UNC45A	chr15	91473542	NA	358	5.05	4.66	1.07 (0.90 - 1.27)	0.452	0.831
UNC45A	chr15	91473569	TARGET	358	5.42	5.13	1.05 (0.89 - 1.24)	0.589	0.871
UNC45A	chr15	91473573	NA	358	5.28	5.14	1.03 (0.87 - 1.22)	0.695	0.907
UNC45A	chr15	91473586	NA	358	5.21	5.49	1.00 (0.85 - 1.19)	0.963	0.978
UNC45A	chr15	91473591	NA	358	5.24	5.25	1.03 (0.87 - 1.21)	0.765	0.932
UNC45A	chr15	91473617	NA	358	6.40	5.87	1.07 (0.90 - 1.27)	0.462	0.837
UNC45A	chr15	91473625	NA	358	5.67	5.69	1.02 (0.86 - 1.21)	0.813	0.940
WAS	chrX	48541924	TARGET	385	40.56	42.49	0.92 (0.78 - 1.10)	0.368	0.831

Appendix 5. Script for the meta-analysis

```
library(glmnet)
library(survival)

# Load probes and model for MI
load("MI_of_ELEE.RData") #load 'model' and 'probes'

#####
# FIRST PART: MI AND BREAST CANCER RISK ANALYSIS #
#####

#load pd, betas, WBC counts from Houseman method
load("")

pd <- pd[, c("caco", "match", "age", "age_menarche", "age_menopause", "nr_preg",
"bf_dur", "alcohol", "smoking", "bmi", "time")]

#HM450K beta-values with probes as rows and samples as columns, keep only probes used
in the MI model
betas <- betas[probes, rownames(pd)]

#WBC data from the Houseman method, with samples as rows and WBCs as columns
wbc <- wbc[rownames(pd),]

#Exclude samples with missing data
samples.na <- apply(cbind.data.frame(pd$match, pd$age, pd$bmi, pd$smoking, pd$alcohol,
wbc$Bcell, wbc$CD4T, wbc$CD8T, wbc$Mono, wbc$Gran, wbc$NK), 1, function(x) {
any(is.na(x)) })

if(any(samples.na==TRUE)){
  remove <- which(samples.na==TRUE)
  pd <- pd[-remove,]
  wbc <- wbc[-remove,]
  betas<-betas[,-remove]
}

#Make sure there are no NA:s in 450K methylation data
if(any(is.na(betas))){
  samples.na <- apply(betas, 1, function(x){ any(is.na(x)) })
  remove <- which(samples.na==TRUE)
  pd <- pd[-remove,]
  wbc <- wbc[-remove,]
  betas<-betas[,-remove]
}

#Remove unmatched pairs
pd$match <- as.character(pd$match)
keep.id.match <- pd$match[duplicated(pd$match)]
keep <- which(pd$match %in% keep.id.match)
pd <- pd[keep,]
wbc <- wbc[keep,]
betas <- betas[,keep]

identical(rownames(pd), rownames(wbc))
identical(rownames(pd), colnames(betas))

#####
# Association MI and Breast Cancer Risk #
#####
MI <- predict(model$finalModel, newx=t(as.data.frame(betas)), s=model$bestTune$lambda,
type="response", standardize=FALSE)
MI<-MI[,1]
```

```

#Variables
caco <- pd$caco
match <- pd$match
age <- pd$age
alcohol <- pd$alcohol
smoking <- pd$smoking
bmi <- pd$bmi
bcells <- wbc$Bcell
CD4T <- wbc$CD4T
CD8T <- wbc$CD8T
mon <- wbc$Mono
gran <- wbc$Gran
NK <- wbc$NK

#Association MI and Breast Cancer - unadjusted model
fit <- clogit(caco ~ MI + strata(match))
OR <- summary(fit)$conf.int["MI", 1]
l <- summary(fit)$conf.int["MI", 3]
u <- summary(fit)$conf.int["MI", 4]
P.OR <- summary(fit)$coefficients["MI", 5]

#Association MI and Breast Cancer - multivariate model
fit <- clogit(caco ~ MI + age + bmi + smoking + alcohol + bcells + CD4T + CD8T + mon +
gran + NK + strata(match))
OR.multi <- summary(fit)$conf.int["MI", 1]
l.multi <- summary(fit)$conf.int["MI", 3]
u.multi <- summary(fit)$conf.int["MI", 4]
P.OR.multi <- summary(fit)$coefficients["MI", 5]

#Save result
n <- nrow(pd)/2
brca.res <- c(n, OR, l, u, P.OR, OR.multi, l.multi, u.multi, P.OR.multi)
names(brca.res) <- c("n", "OR", "l", "u", "P", "OR_multi", "l_multi", "u_multi",
"P_multi")

#####
# Quartile analysis #
#####
ctrls <- which(pd$caco==0)
Q1 <- summary(MI[ctrls])[2]
M <- summary(MI[ctrls])[3]
Q3 <- summary(MI[ctrls])[5]
MI.cat <- cut(MI, breaks=c(0, Q1, M, Q3, 1000), labels=c("A", "B", "C", "D"),
dig.lab=7)
summary(MI.cat[ctrls])

#### Make sure equal number of controls in each group ####

#Quartile analysis - unadjusted model
fit <- clogit(caco ~ MI.cat + strata(match))
OR <- summary(fit)$coefficients[,2]
ci.B <- exp(confint(fit, parm="MI.catB"))
ci.C <- exp(confint(fit, parm="MI.catC"))
ci.D <- exp(confint(fit, parm="MI.catD"))
l <- c(ci.B[1],ci.C[1],ci.D[1])
u <- c(ci.B[2],ci.C[2],ci.D[2])
p <- summary(fit)$coefficients[,5]
min <- c(min(MI), Q1, M, Q3)
max <- c(Q1, M, Q3, max(MI))
n.ctrls <- c(summary(MI.cat[(which(pd$caco==0))]))
n.cases <- c(summary(MI.cat[(which(pd$caco==1))]))

brca.res.quartiles <- cbind.data.frame(min, max, n.cases, n.ctrls, c("NA",OR),
c("NA",l), c("NA",u), c("NA",p))

```

```

#Quartile analysis - multivariate model
fit <- clogit(caco ~ MI.cat + age + bmi + alcohol + smoking + alcohol + bcells + CD4T +
  CD8T + mon + gran + NK + strata(match))
OR <- summary(fit)$coefficients[c(1:3), 2]
ci.B <- exp(confint(fit, parm="MI.catB"))
ci.C <- exp(confint(fit, parm="MI.catC"))
ci.D <- exp(confint(fit, parm="MI.catD"))
l <- c(ci.B[1],ci.C[1],ci.D[1])
u <- c(ci.B[2],ci.C[2],ci.D[2])
p <- summary(fit)$coefficients[c(1:3),5]

brca.res.quartiles <- cbind.data.frame(brca.res.quartiles, c("NA",OR), c("NA",l),
  c("NA",u), c("NA",p))
colnames(brca.res.quartiles) <- c("min", "max", "n_cases", "n_ctrls", "OR", "l", "u",
  "P", "OR_multi", "l_multi", "u_multi", "P_multi")
rownames(brca.res.quartiles)<-c("Q1", "Q2", "Q3", "Q4")

#####
# Time to diagnosis #
#####
time <- pd$time
time[which(pd$caco==0)] <- NA
time[which(is.na(pd$match)))] <- NA
m <- median(time, na.rm=T)

#Pairs with time to diagnosis < median
keep.match <- pd$match[which(time < m)]
keep.short <- which(pd$match %in% keep.match)

#Association MI and Breast Cancer - unadjusted model
fit <- clogit(caco[keep.short] ~ MI[keep.short] + strata(match[keep.short]))
OR <- summary(fit)$conf.int[1]
l <- summary(fit)$conf.int[3]
u <- summary(fit)$conf.int[4]
P.OR <- summary(fit)$coefficients[5]

#Association MI and Breast Cancer - multivariate model
fit <- clogit(caco[keep.short] ~ MI[keep.short] + age[keep.short] + bmi[keep.short] +
  smoking[keep.short] + alcohol[keep.short] + bcells[keep.short] + CD4T[keep.short] +
  CD8T[keep.short] + mon[keep.short] + gran[keep.short] + NK[keep.short] +
  strata(match[keep.short]))
OR.multi <- summary(fit)$conf.int["MI[keep.short]", 1]
l.multi <- summary(fit)$conf.int["MI[keep.short]", 3]
u.multi <- summary(fit)$conf.int["MI[keep.short]", 4]
P.OR.multi <- summary(fit)$coefficients["MI[keep.short]", 5]

brca.res.time <- c(length(keep.short/2), OR, l, u, P.OR, OR.multi, l.multi, u.multi,
  P.OR.multi)

#Pairs with time to diagnosis >= median
keep.match <- pd$match[which(time >= m)]
keep.long <- which(pd$match %in% keep.match)

#Association MI and Breast Cancer - unadjusted model
fit <- clogit(caco[keep.long] ~ MI[keep.long] + strata(match[keep.long]))
OR <- summary(fit)$conf.int[1]
l <- summary(fit)$conf.int[3]
u <- summary(fit)$conf.int[4]
P.OR <- summary(fit)$coefficients[5]

#Association MI and Breast Cancer - multivariate model
fit <- clogit(caco[keep.long] ~ MI[keep.long] + age[keep.long] + bmi[keep.long] +
  smoking[keep.long] + alcohol[keep.long] + bcells[keep.long] + CD4T[keep.long] +
  CD8T[keep.long] + mon[keep.long] + gran[keep.long] + NK[keep.long] +

```



```

strata(match[keep.long]))
OR.multi <- summary(fit)$conf.int["MI[keep.long]", 1]
l.multi <- summary(fit)$conf.int["MI[keep.long]", 3]
u.multi <- summary(fit)$conf.int["MI[keep.long]", 4]
P.OR.multi <- summary(fit)$coefficients["MI[keep.long]", 5]

brca.res.time <- rbind.data.frame(brca.res.time, c(length(keep.long/2), OR, l, u, P.OR,
  OR.multi, l.multi, u.multi, P.OR.multi))
colnames(brca.res.time) <- c("n", "OR", "l", "u", "P", "OR_multi", "l_multi",
  "u_multi", "P_multi")
rownames(brca.res.time) <- c("short", "long")

# Save results
save(brca.res, brca.res.quantiles, brca.res.time, file="MI_ELEE_PART1.RData") res,
brca.res.quantiles, brca.res.time, file="MI_ELEE_PART1.RData")

#####
# SECOND PART: MI AND ELEE ANALYSIS #
#####

#Calculate reproductive time
age <- pd$age
age_menarche <- pd$age_menarche
age_menopause <- pd$age_menopause
repr_time <- rep(NA, nrow(pd))
prem <- which(is.na(age_menopause))
postm <- which(!(is.na(age_menopause)))
repr_time[prem] <- age[prem] - age_menarche[prem]
repr_time[postm] <- age_menopause[postm] - age_menarche[postm]

#Calculate ELEE
age <- pd$age
nr_preg <- pd$nr_preg
bf_dur <- pd$bf_dur
ELEE <- repr_time - nr_preg - bf_dur

#Exclude samples with missing information of ELEE
if(any(is.na(ELEE))){
  r <- which(is.na(ELEE))
  ELEE <- ELEE[-r]
  pd <- pd[-r,]
  betas <- betas[,-r]
}

identical(rownames(pd), colnames(betas))

#####
# Association MI and ELEE #
#####
MI <- predict(model$finalModel, newx=t(as.data.frame(betas)), s=model$bestTune$lambda,
  type="response", standardize=FALSE)
MI<-MI[,1]
r <- cor(ELEE, MI)
P.corr <- cor.test(ELEE, MI)$p.value
res.ELEE <- c(r, P.corr)
names(res.ELEE) <- c("corr", "P_corr")

# Save results
save(res.ELEE, file="MI_ELEE_PART2.RData")

```

Appendix 6: Primer sequences for the bisulphite pyrosequencing assays

In this study the DNA methylation levels for five assays were analysed using bisulphite sequencing. The primers were designed in PyroMark Assay Design 2.0 Software, and include two PCR primers (forward and reverse) where one is biotinylated at the 5' end (marked by [Bt]), and one sequencing primer.

Assay	Primer	Sequence
LINE-1	Forward	GGATTTTTTGAGTTAGGTGTGGG
	Reverse	CAAAAAATCAAAAAATCCCTTTCC [Bt]
	Sequencing	AGGTGTGGGATATAGT
CTNNA2	Forward	GGATGAAGGAGGAGTATTTGTATTTAT [Bt]
	Reverse	AACACCCTTATACCTCACACAATAAC
	Sequencing	ATCTTTTCTTCAACAAAACCTT
GRB10	Forward	GGGTTTTAGTGGGTATAGTTATTAT
	Reverse	CCCCCTCTCCAAATACTCAAATAAACT [Bt]
	Sequencing	ATGGTTATATAATATTGTTTTATGG
RPH3AL	Forward	AGAGGGGTTTGTGTGATGAA
	Reverse	AACCCTCTCTACATTCCTACCAATAC [Bt]
	Sequencing	GGGTGGTAGAGGGTTAT
TINCR	Forward	AGGAGGAGTGGTTATTTAGTATAGAGTT
	Reverse	CTCCCTAAAACCTCCACCATATA [Bt]
	Sequencing	GAGGTTATTTTAGTTAAGGAGA