

This electronic thesis or dissertation has been downloaded from the King's Research Portal at <https://kclpure.kcl.ac.uk/portal/>



## Epigenetic Determinants of Healthy Ageing and Age-Related Disease Risk

Tsai, Pei-Chien

*Awarding institution:*  
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

### END USER LICENCE AGREEMENT



Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. <https://creativecommons.org/licenses/by-nc-nd/4.0/>

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

### Take down policy

If you believe that this document breaches copyright please contact [librarypure@kcl.ac.uk](mailto:librarypure@kcl.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.

**EPIGENETIC DETERMINANTS OF  
HEALTHY AGEING AND AGE-RELATED  
DISEASE RISK**

**Pei-Chien Tsai**

**Thesis submitted in fulfilment of the requirements for the degree of  
Doctor of Philosophy**

**Department of Twin Research and Genetic Epidemiology  
King's College London**

**2014**



I dedicate this thesis to my loving parents

謹將此論文獻給我的父母

Mr. Tsai Che-Wen and Mrs. Lin Hsiang-Ching

蔡哲文先生以及林香卿女士

For modifying my methylome in the best way with their endless love and support

以無比的愛與支持，成就我人生中每一次的向前邁進

## ABSTRACT

Epigenome-wide association scans (EWAS) of human complex traits are a rapidly growing area of research, in part due to recent advances in technology that have allowed for a deeper coverage of the human methylome. One of the unique features of the human methylome is that it is dynamic and previous studies have shown that age can have a strong impact on DNA methylation patterns. The dynamic nature of DNA methylation also influences EWAS methodology, both from a statistical and biological perspective. In this thesis, I explored EWAS methods and applications to ageing and age-related phenotypes. Firstly, I estimated EWAS power under several simulation scenarios and study designs, and my results suggested that the majority of recent EWAS studies lack statistical power to detect small DNA methylation effect sizes. I then applied EWAS to identify differential methylation CpG sites associated with three phenotypes, including ageing, birth weight and smoking. One of the novel findings from this thesis was that hundreds of genome-wide significant ageing-related hyper-methylated regions were identified across multiple tissues in twins. These findings confirm and extend previous work showing that ageing has a strong underlying effect on DNA methylation. Birth weight did not yield significant differential methylation sites, which may be partly explained by low power to detect modest methylation effects. Smoking is a well-known environmental risk factor for disease, and my analyses identified novel impacts of smoking on DNA methylation patterns in adipose tissue, which are of interest to cardiovascular and metabolic disease. I further explored the impacts of smoking by integrating DNA methylation and gene expression profiles in adipose tissue and in whole blood. In addition to identifying novel results, my findings also confirmed that the *AHRR* and *F2RL3* genes showed stable and consistent changes related to smoking in both DNA methylation and gene expression profiles across tissues. My findings explored methodological issues in genome-wide methylation studies and showed that age and smoking have a strong and reproducible effect on DNA methylation across tissues in humans, which suggests that these factors should always be included as covariates in EWAS of human complex traits.

# ACKNOWLEDGEMENTS

I humbly thank God, for leading and blessing me throughout my life.

I would like to express my deepest gratitude to my supervisors, Dr. Jordana Bell, Dr. Ana Valdes, and Prof. Tim Spector. They have foreseen my potential before anyone else, and offered me this marvellous opportunity to initiate my research life. They have been very supportive throughout my PhD study, and inspired me during my research.

Foremost, I am grateful to my primary supervisor, Dr. Jordana Bell, for her full support, patience, expert guidance, considerations, and offered me opportunities to achieve my goals. She not only gave me a lot of freedom to express and to think, but also a lot of advices to help me complete my work.

I would like to thank all my colleagues in the Twin Department, and our participants for their generous contributions to science. A special thank to senior researchers Dr. Kerrin Small, Dr. Wei Yuan, Dr. Ana Viñuela, Dr. Kirsten Ward, and Dr. Chris Bell, who is always willing to discuss and help me to solve problems. Thanks to my friends in the department, Idil, Abhishek, Leonie, Cristina, Juan, Andy, and Lisa, who have given me a colorful time and broaden my vision. Specifically to Idil, who is my best spiritual support and shared many work loadings with me. I would like to thank two of my friends, Tzu-Ching and Hsin-I, who backed me up all the time like my family.

A very special thank to Albert, who has been there with me through all the good and bad times in the past few years, and always supported me. He has also been the motivation for my work and life.

Most importantly, to my family in Taiwan, my parents, my aunt Lin Li-Ching, my brother Tsai Tsung-Hua, who gave me love and strength to go through all the difficult times and emotional moments, and always believe in me unconditionally.

感謝我的父母和所有家人 - 特別是小阿姨林麗卿和哥哥蔡宗樺。雖然遠在台灣，但他們給予了我無條件的愛、支持、信任和勇氣去克服重重難關。感謝天主。

# TABLE OF CONTENTS

<b>ABSTRACT .....</b>	<b>3</b>
<b>ACKNOWLEDGEMENTS.....</b>	<b>4</b>
<b>TABLE OF CONTENTS.....</b>	<b>5</b>
<b>TABLE OF FIGURES .....</b>	<b>10</b>
<b>TABLE OF TABLES .....</b>	<b>13</b>
<b>ABBREVIATIONS .....</b>	<b>15</b>
<b>Chapter 1 Introduction.....</b>	<b>16</b>
1.1 Epigenetics.....	16
1.2 DNA methylation .....	18
1.2.1 Role of DNA methylation during early development .....	19
1.2.1.1 Reprogramming in gametogenesis, embryogenesis and cell lineage differentiation .....	19
1.2.1.2 Genomic imprinting .....	20
1.2.1.3 X-chromosome inactivation .....	21
1.2.1.4 Regulation of transcription and maintenance of genome stability .....	21
1.2.2 DNA methylation variation and genetic factors .....	22
1.2.3 DNA methylation variation and environmental factors .....	25
1.2.3.1 Ageing .....	25
1.2.3.2 Nutrition and caloric intake .....	25
1.2.3.3 Air pollution, smoking, and others .....	26
1.3 Epigenome-Wide Association Scans (EWAS).....	27
1.3.1 The choice of assay platform for methylation-based EWAS .....	27
1.3.2 Study design .....	28
1.3.3 Power estimation for EWAS .....	30
1.3.4 Data quality control analysis .....	30
1.3.5 Replication and validation .....	31
1.4 Value of epigenetic studies in twins .....	32
1.5 TwinsUK cohort .....	32
1.6 Study aims .....	33
<b>Chapter 2 Power and Sample Size Estimation for Epigenome-wide Association Scans to Detect Differential DNA Methylation .....</b>	<b>35</b>
2.1 Introduction .....	35

2.2 Methods .....	38
2.2.1 An epigenetic model of complex disease susceptibility .....	38
2.2.2 DNA methylation distribution .....	41
2.2.3 Study Designs .....	42
2.2.4 Simulation parameters .....	43
2.2.5 Estimation of statistical power .....	45
2.3 Results .....	45
2.3.1 Power of case-control EWAS using mean difference effect estimates .....	45
2.3.2 Power of case-control EWAS using methOR effect estimates with restrictions on the mean differences .....	48
2.3.3 Power of discordant twin and case-control designs for small sample sizes and mean differences .....	48
2.3.4 Sample size required for 80% power in discordant twin and case-control designs for a range of mean differences .....	54
2.3.5 Methylation variance and methOR affect power under the same mean difference in the case-control design .....	55
2.3.6 DNA methylation variance and twin correlation can influence power in the EWAS twin design.....	59
2.4 Discussion.....	60
2.5 Conclusions .....	63

**Chapter 3 Materials and Methods: An Overview of the Methylation Datasets and Quality Control Procedure..... 64**

3.1 Methylation datasets .....	64
3.2 Illumina Infinium HumanMethylation assays .....	65
3.2.1 Illumina 27k array .....	65
3.2.2 Illumina 450k array .....	67
3.3 Quality control of the Illumina array data .....	69
3.3.1 Identification of probes mapping to multiple locations.....	69
3.3.2 Identification of outliers .....	70
3.3.3 Identification of the batch effects and covariates .....	72
3.3.4 Data normalization and adjusting for batch effects .....	73
3.4 Methylation heritability and patterns in twins .....	75
3.4.1 DNA Methylation Patterns in Twins .....	76
3.4.2 Heritability of DNA methylation.....	77

**Chapter 4 DNA Methylation Associates with Age and Age-Related Phenotypes .. 79**

4.1 Introduction .....	79
4.2 Materials and methods .....	82
4.2.1 Illumina 27k dataset .....	82
4.2.2 Illumina 450k dataset .....	82
4.2.3 Ageing-related clinical measurements .....	83
4.2.4 Statistical analyses .....	85
4.2.4.1 Age-differential methylation sites analyses .....	85
4.2.4.1.1 Permutation-based analysis (used on Illumina 27k) .....	85
4.2.4.1.2 Linear mixed effect regression (LMER) model (used on Illumina 27k and 450k) ..	85
4.2.4.2 Analysis of Illumina 27k data .....	86
4.2.4.3 Analysis of Illumina 450k data .....	86
4.2.4.4 Age acceleration analysis in Illumina 450k data .....	87
4.3 Results .....	88
4.3 Section A: Illumina 27k data (blood) .....	88
4.3.1 Overall methylation patterns with age .....	88
4.3.2 Age-related differential methylation site: Permutation-based .....	89
4.3.2.1 Using chronological age to estimate a-DMPs .....	89
4.3.2.2 Using DNA extraction age to estimate a-DMPs .....	90
4.3.3 Age-related differential methylation: LMER model .....	91
4.3 Section B: Illumina 450k datasets (blood, skin, adipose tissue) .....	94
4.3.4 a-DMPs analysis in three tissues .....	94
4.3.5 a-DMPs analysis across multiple tissues .....	95
4.3.5.1 Using a single Bonferroni adjusted P value as the significance threshold .....	95
4.3.5.2 Using FDR as significance criteria .....	96
4.3.6 Age acceleration analysis .....	97
4.3.6.1 DNA methylation age and age acceleration across three tissues .....	97
4.4 Discussion .....	101

## **Chapter 5 Epigenome-Wide Association Scans And Longstanding DNA Methylation Changes Related to Birth Weight in Discordant Monozygotic Twin Pairs .....**

5.1 Introduction .....	107
5.2 Materials and Methods .....	110
5.2.1 Datasets .....	110
5.2.1.1 Discovery dataset (BW discordant MZ twins) .....	110
5.2.1.2 Replication dataset (BW discordant MZ twins) .....	110
5.2.1.3 Verification dataset (unrelated female subjects) .....	111
5.2.2 Phenotypes .....	111

5.2.3 Methylation data .....	111
5.2.4 Statistical analyses .....	112
5.2.4.1 Quality Control for Illumina 450k data .....	112
5.2.4.2 Birth weight differentially methylated positions (BW-DMPs) analysis .....	112
5.2.4.3 Meta-analysis of twin datasets .....	114
5.2.4.4 Age differentially methylated positions (a-DMPs) analysis .....	114
5.2.4.5 Gene clustering analysis .....	114
5.3 Results .....	115
5.3.1 The demographic characteristics of the twin datasets .....	115
5.3.2 Birth weight differentially methylated positions (BW-DMPs) .....	115
5.3.2.1 Identification of BW-DMPs using ‘continuous trait analysis’ .....	115
5.3.2.2 Identification of BW-DMPs using ‘categorical trait analysis’ .....	121
5.3.2.3 Identification of BW-DMPs in 310 unrelated subjects .....	123
5.4 Discussion .....	125
5.4.1 Evidence of neonatal and postnatal variations in DNA methylation .....	125
5.4.2 Replication of BW-DMPs in different twin datasets .....	126
5.4.3 Disease-associated genes .....	127
5.4.4 Strengths and weaknesses of the current study .....	128

**Chapter 6 Tobacco Smoking Induces Coordinated DNA Methylation and Gene Expression Changes Across Multiple Tissues..... 129**

6.1 Introduction .....	129
6.2 Materials and methods .....	133
6.2.1 Datasets.....	133
6.2.1.1 DNA methylation and RNA-seq datasets in adipose tissue .....	133
6.2.1.2 DNA methylation and RNA-seq datasets in whole blood samples .....	133
6.2.2 Phenotype collection .....	134
6.2.3 Statistical analyses .....	135
6.2.3.1 Quality Control for Illumina 450k and RNA-seq data .....	135
6.2.3.2 Smoking differential methylation sites analysis .....	135
6.2.3.3 Smoking differentially expressed gene analyses .....	136
6.2.3.4 Correlation between methylation and gene expression levels in adipose tissue .....	137
6.2.3.5 Conditional analysis between methylation and gene expression levels in adipose tissue .....	137
6.2.3.6 Methylation quantitative trait locus (meQTL) and expression quantitative trait locus (eQTL) .....	138
6.3 Results .....	139
6.3.1 Smoking differentially methylated positions in adipose tissue (smoking-DMPs) ..	139

6.3.2 Smoking differentially expressed regions in adipose tissue .....	141
6.3.3 Comparison between the smoking EWAS and genome-wide expression results ...	144
6.3.3.1 Four genes overlap between methylation and expression adipose results .....	144
6.3.3.2 Correlations between methylation and expression of the 4 overlapping genes .....	145
6.3.4 Genetic contributions to methylation and exon expression levels .....	148
6.3.5 Association among cotinine levels, smoking status, and methylation levels .....	149
6.4 Discussion .....	150
6.4.1 Tissue-shared and adipose-specific smoking-DMPs .....	151
6.4.2 Highly replicated Smoking-DMPs .....	154
6.4.2.1 AHRR (aryl hydrocarbon receptor (AhR) repressor) gene .....	154
6.4.2.2 F2RL3 (coagulation factor II receptor-like 3) gene (also known as PAR-4) .....	155
6.4.2.3 2q37.1 region (close to ALPP/ALPPL2 genes) .....	155
6.4.2.4 Lung cancer related genes .....	156
6.4.2.5 Maternal smoking effect on newborns .....	157
6.4.2.6 Smoking-DMPs and disease .....	157
6.4.3 Cotinine levels, smoking status, and methylation .....	158
6.4.4 Genetic contributions to smoking-DMPs and differentially expressed exons .....	158
6.4.5 Smoking-DMPs or expressed genes overlap with GWAS results .....	159
6.4.6 Conclusion .....	159
<b>Chapter 7 Conclusions, Discussions and Future Perspectives .....</b>	<b>160</b>
<b>Appendix A: Epigenome-Wide Association Scans in Osteoarthritis .....</b>	<b>168</b>
<b>Appendix B: Publications Related to My PhD work .....</b>	<b>182</b>
<b>References .....</b>	<b>183</b>



# TABLE OF FIGURES

## Chapter 1 Introduction

---

Figure 1-1. Waddington’s Epigenetic Landscape (Waddington, 1957).....	17
Figure 1-2. Heritability estimation using twin model .....	23
Figure 1-3. Epigenome-wide association study designs .....	29

## Chapter 2 Power for EWAS

---

Figure 2-1. The hypothesis of statistical power .....	36
Figure 2-2. DNA methylation pattern at (A) the cellular and individual levels and (B) in the proposed methylation distributions in the study .....	40
Figure 2-3. Example of a simulation procedure .....	43
Figure 2-4. Power of large-scale case-control EWAS for a range of sample sizes. ....	47
Figure 2-5. Power of small-scale discordant twin (solid lines) and case-control (dashed lines) designs for a range of sample sizes and mean differences .....	49
Figure 2-6. DNA methylation variance and correlation can impact EWAS power.....	56
Figure 2-7. DNA methylation variances can impact case-control EWAS power.....	57
Figure 2-8. Unequal DNA methylation variances in cases and controls can impact EWAS power .....	58

## Chapter 3 Materials and methods

---

Figure 3-1. Beta values for the genome-wide DNA methylation of (A) autosomes and (B) X-chromosomes of a single subject in the Illumina 27k .....	67
Figure 3-2. Probe designs for type I (Infinium I) and type II (Infinium II) probe .....	68
Figure 3-3. Density of methylation levels of a single subject from the Illumina 450k..	69
Figure 3-4. Beta values in the autosomes from (A) batch 1 and (B) batch 2.....	71
Figure 3-5. Heatmap of beta values in (A) batch 1 and (B) batch 2 .....	72
Figure 3-6. Estimating batch effects using PC1 and PC2 .....	73
Figure 3-7. Density of methylation levels prior and post quantile-normalization in (A) batch 1 and (B) batch 2 .....	74
Figure 3-8. Example of dataset (A) before and (B) after BMIQ transformation in an Illumina 450k array .....	75

Figure 3-9. Correlation of genome-wide methylation levels between a random MZ and DZ twin pair, and unrelated subjects.....	76
Figure 3-10. Pearson’s correlation coefficient ( $r$ ) in MZ, DZ, and singletons for (A) batch 1 and (B) batch 2 .....	77
Figure 3-11. Density of ICC in MZ and DZ for (A) batch 1 and (B) batch 2.....	78

## **Chapter 4 Methylation and age**

---

Figure 4-1. Age acceleration estimates for (A) age acceleration differences and (B) age acceleration residuals .....	87
Figure 4-2. Scatter plot of PC1/PC2 and DNA extraction age .....	89
Figure 4-3. An example of age differential methylation at gene <i>SHANK2</i> .....	91
Figure 4-4. Manhattan plot of EWAS using chronological age at 5% FDR.....	92
Figure 4-5. Two most associated a-DMPs proximal to <i>NHLRC1</i> and <i>IRX5</i> genes.....	93
Figure 4-6. Functional annotation of a-DMPs .....	93
Figure 4-7. Tissue samples that overlapped across the three datasets .....	94
Figure 4-8. Tissue-shared and tissue-specific a-DMPs found in three datasets with same effect direction .....	96
Figure 4-9. Summary of the age acceleration in blood (red), adipose (yellow), and skin (green) tissues.....	98
Figure 4-10. Age acceleration patterns across tissues (blood, adipose, and skin) in (A) age acceleration differences, and (B) age acceleration residuals .....	99

## **Chapter 5 Methylation and birth weight**

---

Figure 5-1. Workflow of datasets and data analysis .....	112
Figure 5-2. Identification of BW-DMPs using the continuous trait and the categorical trait analysis .....	113
Figure 5-3. Manhattan plot of BW EWAS results of the observatory dataset.....	116
Figure 5-4. Manhattan plot of BW EWAS results of the replication dataset.....	117
Figure 5-5. Manhattan plot of meta-analysis results of continuous trait.....	118
Figure 5-6. Top 38 genes involved in biological processes .....	121
Figure 5-7. Manhattan plot of meta-analysis results in categorical trait.....	122
Figure 5-8. Manhattan plot of BW EWAS results of 310 unrelated subjects .....	123

## Chapter 6 Methylation and smoking

---

Figure 6-1. Workflow for the methylation and expression datasets .....	137
Figure 6-2. Manhattan plot of smoking EWAS results in the adipose tissue .....	139
Figure 6-3. An example of the methylation levels on different smoking groups on <i>CYP1A1</i> gene (cg23680900) .....	141
Figure 6-4. Manhattan plot of the smoking associated exon expression in the adipose tissue.....	142
Figure 6-5. Genome-wide smoking results in adipose tissue: comparisons between methylation and expression results .....	144
Figure 6-6. Correlation matrix between methylation and expression of the 4 genes in adipose tissue .....	145
Figure 6-7. Three proposed models of smoking effects on methylation and gene expression levels .....	146
Figure 6-8. The regional plot of the genome-wide results of <i>AHRR</i> gene.....	147
Figure 6-9. Regional plot of the EWAS results on 2q37.1 region .....	148
Figure 6-10. Cotinine levels in different smoking status .....	149

## Chapter 7 Conclusions

---

Figure 7-1. Example of association between different effect size and significant P values using adipose smoking EWAS results .....	162
--	-----

# TABLE OF TABLES

## Chapter 2 Power for EWAS

---

Table 2-1. Power of large-scale case-control EWAS using mean difference effects (1 case : 1 control) .....	46
Table 2-2. Power of large-scale case-control EWAS using mean difference effects (1 case : 2 controls).....	50
Table 2-3. Power of large-scale case-control EWAS using mean difference effects (1 case : 4 controls).....	51
Table 2-4. Power of large-scale case-control EWAS using methOR effects.....	52
Table 2-5. Power of EWAS twin and case-control designs .....	53
Table 2-6. Sample size required for 80% power in EWAS twin and case-control designs .....	54

## Chapter 3 Materials and methods

---

Table 3-1. Summary of the five methylation datasets.....	65
--	----

## Chapter 4 Methylation and age

---

Table 4-1. Summary of Illumina 27k datasets .....	82
Table 4-2. Summary of Illumina 450k datasets .....	83
Table 4-3. List of ageing-related indicators .....	84
Table 4-4. Age-related differential methylation identified with chronological and DNA extraction age .....	90
Table 4-5. Summary of twin pairs included in the three datasets .....	94
Table 4-6. List of significant a-DMPs found in three datasets .....	95
Table 4-7. Pairwise tissue-shared a-DMPs .....	96
Table 4-8. List of significant age accelerated phenotypes in three tissues .....	100
Table 4-9. Pairwise comparison of a-DMPs across six studies (Tsai <i>et al.</i> , 2012) .....	102
Table 4-10. Recent EWAS using chronological age on the Illumina 450k array .....	103

## **Chapter 5 Methylation and birth weight**

---

Table 5-1. Characteristics of the two twin datasets.....	115
Table 5-2. Summary of BW-DMPs discovered in the two twin datasets .....	116
Table 5-3. Top 51 probes found in meta-analysis.....	119
Table 5-4. Diseases associated with top BW-DMPs.....	120
Table 5-5. Summary of the BW-DMPs discovered in twin datasets .....	121
Table 5-6. Top 24 probes found in meta-analysis.....	122
Table 5-7. Significant BW-DMPs identified in 310 subjects .....	123
Table 5-8. Diseases associated with top BW-DMPs.....	124

## **Chapter 6 Methylation and smoking**

---

Table 6-1. Zygosity and smoking status in the adipose dataset .....	133
Table 6-2. Top 39 smoking-DMPs in adipose tissue .....	140
Table 6-3. Top 48 smoking-expressed exons in adipose tissue .....	143
Table 6-4. Results of conditional analysis .....	146
Table 6-5. List of meQTLs found in the top EWAS results .....	149
Table 6-6. Overview of recent smoking EWASs.....	151
Table 6-7. List of well-known smoking-DMPs identified from previous studies .....	152
Table 6-8. Tissue-shared smoking differential sites in adipose EWAS.....	153

## ABBREVIATIONS

Abbreviations	Full name
BMI	Body mass index
bp	Base pairs
BW	Birth weight
CGI	CpG island
CpG	Cytosine-phosphate-guanine site
DMP	Differentially methylated position
DMR	Differentially methylated region
DZ twin	Dizygotic twin (non-identical twin)
EWAS	Epigenome-wide association scan; Epigenome-wide association study
eQTL	Expression quantitative trait loci
FDR	False discovery rate
g	Gram
$h^2$	Heritability
ICC	Intra-class correlation coefficient
Illumina 27k	Illumina Infinium HumanMethylation27
Illumina 450k	Illumina Infinium HumanMethylation450
kb	Kilobase
LMER	Linear mixed effect regression
MeDIP-seq	Methylated DNA immune-precipitation sequencing
meQTL	Methylation quantitative trait loci
mRNA	Messenger ribonucleic acid
MZ twin	Monozygotic twin (identical twin)
N	Number of subjects
OA	Osteoarthritis
r	Pearson's correlation coefficient
rho ( $\rho$ )	Spearman's correlation coefficient
RA	Rheumatoid arthritis
SD	Standard deviation
SNP	Single nucleotide polymorphism
SLE	Systemic lupus erythematosus
T2D	Type 2 diabetes
TES	Transcription end site
TSS	Transcription start site

# Introduction

---

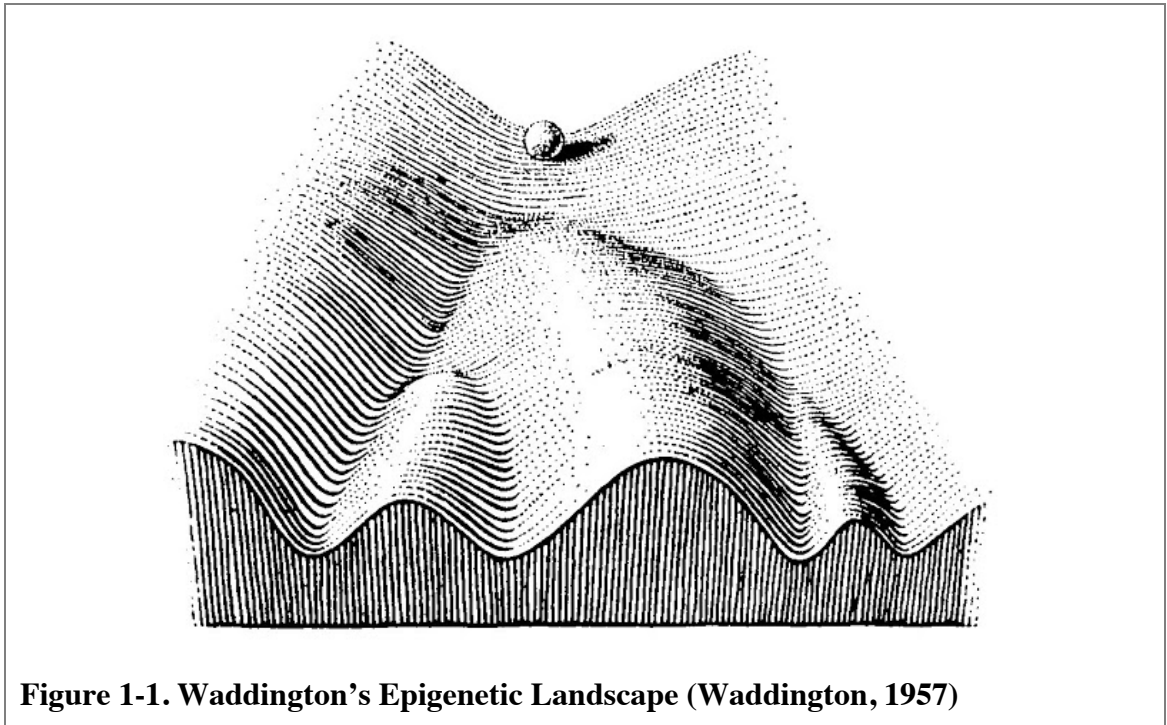
In this chapter, I will introduce the background of epigenetic modifications, and the features of my target measurements, which is the DNA methylation. I further discuss about the commonly used epigenetic method, the epigenome-wide association scan (EWAS), and the considerations for conducting this type of study. I also discuss about the value of using twin study design in the epigenetic studies, and give a brief overview of the data and phenotypes in the TwinsUK cohort. Lastly, I have summarized my study aims for this thesis.

Part of this work has been published as a review article in *Epigenomics* (Tsai *et al.*, 2012)

---

## 1.1 Epigenetics

Epigenetics was first introduced by Conrad Hal Waddington in 1942, who described it as a ‘branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being’ (Waddington, 1942). Epigenetics was explained as a stable mechanistic interplay between the genotype and phenotype without alteration to the DNA sequence. Subsequently, in 1957 (Waddington, 1957) Waddington proposed the idea of an ‘epigenetic landscape’ as a mathematical metaphor for the progression of cellular development, and a cell was analogous to a marble rolling from the hilltop, where some cells after a certain ‘decision-making’ process undergo differentiation (roll down to the valley) to eventually specialize in different function and expression from one another (Figure 1-1).



**Figure 1-1. Waddington's Epigenetic Landscape (Waddington, 1957)**

More recently epigenetics has been re-defined as a mechanism of regulating gene expression without changes in the DNA sequence (Holliday, 1994). The two most studied epigenetic modifications are the chemical alteration of the DNA, by the addition of a methyl group to the cytosine residues of DNA, known as DNA methylation, and modifications of the histone proteins and their tails that format chromatin structure (Qiu, 2006; Goldberg *et al.*, 2007), known as histone modification. Other epigenetic modifications also exist, such as the DNA hydroxymethylation, noncoding RNA regulation of expression (Ponting *et al.*, 2009) and nucleosome positioning (Portela & Esteller, 2010). These epigenetic modifications could interplay to modulate chromatin structure or regulate noncoding RNA to ultimately influence gene expression (Ponting *et al.*, 2009; Kaikkonen *et al.*, 2011). The key role of epigenetic mechanisms is to regulate gene expression, for example, through transcriptional-repression (Reik, 2007) or epigenetic silencing in cancer through demethylation of CpG islands in the promoter regions of tumour suppressor genes (Gonzalez-Zulueta *et al.*, 1995; Herman *et al.*, 1995; Merlo *et al.*, 1995).

Within the past decade, epigenetics has been increasingly studied in the context of complex diseases. Despite the characterization of many disease-related single nucleotide polymorphisms (SNPs) and haplotypes, it is becoming clear that the genetic makeup of an individual only contributes to a fraction of the predisposition to certain



phenotypes. For instance, genetic variation at the *BRCA1/2* in familial breast cancer is estimated to account for 30% of its occurrence. The missing 70% is termed the ‘missing heritability’ and is believed to be due to other factors, which might include epigenetic modifications induced by the environment of the individual.

Variation in the epigenome of an individual could occur prenatally, during early development of primordial germ cells (Reik *et al.*, 2001) or in postnatal life. It might involve three scenarios: (1) inherited changes that are present in all tissues (Antequera & Bird, 1993; Reik *et al.*, 2001; Bird, 2002); (2) stochastic changes that occur in early development (Waterland *et al.*, 2010) or arise in certain tissues during life (Fraga *et al.*, 2005; Z. A. Kaminsky *et al.*, 2009; Gibbs *et al.*, 2010; Ollikainen *et al.*, 2010; Pai *et al.*, 2011); and (3) changes triggered by environmental influences, such as long-lasting nutritional effects (Jaenisch & Bird, 2003; Feil, 2006).

In this thesis, I will focus on DNA methylation, as one of most studied and stable epigenetic processes. In the following sections, I discuss the characteristics of DNA methylation, its patterns in twins, and epigenome-wide association scans (EWAS) using DNA methylation.

## 1.2 DNA methylation

DNA methylation is currently the best understood epigenetic modification in mammals. It is a biochemical process where methyl groups (CH<sub>3</sub>) are added to cytosine bases by the enzymes DNA methyltransferases, which occurs primarily at cytosine-phosphate-guanine (CpG) dinucleotides. Rarely, methylation could also occur at non-CpG sites (Ziller *et al.*, 2011). CpGs are mostly methylated (70-80%) and have been estimated to comprise ~25% of the human genome, and are especially enriched around 15 base pairs (bp) upstream of the transcription start sites (TSS) (Saxonov *et al.*, 2006). Roughly 7% of CpGs cluster into regions known as CpG islands (CGI) (Bird, 2002) characterized as ~1kb regions with over 50% GC content and over 60% observed to expected CpG ratio (Gardiner-Garden & Frommer, 1987):

$$\text{CpG ratio} = \frac{\text{Number of CpG}}{\text{Number of C} \times \text{number of G}} \times \text{number of nucleotides}$$

CpG islands occur at ~1% of the genome and it has been estimated that 60-70% of the human gene promoters are enriched for CGIs, which are mainly unmethylated in normal cells (Weber *et al.*, 2007; Strausman *et al.*, 2009). DNA methylation can also be detected at the regions of lower CpG density that border the CpG islands, known as CGI shores (~2 kb). It has been shown that CGI shores are enriched for functional signals, such as tissue-specificity in DNA methylation profiles (Doi *et al.*, 2009; Irizarry *et al.*, 2009), where the majority of methylation changes during reprogramming tended to occur (Doi *et al.*, 2009), and are highly associated with gene expression linked to disease (Irizarry *et al.*, 2009).

Although DNA methylation is one of the most stable epigenetic mechanisms, evidence shows that dynamic changes in methylation occur during development and differentiation. In the following paragraphs I discuss methylation patterns during development.

### **1.2.1 Role of DNA methylation during early development**

DNA methylation plays a crucial role during development and normal physiological processes of mammals.

#### ***1.2.1.1 Reprogramming in gametogenesis, embryogenesis and cell lineage differentiation***

Reprogramming predominantly occurs during the germ cell stage and pre-implantation (Reik & Walter, 2001a; Santos & Dean, 2004; Buganim *et al.*, 2013; Stower, 2014). In the first stage, highly methylated primordial germ cells lose most of their methylation memory and reacquire it during the expansion phase. In the fertilization stage, the highly methylated primordial germ cells (PGCs) undergo another wave of de-methylation before embryonic day 12.5 when they migrate to the genital ridges. Most of the methylation is erased during this time and followed by *de novo* methylation after the fifth cell cycle (Reik & Walter, 2001b; J. Lee *et al.*, 2002; Yamazaki *et al.*, 2003). Passive de-methylation occurs during DNA replication (Bestor, 2000) since the DNA methyltransferase DNMT1 is required for copying existing methylation patterns, its absence at this stage causes the newly replicated strand to fail to become methylated.

The first differentiation event also occurs and selectively activates the lineage specific genes. The first two cell lineages: the inner cell mass (ICM; embryoblast) and trophoctoderm (TE) are established. In porcine *in vivo* developed latter blastocysts, the former becomes hyper-methylated and gives rise to somatic tissues, while latter stays unmethylated and forms the placenta (Santos & Dean, 2004; Morgan *et al.*, 2005; Kwon *et al.*, 2008; Huang & Fan, 2010).

DNA methylation levels can be tissue-specific or shared across tissues (Gibbs *et al.*, 2010; J. T. Bell *et al.*, 2011; Numata *et al.*, 2012; Gamazon *et al.*, 2013; Lokk *et al.*, 2014). Tissue-specificity occurs as cells transit from a pluripotent state to differential cell lineages during the course of early development. Cells acquire tissue-specific transcriptional programs from interaction with epigenetic mechanisms (Reik *et al.*, 2001; Albert & Peters, 2009; Hemberger *et al.*, 2009) that target DNA regulatory sequences such as promoters (Maston *et al.*, 2006). Thus there could be more methylation differences across tissues of the same individual than in the same cell type of two unrelated individuals (Eckhardt *et al.*, 2006). For example, in 283 samples of human blood, brain, kidney, and skeletal muscle tissues, there were both tissue-specific and tissue-shared sites and regions found across tissues (Day *et al.*, 2013). Tissue-specificity and cell heterogeneity of methylation levels impact choice of the appropriate tissue for EWAS. Generally, the most accessible tissues are saliva and whole blood. It is still unclear whether these samples are good surrogates for methylation from other tissues. Since methylation levels are heterogeneous among tissues and cells (Doi *et al.*, 2009), cell composition should also be adjusted for, especially when using whole blood and methods have been proposed to resolve this (Houseman *et al.*, 2012; Zou *et al.*, 2014).

### ***1.2.1.2 Genomic imprinting***

Genomic imprinting is defined as the allele-specific silencing of imprinted genes, where one parental allele is repressed (by DNA methylation) and the other is activated in a stable manner. The imprinted genes could be functionally different depending on the parental origin (McGrath & Solter, 1984; Surani *et al.*, 1984) and subsequently can affect human genetic diseases differently (Nicholls *et al.*, 1989; Henry *et al.*, 1991). During methylation reprogramming of germ cells, methylation levels are erased and

reset on imprinted differential methylation sites at the imprinting control centres (Tucker *et al.*, 1996). During fertilization, a rapid wave of de-methylation occurs on the paternal non-imprinted genetic sequences (Mayer *et al.*, 2000; Oswald *et al.*, 2000), and the maternal genome is de-methylated passively during the DNA replication stage of embryogenesis (Howlett & Reik, 1991; Reik & Walter, 2001b). The expression of imprinted genes is determined by whether the allele is paternally or maternally inherited, for example, at the *H19/IGF2* region only the maternal copy of *H19* is expressed, whereas only the paternal copy of *IGF2* is expressed (Barlow *et al.*, 1991; Bartolomei *et al.*, 1991; DeChiara *et al.*, 1991). Another similar interesting imprinting example in disease is the chromosome 15q11-q13 region, where two different neurological disorders result from imprinting errors, Angelman syndrome and Prader-Willi syndrome (Cassidy *et al.*, 2000).

### ***1.2.1.3 X-chromosome inactivation***

In females, there are two copies of the X chromosome. One of two X chromosomes is silenced at random, known as X-chromosome inactivation, and the silencing is maintained by DNA methylation. DNA methylation on the X-chromosome in females shows chromosome-wide hemi-methylated levels (Avner & Heard, 2001). The inactivated X chromosome is condensed into the 'Barr body' (Barr & Bertram, 1949) as a result of whole-chromosome silencing (Ohno *et al.*, 1959; Lyon, 1961).

### ***1.2.1.4 Regulation of transcription and maintenance of genome stability***

For decades, DNA methylation has been reported to regulate gene expression by repressing transcription (Holliday & Pugh, 1975; Riggs, 1975). At gene promoters, a negative correlation between methylation and gene expression is observed across multiple tissues (Eckhardt *et al.*, 2006; Ball *et al.*, 2009; Lister *et al.*, 2009). This might occur either by methylation directly blocking access of transcription binding factors to the binding site sequence in the promoter, or indirectly through chromatin remodelling. In the first approach, CpG methylation can block the chromatin boundary element binding protein, transcriptional repressor CTCF, access to DNA and allow deactivation of the promoter activity (A. C. Bell *et al.*, 1999; Ohlsson *et al.*, 2001). In the second approach, the methylated DNA can be bound by Methyl-CpG-binding Proteins, such as

MeCP2 (Lewis *et al.*, 1992; Nan *et al.*, 1997), and resulting in recruitment of histone deacetylase (HDAC) and inactivation of chromatin structure, therefore silencing of the gene (Jones *et al.*, 1998; Nan *et al.*, 1998; Fuks *et al.*, 2003).

However, the correlation between gene body methylation and gene expression is not fully understood. DNA methylation levels in the gene body can be positively correlated with gene expression levels and mid-range expressed genes appear to have the highest methylation levels (Zilberman *et al.*, 2007; Zemach *et al.*, 2010; Jjingo *et al.*, 2012). In a recent study, a negative correlation between adipose DNA methylation and gene expression levels was observed on sites located in the gene-body or 1500 bp upstream of the TSS across 13,532 genes (Grundberg *et al.*, 2013). It is still unclear whether methylation is driving changes in gene expression or if it is itself a consequence of gene regulation (Schubeler, 2012). One recent study suggested that DNA methylation could actively or passively associate with gene expression, depending on where it occurs in the genome (Gutierrez-Arcelus *et al.*, 2013).

### **1.2.2 DNA methylation variation and genetic factors**

During the development process, the methylation marks are thought to be wiped out. Recent evidence shows that methylation levels at different regions of genome are under the influence of genetic variants, suggesting that they are heritable. The heritability of phenotype estimates the proportion of phenotype differences among individuals that are contributed by their genetic differences. There are two types of heritability: broad-sense heritability ( $H^2$ ) and narrow-sense heritability ( $h^2$ ). Each phenotype is composed of a genotype (G) and an environment (E), and the variance of this phenotype is explained by the variance of genotype, environment, and the covariance of genotype and environment as:

$$Var (P) = Var (G) + Var (E) + 2 Cov (G, E)$$

In the broad-sense heritability, all the genetic contributions, such as additive, dominant, epistatic, and parental effects are considered and defined as:

$$H^2 = \frac{Var (G)}{Var (P)}$$

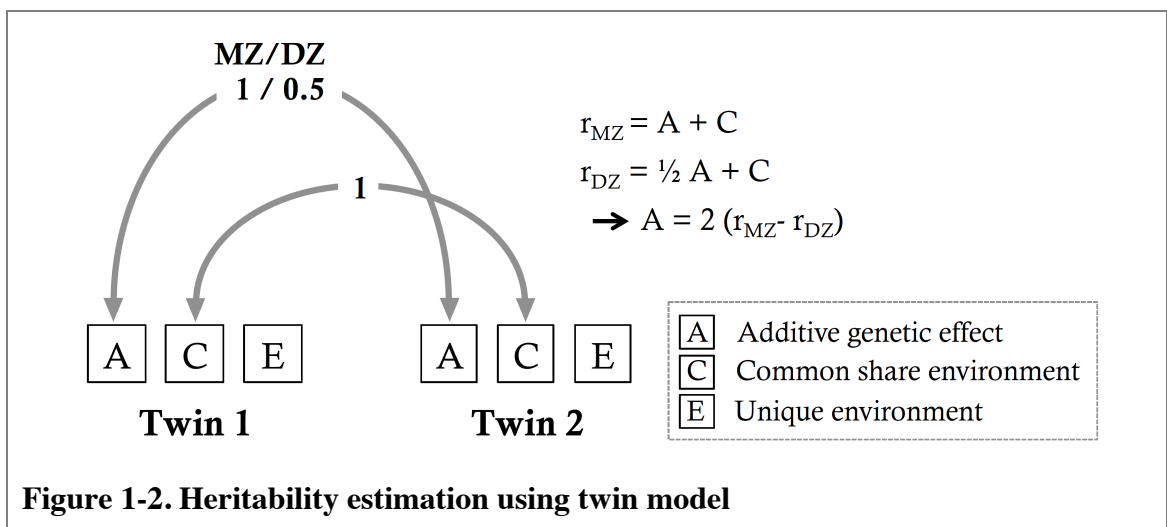
The narrow-sense heritability, which is what I consider in the thesis, only accounts for the additive genetic effect, is defined as:

$$h^2 = \frac{Var(A)}{Var(P)}$$

Using twin study, by comparing genetically identical MZ twins to DZ twins, who share on average 50% of germ line genetic variants, one can estimate: the genetic contribution (A), the shared environmental components (C), and the unique environmental components (E) to the phenotype (Falconer, 1996).

The narrow-sense heritability of the genome-wide methylation in my thesis is estimated from the comparison of the intra-class correlation coefficients (ICC) in MZ and DZ twins (Figure 1-2). I assume that the common shared environment is the same for both MZ and DZ twins, and their heritability is calculated by:

$$heritability (h^2) = 2 \times (ICC_{MZ} - ICC_{DZ})$$



Evidence for DNA methylation heritability comes from the observations that methylation is more similar within MZ twins than DZ twins (Z. A. Kaminsky *et al.*, 2009; J. T. Bell *et al.*, 2012; Gordon *et al.*, 2012) and methylation patterns segregate within families (Bjornsson *et al.*, 2008). The average methylation heritability across the genome was estimated to be about 18% in Illumina 27k blood data (J. T. Bell *et al.*, 2012) and 19%-34% in Illumina 450k adipose tissue data (Grundberg *et al.*, 2013), depending on the probe subset. Genetic variants can impact DNA methylation levels (or

methylation quantitative trait loci, meQTLs) and the majority of associations occur in *cis* (Schilling *et al.*, 2009; Gibbs *et al.*, 2010; J. T. Bell *et al.*, 2011; Gertz *et al.*, 2011; J. T. Bell *et al.*, 2012; Numata *et al.*, 2012; Drong *et al.*, 2013; Gamazon *et al.*, 2013). By comparing meQTLs identified in blood (J. T. Bell *et al.*, 2012), lymphoblastoid cell lines (J. T. Bell *et al.*, 2011) and four brain tissues (Gibbs *et al.*, 2010), nearly 28% and 34% of the 1,537 blood meQTLs overlapped with those in brain and lymphoblastoid cell lines, respectively. The overlap shows that methylation levels at these CpG sites are strongly heritable and conserved across tissues over time.

It is debated whether epigenetic modifications can be passed down to offspring due to 'foetal reprogramming' in early development. During stages of germ cell development and pre-implantation of the embryo, the genome-wide epigenetic marks are erased to restore the totipotency of the fertilized egg (Reik *et al.*, 2001). The epigenetic marks then become mitotically stable after cell differentiation or by the end of the cell cycle (Reik *et al.*, 2001; Morgan *et al.*, 2005). Studies show that some of these epigenetic marks are not always erased and can be transmitted from the parents to offspring through the germ line (Chong & Whitelaw, 2004), which could provide the basis for trans-generational epigenetic inheritance (Daxinger & Whitelaw, 2012; Grossniklaus *et al.*, 2013).

An example of trans-generational epigenetic inheritance could be observed in the agouti viable yellow ( $A^{vy}$ ) gene in mice (Wolff *et al.*, 1998). These agouti mice have the intracisternal A particle (IAP) retrotransposon inserted upstream of the agouti gene. If the  $A^{vy}$  locus were unmethylated then the agouti protein would become overexpressed and produce a viable yellow ( $A^{vy}/a$ ) mouse. The yellow-coated mice are obese, have shorter lifespan and higher risk of cancer compared to their non-yellow coated siblings (Miltenberger *et al.*, 1997). The epigenetic regulation of the offspring agouti mice could be influenced by maternal intake of methyl-supplemented diet (Wolff *et al.*, 1998), and a yellow-coated mother could pass on a higher proportion of the silenced  $A^{vy}$  allele to their offspring through the germ line, thus suggesting that epigenetic modification was not fully erased in germ cells (Morgan *et al.*, 1999). In humans, genetic associations with DNA methylation at many CpG sites, as well as DNA methylation heritability findings in twins, show that methylation at some genomic regions can be influenced by

the underlying genetic sequence, and imply that these regions may show evidence for heritability across generations.

### **1.2.3 DNA methylation variation and environmental factors**

Here, I discuss environmental and lifestyle factors that have been linked to differential methylation sites and potentially disease in humans.

#### ***1.2.3.1 Ageing***

The DNA methylation profile of an individual has been shown to change during the ageing process. Using twins, researchers have shown on a genome-wide scale that methylation changes over time, with younger monozygotic twins sharing more similarities in methylation than older twins (Fraga *et al.*, 2005). This suggests that methylation variability arises from ageing and different lifestyles (Fraga *et al.*, 2005). In another study, the methylation levels of three genes within the same subjects were longitudinally compared over five years. The direction of change over time was gene-specific, and overall changes were different among individuals (Wong *et al.*, 2010). The CpG sites at which methylation levels change over time are known as age-related differential methylation sites (a-DMRs/a-DMPs). A-DMRs/a-DMPs have been identified across different tissues and cells, including white blood cells (Boks *et al.*, 2009; B. C. Christensen *et al.*, 2009; Rakyan *et al.*, 2010; Teschendorff *et al.*, 2010; Adkins *et al.*, 2011; Alisch *et al.*, 2012; J. T. Bell *et al.*, 2012), skin (Gronniger *et al.*, 2010; Koch *et al.*, 2011), saliva (Bocklandt *et al.*, 2011), and human brain tissues (Hernandez *et al.*, 2011; Numata *et al.*, 2012). Some of these differential methylation sites were proposed as stable biomarkers for chronological age prediction (Bocklandt *et al.*, 2011; Koch & Wagner, 2011; Burgess, 2013; Horvath, 2013). More discussion of age-related methylation changes is presented in Chapter 4.

#### ***1.2.3.2 Nutrition and caloric intake***

The maternal nutritional intake in mammals, such as folate, vitamin B6 and B12, and betaine, or caloric restrictions during pregnancy, can influence methylation changes in offspring (Alegria-Torres *et al.*, 2011; Feil & Fraga, 2011). These methylation changes



are often also associated with phenotype changes and metabolic-related diseases, such as low birth weight (see discussion in Chapter 5), obesity and type II diabetes (Seki *et al.*, 2012). For example, the diet of the *agouti* mouse interacts with the *agouti* gene to influence coat colour (Morgan *et al.*, 1999). After feeding the yellow-coated *agouti* mother choline, folic acid, betaine, and vitamin B12 before and during pregnancy, the offspring are predominantly brown and have lower susceptibility to obesity and diabetes (Waterland & Jirtle, 2003). Similarly, in humans, the Dutch famine study also reveals that nutritional insufficiency during pregnancy can influence the methylation status of the offspring. The children who are prenatally exposed during the Dutch famine of 1944-45 have lower methylation level of the imprinted *IGF2* gene when compared to their same-sex siblings who are not exposed to famine *in utero* after six decades. Those subjects also have significantly higher rates of metabolic syndrome later in life (Heijmans *et al.*, 2008). This study suggests that the methylation changes from early-life nutrition condition may persist throughout life.

### ***1.2.3.3 Air pollution, smoking, and others***

Environmental toxins, such as air pollution, benzene, dioxin, and cigarette smoking, can induce DNA methylation changes. Strong evidence for impacts of tobacco smoking on methylation changes at many genomic regions has been identified across populations and tissues (Breitling *et al.*, 2011; Philibert *et al.*, 2013; Shenker *et al.*, 2013; Zeilinger *et al.*, 2013; H. R. Elliott *et al.*, 2014; Y. Zhang *et al.*, 2014). Among these regions, the CpG sites in the *AHRR* gene, a mediator of carcinogenic agents PAHs which causes tobacco-related lung cancer, are identified to be the top associated and replicated regions in smoking (Philibert *et al.* 2013; Shenker *et al.* 2013; Zeilinger *et al.* 2013; Elliott *et al.* 2014). Other environmental agents that associated with methylation changes are alcohol consumption (B. C. Christensen & Marsit, 2011; H. Zhang *et al.*, 2013), UV radiation (Stein, 2012), pain perception (J. T. Bell *et al.*, 2014), psychological stress (Groom *et al.*, 2011), and sunlight (Gronniger *et al.*, 2010).

## 1.3 Epigenome-Wide Association Scans (EWAS)

Recent DNA methylation studies in humans have largely expanded from candidate gene studies to EWAS studies (Rakyan, Down, *et al.*, 2011). In this section, I introduce EWAS and discuss analytical considerations for methylation array-based EWAS studies. To date, EWAS focus on characterizing DNA methylation, however in the near future, there is potential to examine other epigenetic processes, such as histone modification. There are challenges in performing EWAS on human complex traits, discussed below (see also (Tsai *et al.*, 2012)).

### 1.3.1 The choice of assay platform for methylation-based EWAS

Several platforms have been developed to detect genome-wide methylation levels. They could be categorized into three main groups (J. T. Bell & Spector, 2011; Heyn & Esteller, 2012): microarray-based, enrichment-based followed by sequencing, and bisulfite sequence-based. The most cost-effective platform is the microarray-based approach, for example, Illumina Infinium® HumanMethylation27 (Illumina 27k) (Bibikova *et al.*, 2009) and Illumina Infinium® HumanMethylation450 (Illumina 450k) (Dedeurwaerder *et al.*, 2011) bead arrays, and comprehensive high-throughput arrays for relative methylation (CHARM) (Irizarry *et al.*, 2008). On these hybridization-based arrays, DNA samples are first bisulfite treated followed by bead-anneal genotyping (Illumina system) or have restriction enzyme application for methylation detection (CHARM). Among the Illumina microarrays, the old version is the Illumina GoldenGate Methylation Cancer Panel I (Illumina GoldenGate) (Bibikova *et al.*, 2006) that targeted 1,500 cancer-related CpG sites. The next version is the Illumina 27k, where coverage increased to 27,000 CpG sites in promoter-specific regions that were predominantly unmethylated. The latest version is the Illumina 450k that covers ~485,000 CpGs that are predominantly located near genes, and represent 5% out of approximately  $10^7$  possible CpG sites across the genome. One potential benefit of the continual use of these standardized array-based platforms is to minimize variation caused during the methylation detection and their wide use also allows for the possibility of meta-analysis across studies.

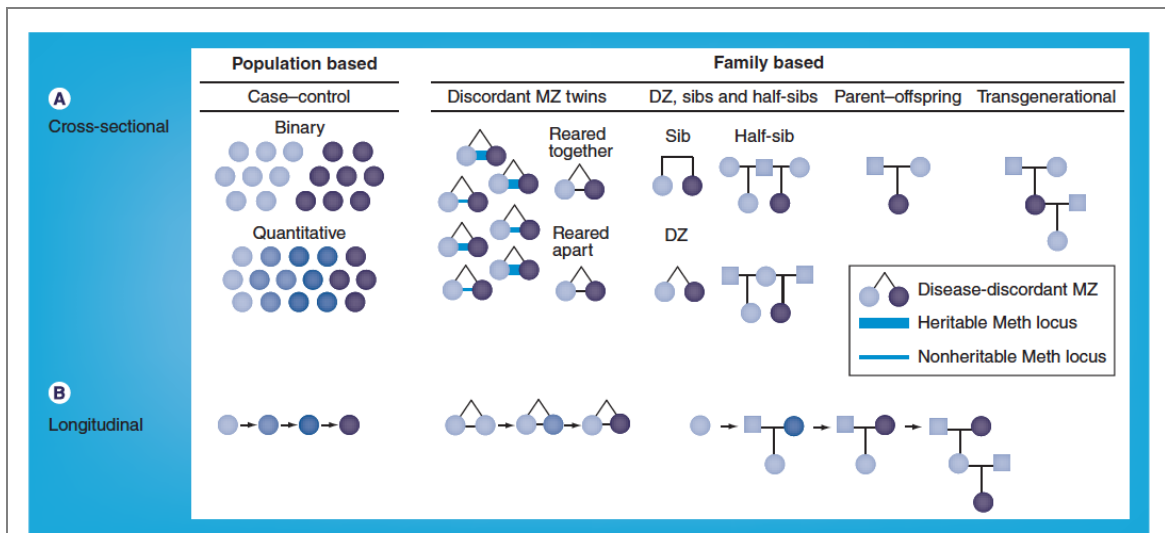
The enrichment-based platforms captures the methylated portions of the genome by array or sequencing-based methods, and include methylated DNA immune-precipitation sequencing (MeDIP-seq), methylated DNA capture by affinity purification sequencing (MeCAP-seq), and methylated DNA binding domain sequencing (MBD-seq). Here, the methylation levels are at the resolution level of DNA fragment size (usually up to 500 bp) instead of at a single CpG level resolution. The bisulfite sequencing-based methods include whole-genome bisulfite sequencing (WGBS), which is currently the gold standard. WGBS has the widest coverage and also gives single CpG resolution, however, this method is costly and cannot differentiate between 5-hydroxymethylation and 5-methylation.

The choice of platforms drastically affects the coverage and resolution across the genome, and the sensitivity. Sample throughput and genome coverage across multiple platforms has previously been discussed (Laird, 2010), and cost, sample size, and coverage are mutually dependent. All methods have strengths and weaknesses, therefore, it is important to validate results using multiple methylation assays (Mensaert *et al.*, 2014).

### **1.3.2 Study design**

There are several study designs that can be used in an EWAS setting (Rakyan, Down, *et al.*, 2011) and some are summarized in Figure 1-3 (Tsai *et al.*, 2012).

The two major study designs for EWAS are the cross-sectional and longitudinal designs. To determine causality between methylation changes and disease occurrence, longitudinal studies would be required. Because DNA methylation levels change over time, long-term monitoring of methylation changes prior to and after phenotype changes or disease onset would be optimal. The cross-sectional study is a widely used EWAS design due to sample availability. Because methylation levels are only sampled at one time point in the cross-sectional study (usually after disease occurs), it is difficult to discern if the methylation changes are the cause or consequence of the phenotype changes.



**Figure 1-3. Epigenome-wide association study designs**

(A) Cross-sectional and (B) longitudinal study designs for EWAS using population- and family-based samples. Circles and squares are females and males, where dark purple data points are individuals with disease. For quantitative phenotypes, the colour scale is the quantile of the phenotypic distribution. Lines between the MZ twin pairs show different levels of DNA methylation heritability at the CpG site of interest.

DZ: Dizygotic; Meth: Methylation; MZ: Monozygotic; sib: Sibling. Reproduced from Tsai *et al.* 2012, Figure 1.

The cross-sectional study design can further be divided into population-based and family-based. In the population-based case-control design, the disease status of cases is examined retrospectively and controls are typically chosen with matched age and gender. The methylation levels are then compared between cases and controls to identify disease-related differential methylation sites. In family-based designs, the twin design is well-characterized (J. T. Bell & Saffery, 2012). The disease discordant MZ twin design is one optimal design because identical twins share nearly 100% of their genetic variants and are matched for age and gender, and share similar embryonic and early developmental environment. The purpose of using MZ twins is primarily to identify the environmentally driven or stochastic methylation changes in the case. This design can be MZ twins raised in the same environment until a certain age, or reared-apart disease discordant MZ twins that are valuable in detecting environmental effects. Furthermore, the disease discordant MZ twin design is applicable to non-twin studies, i.e. in cancer research. The general approach is to acquire the tissue samples from both the cancerous region and healthy region of the same subject, then compare the methylation differences between samples.

Other family-based studies include comparisons across MZ, DZ, sibling, half-sibling, parent-offspring, and multigenerational families. The parent-offspring and trans-generational comparisons are useful for detecting methylation heritable regions.

### **1.3.3 Power estimation for EWAS**

The power of an EWAS study depends on the sample size, significance level, and effect size of the target loci. Few studies have addressed power for a case-control design or MZ twin design. The methylation differences of differential methylation positions identified in MZ pairs tend to be quite small, perhaps because twins share more similar methylation patterns compared to unrelated subjects. Also, in general disease discordant MZ twins are rare, and small samples will have low power to detect DMPs of modest effects. Using a simulation-based approach, I estimated power for EWAS study designs and found that in addition to the common determinates, other factors can also impact power, such as study design, test statistic, and the underlying methylation structure (see Chapter 2).

### **1.3.4 Data quality control analysis**

Multiple methods have been proposed quality control assessment of array-based methylation data and several R packages were developed for normalization. To avoid potential batch effects, the appropriate experimental design should be used. Randomizing samples in the experimental design and performing analyses to adjust for confounders should be a routine for EWAS study. The general procedure for data quality control should be to firstly identify the outliers and batch effects. An initial data check of the raw methylation patterns is highly recommended prior to analysis, including an assessment of the correlation patterns in the genome-wide methylation estimates across entire samples, cases and controls, across genome-wide loci, within autosomes and sex chromosomes.

A notable issue in EWAS analysis has been the correction for multiple testing. The significance level depends on the total number of CpG sites examined, for example, the Bonferroni adjusted P value of  $10^{-7}$  is one significance threshold for the Illumina 450k

array. However, patterns of co-methylation across the genome indicate non-independence in methylation levels at CpG sites that are located close together, and therefore the Bonferroni correction can be over-conservative. More discussion on quality control and multiple testing corrections are presented in Chapter 2 and Chapter 3.

### **1.3.5 Replication and validation**

Similar to GWAS study, it is important to replicate the EWAS results in an independent sample. This is particularly important in the context of determining whether the differential methylation is causal or consequential of the disease status. Replication guidelines for the initial stages of EWAS have previously been discussed (Tsai *et al.*, 2012). Briefly, the replication of identified differential methylation should be implemented using the same ethnic population and tissue. The validation of the DNA methylation signal in the region of interest should be performed using different technologies. Current validations for Illumina array data are typically performed using bisulfite sequencing or pyrosequencing on the regions of interest. Custom validation assays could be chosen for specific diseases or phenotypes, for example, promoter-rich assays, or CpG-shore-rich assays to study cancer or tissue specific regions.

An example showing the importance of appropriate analysis and validation of high throughput technology results is that of recent findings of allele-specific expression and RNA editing. Several studies have commented on the identification of imprinting, and mismatch of mRNA-DNA sequence difference (RDD) sites in one individual (DeVeale *et al.*, 2012; Kelsey & Bartolomei, 2012; Kleinman & Majewski, 2012; W. Lin *et al.*, 2012; Pickrell *et al.*, 2012). In the original studies, researchers found an extremely high frequency of genomic imprinting loci and RDD sites, where subsequent studies found an alarmingly high proportion of false positives from the noise created by analysis errors (e.g. mapping error caused by sequence alignment and sequencing errors). This highlights the importance that appropriate efforts need to be invested into the proper methodological approaches to analyse large-scale high-throughput datasets.

## 1.4 Value of epigenetic studies in twins

Studying twins allows us to better understand the biological processes underlying the regulation of methylation, as well as understanding the heritability of methylation. The evidence that DNA methylation is heritable comes from the fact that MZ twins share similar methylation levels compared to dizygotic (DZ) twins (Z. A. Kaminsky *et al.*, 2009) and from a family clustering study where methylation patterns are segregated within the family (Bjornsson *et al.*, 2008).

Phenotypic differences between MZ twins are commonly thought to be an outcome of environmental contributions. Accumulated phenotypic differences between MZ twins illustrate how environment and lifestyle can together change the susceptibility to disease. These might be identified using the discordant MZ twin design. Previous studies using discordant MZ twins aimed to find phenotype-related differentially methylated positions (DMPs) or regions (DMRs), such as in systemic lupus erythematosus (SLE), rheumatoid arthritis (RA) (Javierre *et al.*, 2010), and multiple sclerosis (Handunnetthi *et al.*, 2010). The disease discordance rates have been observed to vary from 5% to 75% in MZ twins (Petronis *et al.*, 2003; Ballestar, 2010), for example, the discordant rate in MZ for osteoarthritis (OA) is 40% (Spector *et al.*, 1996), which suggests that there is a strong environmental or epigenetic component.

The twin study design is a powerful method to find the non-genetic risk factors in a study design matched for genetics, age, sex, and similar environmental exposures in early development (Snieder, 2010; van Dongen *et al.*, 2012). It enables researchers to answer the following questions: (1) are the causes of the discordant disease status in genetically identical twins epigenetic, and (2) what is the heritability of epigenetic factors?

## 1.5 TwinsUK cohort

The TwinsUK cohort was established since 1992 to recruit MZ and DZ same-sex twins (Moayyeri *et al.*, 2013). The majority of participants are healthy female Caucasians (age range from 16 to 98 years old). There are more than 13,000 twin participants from all regions across the United Kingdom and many have multiple visits over the years.

Participants were asked to complete health questionnaires at their visits or by postal service during their follow-up period. These include information about their health, self and family disease history, medication use, and habitual behaviours, such as smoking and alcohol consumption. There are collections of clinical (e.g. bone mineral density) and phenotype measurements (e.g. blood pressure and lung function), and biochemical measures from biological samples (e.g. whole blood and urine samples).

Furthermore, there are also extensive -omic data available, such as genomic, epigenomic, and gene expression data in multiple tissues. The major research interest has been to discover the association between these -omic tools and healthy ageing, and with age-related phenotypes. For example, in this thesis, I have used methylation array data to study epigenetics in smoking and follow-up the top results with an integrative -omics approach in order to understand the underlying mechanisms. The major study design is a population-based design to improve study power, as well as undertaking a disease discordant MZ twin design in the birth weight chapter (Chapter 5).

## **1.6 Study aims**

Epigenetics has given us a better understanding of our genome, its interactions and functions, helping to find out the etiology and disease mechanisms that are not fully explained by genetic sequence changes.

We now know that epigenetic modifications play a crucial role in prenatal/postnatal development and cell differentiation through various means, such as histone modification and differential methylation on the imprinted genes. Recent EWAS studies are finding associations between epigenetic modifications and phenotype changes for complex traits. Correspondingly more demands have been placed onto sequencing technology and enabled a systematic assessment of the DNA methylome, allowing more detailed epigenome-wide scans. The standardized platform (i.e. Illumina 450k) now yields a considerable amount of coverage and high-throughput scans of the methylome, and it has allowed for rapid replication studies and meta-analysis of different data resources.



The key focus of EWAS studies has been the identification of differential methylated positions or regions. These can be long-term epidemiological biomarkers for disease or environmental exposure indicators. Longitudinal EWAS study is an important design to understand the disease mechanisms, epigenetic mediation of disease, and important biological pathways. The outcome of this research can ultimately contribute to modern medicine, for example, (1) to inform prognosis; (2) find a treatment to reverse epigenetic changes by repressing the transcription and hence slowing the disease progression; (3) identify the risk factors for prediction and prevention; (4) identify the biomarkers for treatment efficacy. Moreover, specific differential methylation patterns are showing promise of profiling phenotypes quite precisely, for example, chronological age, gender, tissue samples, and smoking status.

The goal of my PhD is to understand and apply EWAS analytical methodology to age, age-related phenotypes and disease risk. From this, I also expect to find out more the ways in which differential methylation can affect phenotypes along with integrating other -omics data, such as gene expression data. I hope to offer new insights into methylation studies in the related phenotype field.

The chapters in this thesis are arranged in the following order. Firstly, an overview of simulation-based estimate of the power of an EWAS study, and discussions of the power differences under a wide range of EWAS parameter settings (Chapter 2). Followed by a review of the methodological considerations in EWAS study, and discussions of the current pipeline for the quality control and analytical methods in array-based methylation data (Chapter 3). The subsequent three chapters describe our findings and results of our EWAS study of three phenotypes: ageing and age-related phenotypes (Chapter 4), birth weight (Chapter 5), and smoking (Chapter 6). In my early work, I have performed an EWAS for osteoarthritis with preliminary results (see Appendix A). The last chapter is a discussion and concluding remarks of my thesis work (Chapter 7).

# Power and Sample Size Estimation for Epigenome-wide Association Scans to Detect Differential DNA Methylation

---

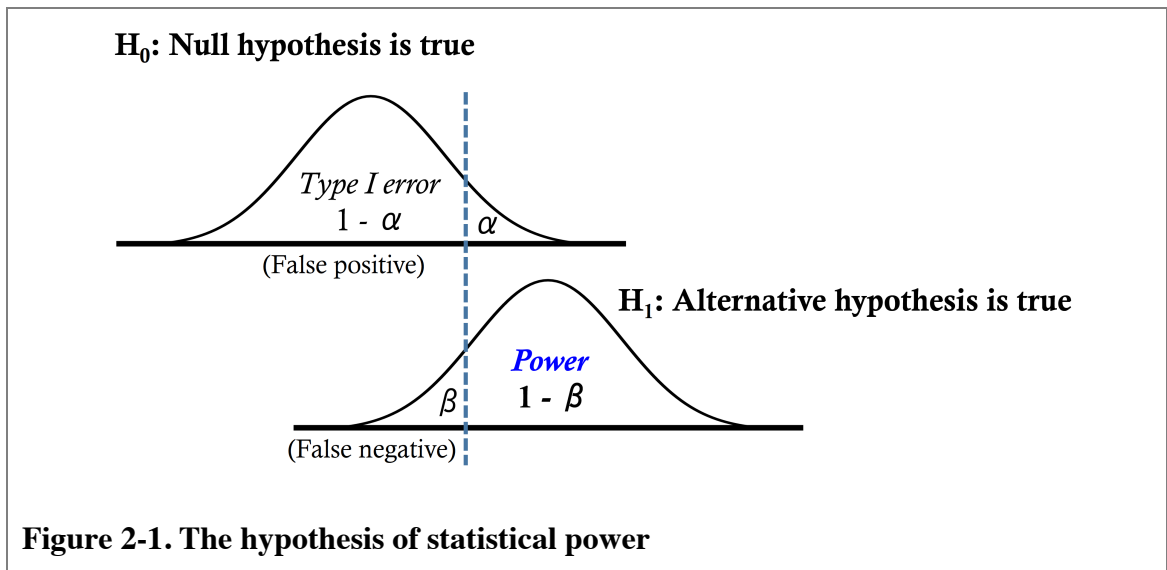
Epigenome-wide association studies (EWAS) are under way for many complex human traits, however EWAS power has not been fully assessed. I have investigated power of EWAS to detect differential methylation using case-control and disease discordant MZ twin designs with genome-wide DNA methylation arrays. In this chapter I provide power estimates for array-based DNA methylation EWAS under these two study designs and using both parametric and non-parametric analysis, and explore the multiple factors that impact on EWAS power.

This work has been published as a research article in International Journal of Epidemiology (Tsai & Bell, 2015)

---

## 2.1 Introduction

Statistical power refers to a statistical test to see the probability that it correctly rejects the null hypothesis ( $H_0$ ) when it is false (Figure 2-1). A sufficient study power, mostly suggested as 80%, represents a study having sufficient sample size to detect the minimum effect size between the case and control changes. Without a sufficient power, the chance of the false discoveries can be increased. In addition, results may only occur in the studied population, and not applicable to the other populations.



Several factors should be taken into account when considering EWAS power, including EWAS study design, the effect size, and multiple testing correction. First, the appropriate study design will determine the analysis method and is crucial to sample size estimation and power. The case-control study is the most widely performed disease association design. However, the disease discordant MZ twin study is often considered to be optimal in epigenetics, because genetic contributions can be adjusted for as co-twins share identical genetic variants over most of their genome (Rakyan, Down, *et al.*, 2011). The standard analyses for these designs are to compare paired or unpaired mean methylation differences or ranks of methylation levels between groups. Second, the DMP effect size is clearly a major factor determining the power of the study. In EWAS, effect size has typically been measured as the mean methylation difference between the groups ( $\text{Mean Meth}_{\text{case}} - \text{Mean Meth}_{\text{control}}$ ), or an alternative measure, the methylation odds ratio (methOR), has also been proposed for discrete traits and is calculated as:

$$\text{methOR} = \frac{\text{Mean Methylation}_{\text{case}} \times (1 - \text{Mean Methylation}_{\text{control}})}{(1 - \text{Mean Methylation}_{\text{case}}) \times \text{Mean Methylation}_{\text{control}}}$$

A CpG site with a 0.2 (20%) median methylation difference between groups is considered to be a DMP with a 99% confidence from a previous technology report (Bibikova *et al.*, 2011; Touleimat & Tost, 2012). However, this threshold might be too stringent for the discordant MZ twin design because MZ twins share more similar methylation levels (J. T. Bell *et al.*, 2012). Recent EWAS of discordant MZ twins identified diseases-related DMPs with effects as small as 2%, which is the identified

effect at the promoter of the *DOK7* gene in breast cancer using 15 MZ pairs (Heyn *et al.*, 2013). Other studies also reported changes of a small magnitude at disease-related DMPs with 0.13% to 6.6% in type 1 diabetes (Rakyan, Beyan, *et al.*, 2011), 10% in pain (J. T. Bell *et al.*, 2014), and > 10% difference in SLE (Javierre *et al.*, 2010). Third, the EWAS significance level requires adjustment for multiple comparisons. Adjusted thresholds depend in part on the methylation array coverage or the significance estimated by false discovery rate (FDR). With an increasing variety of arrays, there is a need to determine the minimum sample size required to reach sufficient power while incorporating all of the above factors.

Although power plays a crucial role in EWAS studies, only few studies have addressed it in detail. Recently, two studies explored power for EWAS case-control studies via simulations. Wang (S. Wang, 2011) assumed that each methylation marker is composed of three distribution categories: unmethylated, hemi-methylated, and methylated levels, to represent Uniform distribution, Normal distribution, and Uniform distribution, respectively. Different parameter settings, such as the proportion of the three distributions, and mean and standard deviation of the normal distribution were tested under three scenarios. The author found that the t-test was not powered to detect small mean differences between cases and controls when the proportion of the three contributing distributions differed in the two groups. However, the t-test was adequate to detect large mean differences between the groups when the methylation distribution between the case and control group was similar. The second case-control EWAS power study by Rakyan *et al.* (Rakyan, Down, *et al.*, 2011) proposed that the distribution of methylation variable positions (MVPs) should follow the beta distribution, where the majority of controls are methylated and cases composed of varying proportions of unmethylated, hemi-methylated, and methylated levels. Similar to Wang's study, the authors suggested that the methylation variance can affect power and should be accounted for. They concluded that sufficient power is attainable if a locus is less variable in both case and control groups, and the methOR was suggested to be a better predictor of effect instead of the mean difference.

In the twin design, most power estimates were based on array-based EWAS, and suggested that a sample of 10 to 25 MZ pairs was sufficient in some cases to reach up to 80% power (E. Dempster *et al.*, 2010b; Rakyan, Beyan, *et al.*, 2011; J. T. Bell *et al.*,

2012; Gervin *et al.*, 2012; Hasler *et al.*, 2012) to detect differential methylation. The first power estimation in twins was based on methylation changes at the *DLX1* gene in 9 MZ twin pairs (Z. Kaminsky *et al.*, 2008) using a specific DNA methylation array. A spot-wise standard deviation of the methylation difference was calculated within the pairs, followed by power analysis performed on these standard deviation (SD) distributions to detect the proportion of loci that have 80% power and sample size (pairs of twins) required to achieve the fixed effect size. Using a significance level of 0.001 based on the family-wise error rate and at  $4.1 \times 10^{-6}$  for a 1.2-fold change after Bonferroni correction, the authors found that 25 twin pairs were sufficient to reach 80% power to detect a 1.2 fold change in DNA methylation using the particular array, which did not provide single-CpG resolution data. More recent studies report low (35%) to good (> 80%) power to detect DMPs at single CpG-sites with methylation differences of 5-6% between affected and unaffected twins in 20-22 disease discordant twin pairs (E. Dempster *et al.*, 2010a; J. T. Bell *et al.*, 2012).

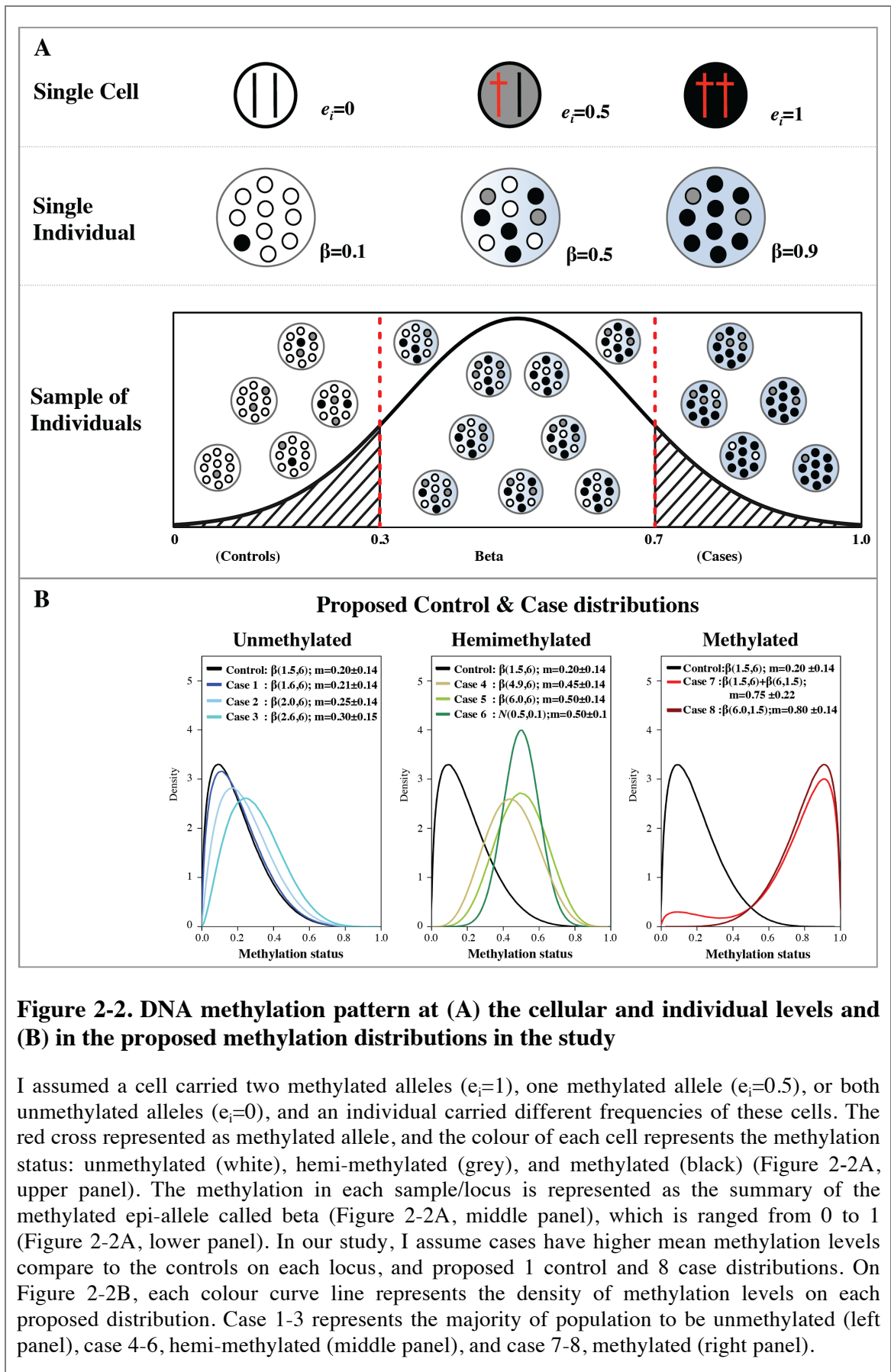
Here, I estimate power of EWAS study to detect the differential methylation under methylation platforms, such as the Illumina 450k, and estimate power for the case-control and disease discordant MZ twin study designs. I also evaluate the sample size that is required to achieve 80% power under a variety of methylation differences for the two study designs. Finally, I identify potential factors that impact EWAS power.

## 2.2 Methods

### 2.2.1 An epigenetic model of complex disease susceptibility

I assume that disease risk is affected by the DNA methylation status at a single locus,  $l$  (Figure 2-2A, upper panel), where  $l$  represents a single CpG-site in the genome. The methylation status at locus  $l$  in a single cell can be represented as a biallelic marker, where epi-allele 1 represents the presence of the methylated mark, and epi-allele 0 represents the absence of methylation. I assume that the disease-associated methylation mark occurs prior to the onset of disease and is faithfully transmitted through mitotic cell division. I denote DNA methylation status (epi-genotype) at locus  $l$  as  $e_j$ , where the  $e_j$  takes the value of 0, 0.5, and 1 to correspond to unmethylated, hemi-methylated, and

methyated states for a single cell, while the frequencies of the methylation status were represented as  $f(e_0)$ ,  $f(e_{0.5})$ , and  $f(e_1)$ , respectively. Each individual cell can consist of unmethylated, hemi-methylated, and methylated epi-genotypes with probabilities of  $p_1$ ,  $p_2$ , and  $p_3$ , where  $p_1 + p_2 + p_3 = 1$  (Figure 2-2A, upper panel). A sample from an individual  $i$ , represents a population of cells (Figure 2-2A, middle panel), and I assume that the contribution of each cell to the population is constant and without bias. The sample-level DNA methylation estimate is a function of the methylation levels of the composition of cells (Figure 2-2A, lower panel), and can be described by different functions or epigenetic models (Slatkin, 2009). In this study, a threshold model was proposed where the sample-level DNA methylation estimate reflects the allele frequency of the methylated epi-allele 1 in the cell population. That is, DNA methylation level for each sample is denoted as  $\beta$ , which represents the sum of its fully methylated cells plus half of its hemi-methylated cells. In addition to the proposed DNA methylation threshold model, dominant and recessive models may also be applied, for example, as proposed for genetic disease susceptibility risk (Risch, 1990c, 1990b, 1990a).



## 2.2.2 DNA methylation distribution

Multiple methods can profile DNA methylation patterns across the genome. Here I focus on microarray-based datasets, such as those generated by the Illumina 450k array. At each CpG-site, the Illumina 450k array-based DNA methylation levels are characterized as a finite bounded quantitative trait, represented as  $\beta$ , calculated as:

$$\beta = \frac{\text{Methylated signal}}{\text{Methylated signal} + \text{Unmethylated signal} + 100}$$

The methylation distribution in one subject can range from 0 (unmethylated) to 1 (methylated). Previous work has proposed that a single or bimodal beta distribution can be used to describe the single-locus distribution of DNA methylation levels on the Illumina 450k array (Rakyan, Down, *et al.*, 2011). I therefore propose 9 potential single-locus DNA methylation distributions in the context of our epigenetic disease susceptibility models. I assume that the absence of methylation is linked to the absence of disease, and propose that the locus of interest follows an unmethylated distribution in unaffected individuals (Control distribution, black line in Figure 2-2B), which is described by  $\beta(1.5,6)$  with a mean methylation level of 0.2. In our model affected individuals will show higher levels of DNA methylation at the locus of interest relative to controls, and I therefore propose 8 possible single-locus methylation distributions in affected individuals (Case 1 – Case 8 distributions, multiple colour lines in Figure 2-2B), with increasing levels of sample-wide DNA methylation. The 8 case distributions had increasing ordinal mean methylation difference with the control distribution that ranged from 1% to 60% in mean DNA methylation level. The 8 case distributions included 3 distributions (Case 1 – Case 3) with mean methylation levels  $\leq 0.3$  (unmethylated), 3 distributions (Case 4 – Case 6) with mean methylation levels  $\geq 0.45$  and  $\leq 0.5$  (hemi-methylated), and 2 distributions with mean methylation levels  $\geq 0.75$  (methylated). The three proposed unmethylated case distributions, Case 1 to 3, follow  $\beta(1.6,6)$ ,  $\beta(2,6)$ , and  $\beta(2.6,6)$  with a mean methylation level of 0.21, 0.26, and 0.30 respectively, and mean methylation difference of 1%, 5%, 10%, with the control distribution, respectively. Case 4 and Case 5 characterize hemi-methylated distributions of  $\beta(4.9,6)$  and  $\beta(6.0,6)$  with mean methylation levels of 0.45 and 0.5, respectively, and mean methylation differences of 25% and 30%. Case 6 is also hemi-methylated, but



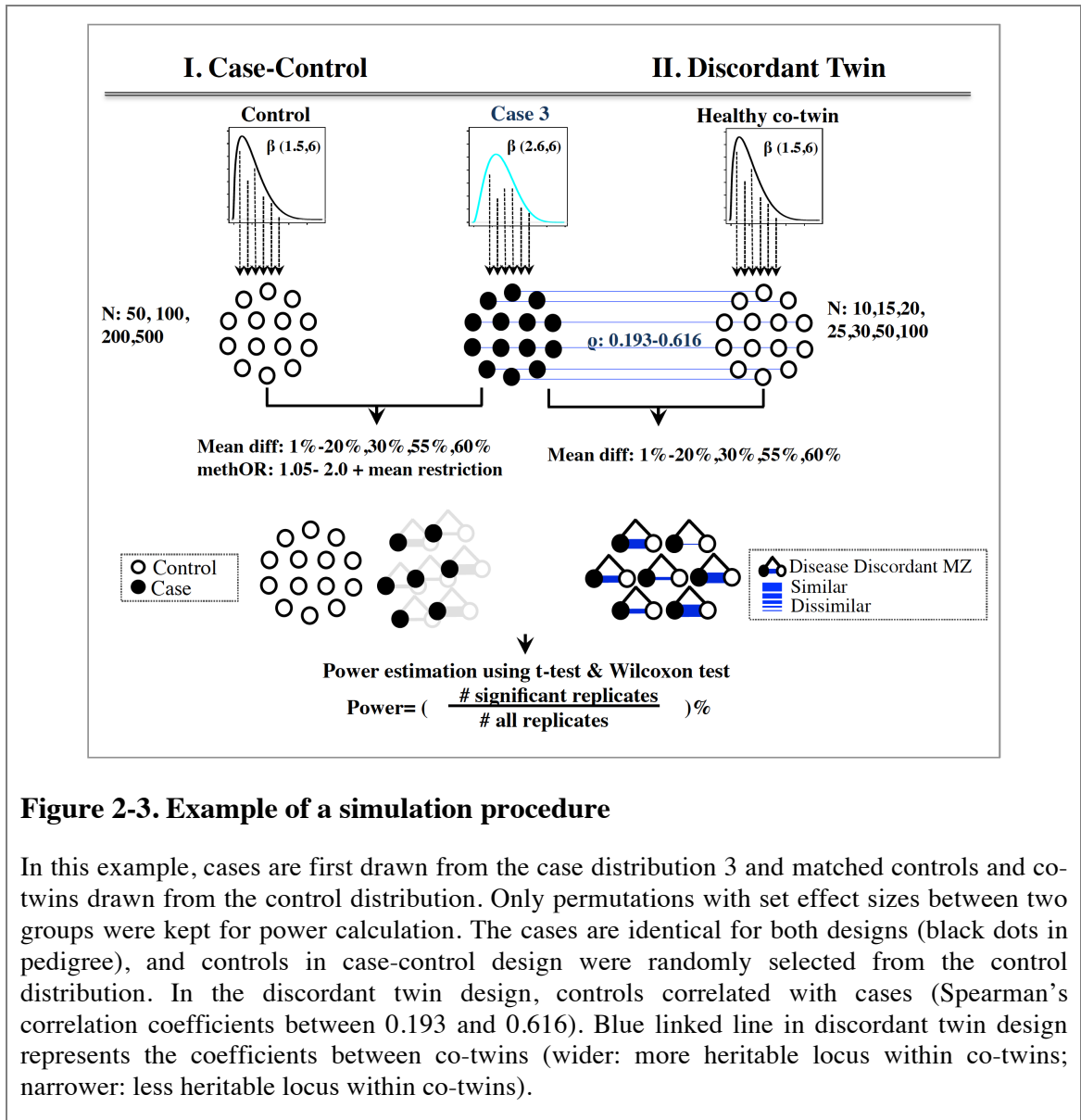
follows the normal distribution  $N(0.5,0.1)$ , and has the same mean methylation level as Case 5, but a smaller standard deviation. Case 7 follows the combination of 9% of  $\beta(1.5,6)$  and 91% of  $\beta(6,1.5)$  with a mean methylation of 0.75, and methylated Case 8 follows the  $\beta(6,1.5)$  with a mean of 0.8 that is diametrically opposed the control distribution. The mean methylation difference between Case 7 and Case 8 with the control distribution was 55% and 60%, respectively.

### 2.2.3 Study Designs

Two EWAS study designs were considered: case-control and discordant disease monozygotic (MZ) twins. MZ twins share nearly all of their genetic variants, and are also matched for age, gender, cohort effects, *in utero* and maternal effects, and many early life environmental factors. All of these factors have either been shown or are hypothesized to influence DNA methylation levels throughout the genome. Therefore, MZ twins are a much more homogeneous sample relative to genetically heterogeneous unrelated individuals who are exposed to different environments throughout life, and correspondingly MZ twins have been shown to have much more similar levels of DNA methylation compared to DZ co-twins and unrelated pairs of individuals (Z. A. Kaminsky *et al.*, 2009; J. T. Bell *et al.*, 2012). It is difficult to incorporate all of these factors in our simulation study, therefore in an attempt to minimize some of these effects, I assumed that all individuals in our study were the same age, gender, and were exposed to similar cohort effects. This will bias the case-control sample towards homogeneity and may give inflated power estimates for the case-control design.

To compare power under the same parameters in the case-control and twin designs, I assumed that the cases were identical in both studies and their matched controls and unaffected co-twins were sampled based on the locus-specific correlation in DNA methylation levels between groups. Cases were selected from one of the 8 Case distributions, and for the disease discordant MZ twin design unaffected co-twins were sampled from the control distribution if: (1) the mean difference within the co-twins matched the pre-specified effect size and (2) the Spearman's correlation coefficient within MZ pairs was between 0.193 and 0.616, which represented the genome-wide mean correlation coefficients  $\pm 1$  SD in a previously published set of 21 MZ twins using Illumina 27k (J. T. Bell *et al.*, 2012). Once MZ twin pairs were selected, for each

affected twin (or case) I also sampled a matched healthy unrelated control sample from the control distribution if the mean difference between the cases and controls matched the pre-specified effect size. Figure 2-3 shows an example simulation procedure by selecting the cases from distribution 3 and both matched unrelated controls and matched healthy co-twins from the control distribution.



**Figure 2-3. Example of a simulation procedure**

In this example, cases are first drawn from the case distribution 3 and matched controls and co-twins drawn from the control distribution. Only permutations with set effect sizes between two groups were kept for power calculation. The cases are identical for both designs (black dots in pedigree), and controls in case-control design were randomly selected from the control distribution. In the discordant twin design, controls correlated with cases (Spearman's correlation coefficients between 0.193 and 0.616). Blue linked line in discordant twin design represents the coefficients between co-twins (wider: more heritable locus within co-twins; narrower: less heritable locus within co-twins).

## 2.2.4 Simulation parameters

A range of sample sizes of disease discordant MZ twin pairs and case-control samples were considered. As MZ twins are more difficult to recruit than unrelated cases and controls I used a smaller sample size for the twin design, specifically 10, 15, 20, 25, 30,

and 50 MZ twin pairs and case-control pairs. Power calculation was also performed for larger case-control sample sizes of 50, 100, 200, and 500 pairs of unrelated individuals (that is, altogether 100 to 1000 individuals in the sample). As an estimate of effect size I used two approaches. First, I used the mean difference in methylation levels between affected and unaffected individuals, which ranged from 1% to 20%, 25%, 30%, 55%, and 60%. The selection of effect sizes and sample sizes was based on recently published EWAS findings, described in the introduction, and we further extended it to cover a broader range. In the case-control simulation results (Tables 1a - 1c), because I did not have power to detect the effects at 1% methylation difference at single locus significant ( $P < 0.05$ ) with 500 cases and controls, therefore the simulations with methylation differences less than 1% were not performed. The mean difference was used to estimate effect size for both the twin and case-control designs. For the case-control design I also calculated effect sizes under the methOR, which previously (Rakyan, Down, *et al.*, 2011) was defined as:

$$methOR = \frac{Mean\ Methylation_{Case} \times (1 - Mean\ Methylation_{Control})}{(1 - Mean\ Methylation_{Case}) \times Mean\ Methylation_{Control}}$$

Given the pre-specified range of mean methylation differences (1% to 60%), I also calculated the methOR that ranged from 1.05 to 2.0, and combined these with a certain maximum mean difference value to minimize methylation effect variability. This was done because the range of mean differences tends to be narrower for larger samples. For example, for a methOR = 1.2, the range of mean differences is 2.63% to 3.68% in 50 case-controls, whereas the range is 2.78% to 3.38% in 500 case-controls. To reduce the bias caused by the variation of mean difference, a cut-off of 3% mean difference was set along with methOR = 1.2. To detect potential factors that affect power, pooled standard deviations (pooled SD) of groups were calculated by:

$$Pooled\ SD\ (SD_{Case,Control}) = \sqrt{\frac{(N_{Case} - 1) * SD_{Case}^2 + (N_{Control} - 1) * SD_{Control}^2}{(N_{Case} + N_{Control} - 2)}}$$

I also assessed the correlation in DNA methylation profiles between cases and controls, and between affected twins and healthy co-twins. I calculated the between-group correlation using Spearman's correlation coefficients ( $\rho$ ). The statistical significance was set at 0.05 for single locus gene analysis, and a P value threshold of  $10^{-6}$  was used

for genome-wide significance. This threshold was selected using a Bonferroni correction based on a subset of the number of probes on the Illumina 450k array, because some regions show evidence for co-methylation pattern. Furthermore, because recent EWAS using Illumina 450k data have reported an FDR-based thresholds of 1% to 5% FDR with corresponding P values close to  $P = 1 \times 10^{-4}$  (Grundberg *et al.*, 2013; Nardone *et al.*, 2014). Therefore, the stringent significance level was relaxed from  $10^{-7}$  to  $10^{-6}$  in the study.

### **2.2.5 Estimation of statistical power**

Power estimation was based on simulations, where for example, if 800 out of 1000 simulations surpassed the pre-specified significance level I would estimate 80% power. A t-test with a prior F-test for equal variance was performed in the case-control design and a paired t-test was performed in the twin study design. All of the case-control simulations include equal and unequal variances between cases and controls with the exception of one case-multiple control scenario with a greater proportion of unequal variances. Table 2-1, Table 2-2, Table 2-3 show results from simulations with equal variances between cases and controls. The corresponding non-parametric analyses (Mann Whitney U test and Wilcoxon rank sum test) were also performed.

## **2.3 Results**

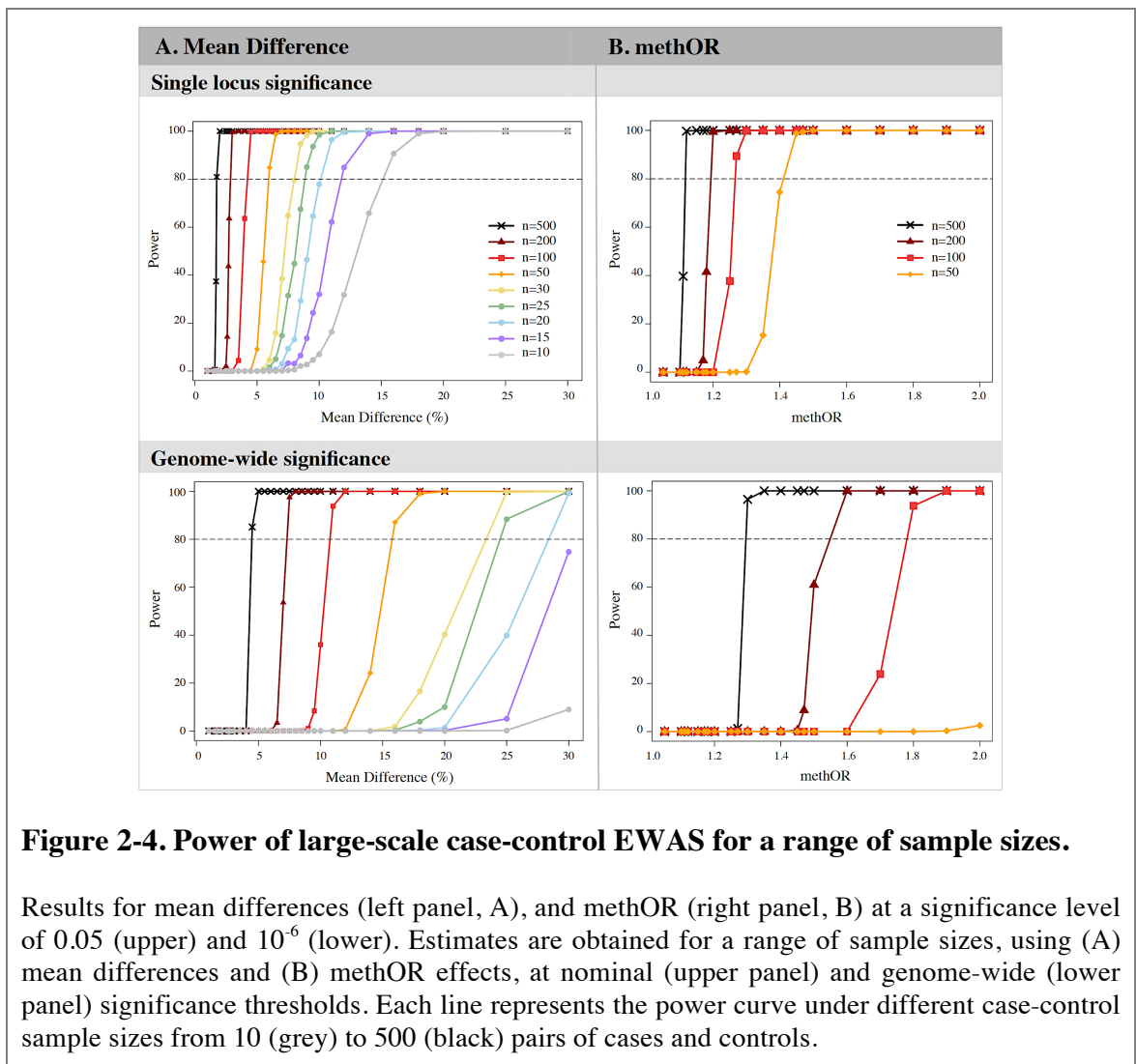
### **2.3.1 Power of case-control EWAS using mean difference effect estimates**

The mean test statistics of 1000 permutations were obtained for eight case distributions and one control distribution, by sampling effect sizes (using the mean difference) of 1% to 20%, 30%, 55%, and 60% and with increasing sample sizes from 10 to 500 pairs of cases and controls, that is, 20 to 1000 individuals altogether. Table 2-1 shows the EWAS power with increasing sample sizes from 10 to 100.

**Table 2-1. Power of large-scale case-control EWAS using mean difference effects (1 case : 1 control)**

Diff (%)	Dist <sup>1</sup>	N=10										N=20										N=25									
		P < 0.05			P < 10 <sup>-6</sup>			P < 10 <sup>-6</sup>			P < 0.05			P < 10 <sup>-6</sup>			P < 10 <sup>-6</sup>			P < 0.05			P < 10 <sup>-6</sup>			P < 10 <sup>-6</sup>					
		T <sup>2</sup>	W <sup>2</sup>	W <sup>0</sup>	T <sup>2</sup>	W <sup>2</sup>	W <sup>0</sup>	T <sup>2</sup>	W <sup>2</sup>	W <sup>0</sup>	T <sup>2</sup>	W <sup>2</sup>	W <sup>0</sup>	T <sup>2</sup>	W <sup>2</sup>	W <sup>0</sup>	T <sup>2</sup>	W <sup>2</sup>	W <sup>0</sup>	T <sup>2</sup>	W <sup>2</sup>	W <sup>0</sup>	T <sup>2</sup>	W <sup>2</sup>	W <sup>0</sup>	T <sup>2</sup>	W <sup>2</sup>	W <sup>0</sup>			
1	c1	1.07 (1.07, 1.11)	0.13 (0.09, 0.34)	0	0	0	1.06 (1.06, 1.09)	0.14 (0.09, 0.30)	0	0	0	1.06 (1.06, 1.09)	0.14 (0.08, 0.20)	0	0	0	1.06 (1.06, 1.09)	0.14 (0.08, 0.20)	0	0	0	1.06 (1.06, 1.09)	0.14 (0.08, 0.20)	0	0	0	0	0	0		
1.5	c1	1.10 (1.07, 1.11)	0.13 (0.09, 0.34)	0	0	0	1.11 (1.08, 1.14)	0.14 (0.07, 0.20)	0	0	0	1.10 (1.08, 1.14)	0.14 (0.07, 0.20)	0	0	0	1.10 (1.08, 1.14)	0.14 (0.07, 0.20)	0	0	0	1.10 (1.08, 1.14)	0.14 (0.07, 0.20)	0	0	0	0	0	0		
1.6	c1	1.11 (1.08, 1.23)	0.14 (0.07, 0.22)	0	0	0	1.11 (1.09, 1.16)	0.14 (0.07, 0.21)	0	0	0	1.11 (1.09, 1.16)	0.14 (0.07, 0.21)	0	0	0	1.11 (1.09, 1.16)	0.14 (0.07, 0.21)	0	0	0	1.11 (1.09, 1.16)	0.14 (0.07, 0.21)	0	0	0	0	0	0		
1.7	c1	1.11 (1.09, 1.19)	0.14 (0.07, 0.22)	0	0	0	1.11 (1.10, 1.20)	0.14 (0.07, 0.20)	0	0	0	1.11 (1.10, 1.20)	0.14 (0.07, 0.20)	0	0	0	1.11 (1.10, 1.20)	0.14 (0.07, 0.20)	0	0	0	1.11 (1.10, 1.20)	0.14 (0.07, 0.20)	0	0	0	0	0	0		
2	c1	1.14 (1.10, 1.23)	0.13 (0.06, 0.22)	0	0	0	1.13 (1.10, 1.25)	0.14 (0.09, 0.20)	0	0	0	1.13 (1.11, 1.18)	0.14 (0.08, 0.20)	0	0	0	1.13 (1.11, 1.18)	0.14 (0.08, 0.20)	0	0	0	1.13 (1.11, 1.18)	0.14 (0.08, 0.20)	0	0	0	0	0	0		
2.5	c1	1.17 (1.12, 1.31)	0.13 (0.07, 0.22)	0	0	0	1.17 (1.13, 1.23)	0.14 (0.07, 0.21)	0	0	0	1.17 (1.13, 1.23)	0.14 (0.07, 0.21)	0	0	0	1.17 (1.13, 1.23)	0.14 (0.07, 0.21)	0	0	0	1.17 (1.13, 1.23)	0.14 (0.07, 0.21)	0	0	0	0	0	0		
2.6	c1	1.18 (1.13, 1.27)	0.13 (0.06, 0.22)	0	0	0	1.18 (1.13, 1.27)	0.14 (0.07, 0.21)	0	0	0	1.18 (1.13, 1.27)	0.14 (0.07, 0.21)	0	0	0	1.18 (1.13, 1.27)	0.14 (0.07, 0.21)	0	0	0	1.18 (1.13, 1.27)	0.14 (0.07, 0.21)	0	0	0	0	0	0		
2.7	c1	1.21 (1.15, 3.09)	0.13 (0.06, 0.22)	0	0	0	1.21 (1.14, 1.29)	0.14 (0.08, 0.22)	0	0	0	1.21 (1.14, 1.29)	0.14 (0.08, 0.22)	0	0	0	1.21 (1.14, 1.29)	0.14 (0.08, 0.22)	0	0	0	1.21 (1.14, 1.29)	0.14 (0.08, 0.22)	0	0	0	0	0	0		
3	c1	1.25 (1.18, 1.49)	0.13 (0.07, 0.23)	0	0	0	1.25 (1.18, 1.34)	0.14 (0.08, 0.20)	0	0	0	1.25 (1.18, 1.34)	0.14 (0.08, 0.20)	0	0	0	1.25 (1.18, 1.34)	0.14 (0.08, 0.20)	0	0	0	1.25 (1.18, 1.34)	0.14 (0.08, 0.20)	0	0	0	0	0	0		
3.5	c1	1.27 (1.20, 1.46)	0.14 (0.06, 0.23)	0	0	0	1.27 (1.21, 1.37)	0.14 (0.08, 0.20)	0	0	0	1.27 (1.21, 1.37)	0.14 (0.08, 0.20)	0	0	0	1.27 (1.21, 1.37)	0.14 (0.08, 0.20)	0	0	0	1.27 (1.21, 1.37)	0.14 (0.08, 0.20)	0	0	0	0	0	0		
4	c2	1.30 (1.23, 1.45)	0.14 (0.07, 0.22)	0	0	0	1.30 (1.23, 1.43)	0.14 (0.08, 0.21)	0	0	0	1.30 (1.23, 1.43)	0.14 (0.08, 0.21)	0	0	0	1.30 (1.23, 1.43)	0.14 (0.08, 0.21)	0	0	0	1.30 (1.23, 1.43)	0.14 (0.08, 0.21)	0	0	0	0	0	0		
4.5	c2	1.34 (1.26, 1.55)	0.14 (0.07, 0.22)	0	0	0	1.34 (1.26, 1.50)	0.14 (0.08, 0.20)	0	0	0	1.34 (1.26, 1.50)	0.14 (0.08, 0.20)	0	0	0	1.34 (1.26, 1.50)	0.14 (0.08, 0.20)	0	0	0	1.34 (1.26, 1.50)	0.14 (0.08, 0.20)	0	0	0	0	0	0		
5	c2	1.38 (1.28, 1.62)	0.14 (0.07, 0.22)	0	0	0	1.38 (1.28, 1.55)	0.14 (0.08, 0.20)	0	0	0	1.38 (1.28, 1.55)	0.14 (0.08, 0.20)	0	0	0	1.38 (1.28, 1.55)	0.14 (0.08, 0.20)	0	0	0	1.38 (1.28, 1.55)	0.14 (0.08, 0.20)	0	0	0	0	0	0		
5.5	c2	1.43 (1.31, 1.68)	0.14 (0.07, 0.21)	0	0.2	0	1.42 (1.33, 1.59)	0.14 (0.08, 0.20)	0.1	0.5	0	1.42 (1.33, 1.59)	0.14 (0.08, 0.20)	0.6	1.1	0	1.42 (1.33, 1.55)	0.14 (0.08, 0.20)	0.6	1.1	0	1.42 (1.33, 1.55)	0.14 (0.08, 0.20)	0.6	1.1	0	0	0			
6	c2	1.47 (1.35, 1.80)	0.14 (0.07, 0.21)	0	0.2	0	1.46 (1.37, 1.74)	0.14 (0.08, 0.20)	0.2	0.6	0	1.46 (1.37, 1.74)	0.14 (0.08, 0.20)	0.7	1.3	0	1.46 (1.37, 1.74)	0.14 (0.08, 0.20)	0.7	1.3	0	1.46 (1.37, 1.74)	0.14 (0.08, 0.20)	0.7	1.3	0	0	0			
6.5	c2	1.51 (1.37, 1.90)	0.14 (0.07, 0.21)	0	0.3	0	1.50 (1.41, 1.82)	0.14 (0.08, 0.20)	0.3	0.7	0	1.50 (1.41, 1.82)	0.14 (0.08, 0.20)	0.4	0.9	0	1.50 (1.41, 1.82)	0.14 (0.08, 0.20)	0.4	0.9	0	1.50 (1.41, 1.82)	0.14 (0.08, 0.20)	0.4	0.9	0	0	0			
7	c2	1.56 (1.41, 1.92)	0.14 (0.06, 0.23)	0	0.3	0	1.55 (1.44, 1.75)	0.14 (0.08, 0.21)	0.4	0.7	0	1.55 (1.44, 1.75)	0.14 (0.08, 0.21)	0.5	1.0	0	1.55 (1.44, 1.75)	0.14 (0.08, 0.21)	0.5	1.0	0	1.55 (1.44, 1.75)	0.14 (0.08, 0.21)	0.5	1.0	0	0	0			
7.5	c2	1.56 (1.42, 1.89)	0.14 (0.06, 0.23)	0	0.4	0	1.55 (1.43, 1.78)	0.14 (0.08, 0.21)	0.4	0.7	0	1.55 (1.43, 1.78)	0.14 (0.08, 0.21)	0.5	1.0	0	1.55 (1.43, 1.78)	0.14 (0.08, 0.21)	0.5	1.0	0	1.55 (1.43, 1.78)	0.14 (0.08, 0.21)	0.5	1.0	0	0	0			
8	c3	1.60 (1.47, 1.99)	0.14 (0.07, 0.21)	0.4	1.9	0	1.59 (1.47, 1.82)	0.14 (0.07, 0.22)	0.5	1.2	0	1.59 (1.47, 1.82)	0.14 (0.07, 0.22)	0.6	1.3	0	1.59 (1.47, 1.82)	0.14 (0.07, 0.22)	0.6	1.3	0	1.59 (1.47, 1.82)	0.14 (0.07, 0.22)	0.6	1.3	0	0	0			
8.5	c3	1.64 (1.48, 2.13)	0.14 (0.07, 0.21)	0.5	2.8	0	1.63 (1.48, 1.93)	0.14 (0.07, 0.22)	0.6	1.4	0	1.63 (1.48, 1.93)	0.14 (0.07, 0.22)	0.7	1.5	0	1.63 (1.48, 1.93)	0.14 (0.07, 0.22)	0.7	1.5	0	1.63 (1.48, 1.93)	0.14 (0.07, 0.22)	0.7	1.5	0	0	0			
9	c3	1.68 (1.52, 2.13)	0.14 (0.07, 0.22)	0.5	3.2	0	1.68 (1.54, 1.98)	0.14 (0.07, 0.22)	0.7	1.6	0	1.68 (1.54, 1.98)	0.14 (0.07, 0.22)	0.8	1.7	0	1.68 (1.54, 1.98)	0.14 (0.07, 0.22)	0.8	1.7	0	1.68 (1.54, 1.98)	0.14 (0.07, 0.22)	0.8	1.7	0	0	0			
9.5	c3	1.72 (1.56, 2.23)	0.14 (0.07, 0.22)	0.5	3.6	0	1.71 (1.57, 2.04)	0.14 (0.08, 0.21)	0.8	1.8	0	1.71 (1.57, 2.04)	0.14 (0.08, 0.21)	0.9	1.9	0	1.71 (1.57, 2.04)	0.14 (0.08, 0.21)	0.9	1.9	0	1.71 (1.57, 2.04)	0.14 (0.08, 0.21)	0.9	1.9	0	0	0			
10	c3	1.77 (1.61, 2.33)	0.14 (0.07, 0.22)	0.5	4.0	0	1.76 (1.62, 2.10)	0.14 (0.08, 0.21)	0.9	2.0	0	1.76 (1.62, 2.10)	0.14 (0.08, 0.21)	1.0	2.1	0	1.76 (1.62, 2.10)	0.14 (0.08, 0.21)	1.0	2.1	0	1.76 (1.62, 2.10)	0.14 (0.08, 0.21)	1.0	2.1	0	0	0			
11	c3	1.83 (1.70, 2.55)	0.14 (0.07, 0.23)	0.5	4.4	0	1.82 (1.72, 2.30)	0.14 (0.08, 0.21)	1.0	2.2	0	1.82 (1.72, 2.30)	0.14 (0.08, 0.21)	1.1	2.3	0	1.82 (1.72, 2.30)	0.14 (0.08, 0.21)	1.1	2.3	0	1.82 (1.72, 2.30)	0.14 (0.08, 0.21)	1.1	2.3	0	0	0			
12	c3	1.89 (1.74, 2.66)	0.14 (0.07, 0.23)	0.5	4.8	0	1.87 (1.77, 2.45)	0.14 (0.08, 0.21)	1.1	2.4	0	1.87 (1.77, 2.45)	0.14 (0.08, 0.21)	1.2	2.5	0	1.87 (1.77, 2.45)	0.14 (0.08, 0.21)	1.2	2.5	0	1.87 (1.77, 2.45)	0.14 (0.08, 0.21)	1.2	2.5	0	0	0			
13	c3	2.04 (1.74, 2.66)	0.14 (0.06, 0.23)	0.5	5.2	0	2.03 (1.79, 2.75)	0.14 (0.09, 0.21)	1.2	2.6	0	2.03 (1.79, 2.75)	0.14 (0.09, 0.21)	1.3	2.7	0	2.03 (1.79, 2.75)	0.14 (0.09, 0.21)	1.3	2.7	0	2.03 (1.79, 2.75)	0.14 (0.09, 0.21)	1.3	2.7	0	0	0			
14	c3	2.17 (1.89, 3.21)	0.14 (0.07, 0.23)	0.5	5.6	0	2.14 (1.89, 2.72)	0.14 (0.08, 0.21)	1.3	2.8	0	2.14 (1.89, 2.72)	0.14 (0.08, 0.21)	1.4	2.9	0	2.14 (1.89, 2.72)	0.14 (0.08, 0.21)	1.4	2.9	0	2.14 (1.89, 2.72)	0.14 (0.08, 0.21)	1.4	2.9	0	0	0			
15	c3	2.28 (1.95, 4.47)	0.14 (0.07, 0.23)	0.5	6.0	0	2.27 (1.96, 3.17)	0.14 (0.09, 0.20)	1.4	3.0	0	2.27 (1.96, 3.17)	0.14 (0.09, 0.20)	1.5	3.1	0	2.27 (1.96, 3.17)	0.14 (0.09, 0.20)	1.5	3.1	0	2.27 (1.96, 3.17)	0.14 (0.09, 0.20)	1.5	3.1	0	0	0			
25	c4	3.33 (2.85, 4.47)	0.14 (0.07, 0.23)	0.5	10.0	0	3.32 (2.88, 4.08)	0.14 (0.09, 0.20)	1.0	10.0	0	3.32 (2.88, 4.08)	0.14 (0.09, 0.20)	1.0	10.0	0	3.32 (2.88, 4.08)	0.14 (0.09, 0.20)	1.0	10.0	0	3.32 (2.88, 4.08)	0.14 (0.09, 0.20)	1.0	10.0	0	0	0			
30	c5	4.07 (3.48, 5.38)	0.14 (0.06, 0.20)	0.0	10.0	3	4.03 (3.56, 5.74)	0.14 (0.09, 0.20)	0.0	10.0	38.7	4.03 (3.56, 5.74)	0.14 (0.09, 0.20)	0.0	10.0	38.8	4.03 (3.56, 5.74)	0.14 (0.09, 0.20)	0.0	10.0	38.8	4.03 (3.56, 5.74)	0.14 (0.09, 0.20)	0.0	10.0	38.8	23.4	0			
30	c6	4.06 (3.48, 5.49)	0.12 (0.05, 0.19)	0.0	10.0	3	4.02 (3.65, 5.00)	0.12 (0.07, 0.18)	0.0	10.0	81.7	4.02 (3.65, 5.00)	0.12 (0.07, 0.18)	0.0	10.0	81.7	4.02 (3.65, 5.00)	0.12 (0.07, 0.18)	0.0	10.0	81.7	4.02 (3.65, 5.00)	0.12 (0.07, 0.18)	0.0	10.0	81.7	0	0			
35	c5	4.76 (4.06, 6.46)	0.14 (0.06, 0.20)	0.0	10.0	5	4.71 (4.24, 6.46)	0.14 (0.09, 0.20)	0.0	10.0	48.6	4.71 (4.24, 6.46)	0.14 (0.09, 0.20)	0.0	10.0	48.6	4.71 (4.24, 6.46)	0.14 (0.09, 0.20)	0.0	10.0</											

Figure 2-4A shows the mean difference required to achieve 80% power with different sample sizes at P value thresholds of 0.05 (Figure 2-4A, upper left) and  $10^{-6}$  (Figure 2-4A, lower left). For example, a sample size of 100 cases and 100 controls results in over 80% power to detect a 4.5% mean difference (mean methOR = 1.32) in methylation at nominal significance ( $P = 0.05$ ). However, at genome-wide threshold  $P = 10^{-6}$  the same sample size gives over 80% power to detect a much larger effect size of 11% mean difference (mean methOR = 1.81).



The results of the Wilcoxon test are also shown in Table 2-1 to Table 2-4. I also performed power estimation under the one case – multiple controls scenario. I show the results from one case : two controls and one case : four controls study design (Table 2-2 and Table 2-3) and as expected power increases when the sample size of the control group increases. Compared to the t-test, the Wilcoxon test was outperformed in the

small sample size and with the smaller mean difference or methOR. Both tests easily reached 80% at genome-wide significance level with larger sample sizes.

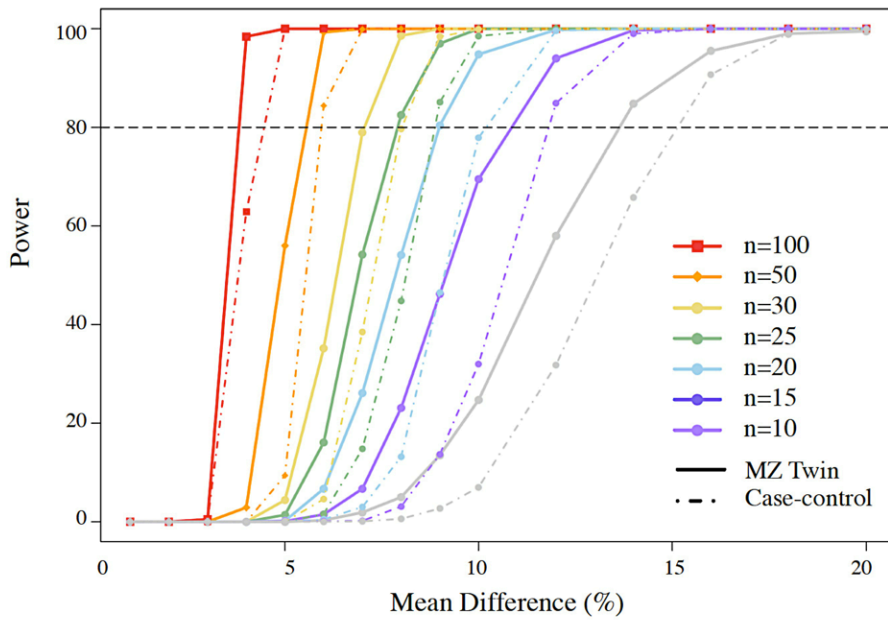
### **2.3.2 Power of case-control EWAS using methOR effect estimates with restrictions on the mean differences**

When the methOR was the selected effect size (Table 2-4, Figure 2-4B), to achieve 80% power to detect a methOR of 1.15 to 1.45 at  $P = 0.05$ , 50 pairs to 500 pairs of cases and controls were required. At a genome-wide significance level, samples greater than 100 pairs of cases and controls could detect methORs of 1.3 to 1.8 with over 80% power, but a smaller sample size of 50 had no power to detect effects within this range.

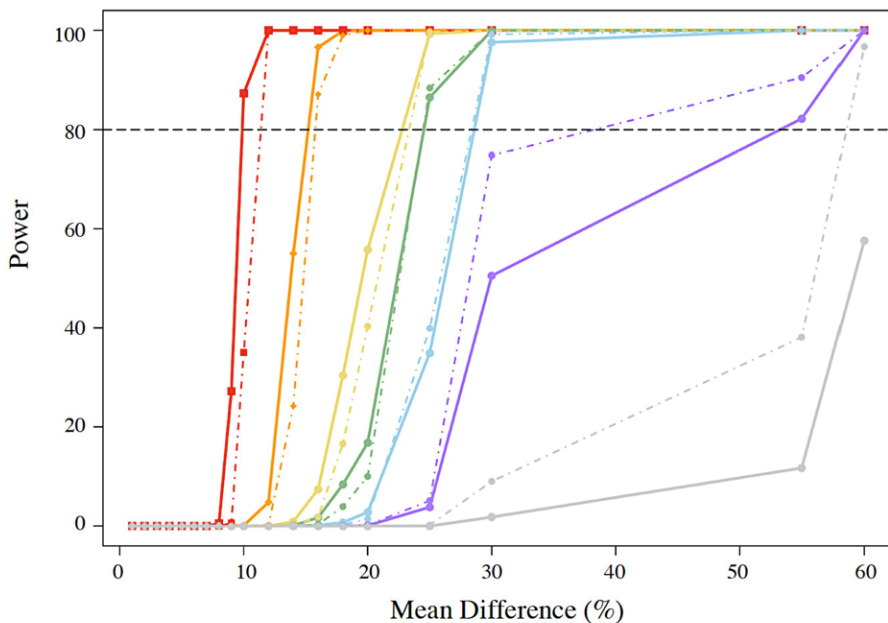
### **2.3.3 Power of discordant twin and case-control designs for small sample sizes and mean differences**

Figure 2-5 and Table 2-5 show the mean difference required to detect 80% power in smaller sample sizes of 10, 15, 20, 25, 30, 50 and 100 pairs of either MZ co-twins or cases and controls. Generally, the twin design outperformed the case-control design to reach 80% power at a significance level of 0.05. A 14% mean difference was necessary to attain 80% power with a small sample size of 10 pairs for both co-twins and case-controls at  $P < 0.05$  (Figure 2-5A). However, at genome-wide significance, at least 15 pairs of subjects were required to reach 80% power in both designs for the mean difference of 14% (Figure 2-5B), and over 20% mean difference was required to reach 80% power if the sample size was smaller than 30 pairs. However, these simulations were not designed for a formal comparison between case-control and twin power, because we assume that twins and case-control samples are equally well matched for factors that can influence differential methylation, including age, sex, and cohort effects, and unrelated samples are typically more heterogeneous than MZ twins.

### A. Twin Design (single-locus significance)



### B. Twin Design (genome-wide significance)



**Figure 2-5. Power of small-scale discordant twin (solid lines) and case-control (dashed lines) designs for a range of sample sizes and mean differences**

Results at a significance level of 0.05 (upper) and  $10^{-6}$  (lower). Each colour line represents the power curve under different sample size from 10 to 100. Solid-lines represent the power under discordant MZ twin design, and the dash line with the same colour represents the power under case-control design.







**Table 2-4. Power of large-scale case-control EWAS using methOR effects**

methOR/diff <sup>1</sup>	D <sup>2</sup>	N = 500															
		N = 100				N = 200				N = 500							
		Mean diff (range) <sup>3</sup>	T <sup>4</sup>	W <sup>5</sup>	P < 1×10 <sup>-6</sup>	Mean diff (range) <sup>3</sup>	T <sup>4</sup>	W <sup>5</sup>	P < 1×10 <sup>-6</sup>	Mean diff (range) <sup>3</sup>	T <sup>4</sup>	W <sup>5</sup>	P < 1×10 <sup>-6</sup>				
1.05, <1%	C1	0.79 (0.65, 0.98)	0	0	0	0.80 (0.67, 0.92)	0	0	0	0.80 (0.68, 0.92)	0	0	0	0.800 (0.7, 0.89)	0	0	0
1.10, <2%	C1	1.55 (1.31, 1.87)	0	0	0	1.55 (1.33, 1.76)	0	0	0	1.55 (1.39, 1.76)	0	0	0	1.55 (1.42, 1.69)	0	33.0	0
1.11, <2%	C1	1.69 (1.38, 1.98)	0	0	0	1.70 (1.49, 1.93)	0	0	0	1.70 (1.53, 1.88)	0	1.1	0	1.70 (1.58, 1.84)	39.7	53.4	0
1.15, <2.5%	C1	2.26 (1.94, 2.50)	0	0	0	2.27 (1.99, 2.50)	0	0.4	0	2.28 (2.07, 2.49)	0	13.5	0	2.27 (2.12, 2.44)	100	98.9	0
1.20, <3%	C2	2.92 (2.63, 3.00)	0	0.7	0	2.90 (2.64, 3.00)	0	5.8	0	2.92 (2.72, 3.00)	99.4	69.8	0	2.94 (2.78, 3.00)	100	100	0
1.25, <4%	C2	3.81 (3.37, 4.00)	0	3.5	0	3.83 (3.45, 3.99)	37.7	65.7	0	3.87 (3.58, 3.99)	100	100	0	3.88 (3.67, 3.99)	100	100	0
1.30, <5%	C2	4.55 (3.80, 5.00)	0.2	17.2	0	4.56 (4.08, 4.95)	99.9	94.2	0	4.55 (4.26, 4.88)	100	100	0	4.56 (4.24, 4.83)	100	100	96.5
1.35, <5.5%	C2	5.17 (4.51, 5.50)	15.3	45.5	0	5.20 (4.65, 5.50)	100	99.5	0	5.21 (4.83, 5.50)	100	100	0	5.22 (4.96, 5.49)	100	100	100
1.40, <6%	C2	5.70 (4.89, 6.00)	74.5	70.8	0	5.77 (5.29, 5.99)	100	100	0	5.81 (5.25, 5.99)	100	100	0	5.83 (5.54, 5.99)	100	100	100
1.45, <6.5%	C2	6.24 (5.58, 6.50)	98.8	88.8	0	6.31 (5.70, 6.50)	100	100	0	6.36 (5.93, 6.50)	100	100	0.5	6.39 (6.04, 6.50)	100	100	100
1.50, <7.5%	C2	6.97 (5.87, 7.50)	100	97.2	0	7.00 (6.24, 7.49)	100	100	0	7.01 (6.51, 7.46)	100	100	61.0	7.01 (6.74, 7.36)	100	100	100
1.60, <9%	C3	8.61 (7.35, 9.00)	100	100	0	8.70 (2.95, 8.99)	100	100	0	8.73 (8.19, 8.99)	100	100	100	8.77 (8.43, 8.99)	100	100	100
1.70, <10%	C3	9.64 (8.43, 10.00)	100	100	0	9.73 (8.93, 9.99)	100	100	23.9	9.79 (9.08, 9.99)	100	100	100	9.84 (9.38, 9.99)	100	100	100
1.80, <11%	C3	10.62 (9.50, 11.00)	100	100	0	10.72 (9.90, 10.99)	100	100	93.8	10.79 (10.20, 10.99)	100	100	100	10.84 (10.40, 10.99)	100	100	100
1.90, <12%	C3	11.59 (10.24, 12.00)	100	100	0.3	11.69 (10.79, 11.99)	100	100	99.8	11.76 (11.12, 11.99)	100	100	100	11.83 (11.46, 11.99)	100	100	100
2.00, <14%	C3	12.76 (11.12, 13.97)	100	100	2.5	12.80 (11.71, 13.70)	100	100	100	12.79 (12.00, 13.63)	100	100	100	12.78 (12.22, 13.30)	100	100	100

<sup>1</sup>methOR/diff: Mean methylation odds ratio and set value of mean difference between cases and controls; <sup>2</sup>D: Case distributions of sample draw, C1 to C3 corresponding to case distribution 1 to case distribution 3; <sup>3</sup>Mean diff: Mean methylation difference between cases and controls; <sup>4</sup>T: two sample t-test; <sup>5</sup>W: Wilcoxon rank-sum test

**Table 2-5. Power of EWAS twin and case-control designs**

Diff <sup>1</sup>	N = 10						N = 20						N = 25						N = 30						N = 50					
	TWIN <sup>2</sup>		CACO <sup>3</sup>		TWIN <sup>2</sup>		CACO <sup>3</sup>		TWIN <sup>2</sup>		CACO <sup>3</sup>		TWIN <sup>2</sup>		CACO <sup>3</sup>		TWIN <sup>2</sup>		CACO <sup>3</sup>		TWIN <sup>2</sup>		CACO <sup>3</sup>		TWIN <sup>2</sup>		CACO <sup>3</sup>			
	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>	P<0.05	P<10 <sup>-6</sup>		
1%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
2%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
3%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
4%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
5%	0	0	0	0	0.1	0	0	0.2	0	0	0	0	1.4	0	0	0	0	4.4	0	0	0	0	0	2.9	0	0	0	0		
6%	0.3	0	0	0	1.5	0	0.1	6.7	0	0.5	0	16.1	0	1.6	0	1.6	0	35.2	0	4.6	0	4.6	0	56.0	0	9.4	0	0		
7%	1.9	0	0.1	0	6.7	0	0.2	26.1	0	3.0	0	54.2	0	14.8	0	14.8	0	79.0	0	38.5	0	38.5	0	100	0	99.9	0	0		
8%	5.0	0	0.6	0	23.1	0	3.1	54.1	0	13.2	0	82.5	0	44.8	0	44.8	0	98.6	0	79.7	0	79.7	0	100	0	100	0	0	0	
9%	13.5	0	2.7	0	46.2	0	13.7	80.4	0	46.5	0	97	0	85.1	0	85.1	0	100	0	98.4	0	98.4	0	100	0.1	100	0	0	0	
10%	24.7	0	7.0	0	69.5	0	32.0	94.8	0	77.9	0	100	0	98.5	0	98.5	0	100	0	100	0	100	0	100	0.1	100	0	0	0	
12%	58.0	0	31.8	0	94.0	0	84.9	99.8	0	99.6	0	100	0	100	0	100	0	100	0	100	0	100	0	100	4.8	100	0.2	0	0	
14%	84.8	0	65.8	0	99.7	0	99.0	100	0	100	0	100	0	100	0	100	0	100	0	100	0	100	0	100	55.0	100	24.2	0	0	
16%	95.5	0	90.7	0	100	0	100	0	100	0	100	0.2	100	0	100	0	100	0.3	100	7.4	100	1.8	100	99.9	100	99.0	0	0	0	
18%	99.0	0	99.0	0	100	0	100	0	100	0	100	0.7	100	0.3	100	8.4	100	3.9	100	30.4	100	16.6	100	100	100	100	100	0	0	
20%	99.5	0	99.9	0	100	0.1	100	0.2	100	1.4	100	16.8	100	10.0	100	16.8	100	10.0	100	55.8	100	40.3	100	100	100	100	100	0	0	
25%	100	0	100	0.2	100	3.8	100	5.1	100	34.9	100	86.5	100	88.4	100	86.5	100	88.4	100	99.3	100	99.7	100	100	100	100	100	0	0	
30%	100	0.5	100	2.3	100	25.2	100	38.7	100	86.8	100	95.2	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	
30%	100	1.8	100	9.0	100	50.5	100	74.8	100	97.6	100	99.2	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	
55%	100	11.7	100	38.1	100	82.2	100	90.5	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	
60%	100	57.6	100	96.7	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	

<sup>1</sup>Diff: Mean methylation difference between affected and unaffected individuals; <sup>2</sup>Twin: discordant MZ twin design (paired t-test); <sup>3</sup>CACO: case-control design (two-sample t-test).

### 2.3.4 Sample size required for 80% power in discordant twin and case-control designs for a range of mean differences

Table 2-6 shows the sample size required to reach 80% power with a mean difference of 7%, 8%, 9%, 10% and 15% for both case-controls and co-twins at a significance level of 0.05 and  $10^{-6}$ . The correlation coefficients with larger sample sizes were consistently lower than the set correlation requirement (range 0.19 to 0.62) in the study, therefore simulations of the mean difference less than 7% were not considered because the required sample size was greater than 200 pairs of twins. Overall, compared to the case-controls, the co-twins required a smaller sample size to reach 80% power. In general, sample sizes required to detect larger mean differences were similar between co-twins and case-controls, but quite different for smaller mean differences. For example, to detect a mean difference of 7% at genome-wide significance 178 pairs of MZ twins were required, while 211 case-control pairs were need, that is 66 additional individuals for the unrelated design. Similar sample sizes were found using the nonparametric Wilcoxon rank-sum test.

**Table 2-6. Sample size required for 80% power in EWAS twin and case-control designs**

Diff <sup>1</sup>	Twin				Case-control			
	P < 0.05		P < $1 \times 10^{-6}$		P < 0.05		P < $1 \times 10^{-6}$	
	t-test <sup>2</sup>	Wilcox <sup>3</sup>	t-test <sup>2</sup>	Wilcox <sup>3</sup>	t-test <sup>4</sup>	Wilcox <sup>5</sup>	t-test <sup>4</sup>	Wilcox <sup>5</sup>
7%	30	30	178	178	37	37	211	211
8%	25	25	145	149	30	30	169	169
9%	20	20	117	117	24	24	137	137
10%	17	18	98	102	20	21	112	110
11%	15	15	81	83	17	18	96	95
12%	13	13	71	71	15	16	80	80
13%	11	12	63	69	13	13	70	70
14%	10	11	55	62	11	13	61	63
15%	9	10	50	57	10	11	54	57

<sup>1</sup>Diff: Mean methylation difference between affected and unaffected individuals; <sup>2</sup>t-test: Paired t-test; <sup>3</sup>Wilcox: Wilcoxon signed-rank test; <sup>4</sup>t-test: Two sample t-test; <sup>5</sup>Wilcox: Wilcoxon rank-sum test

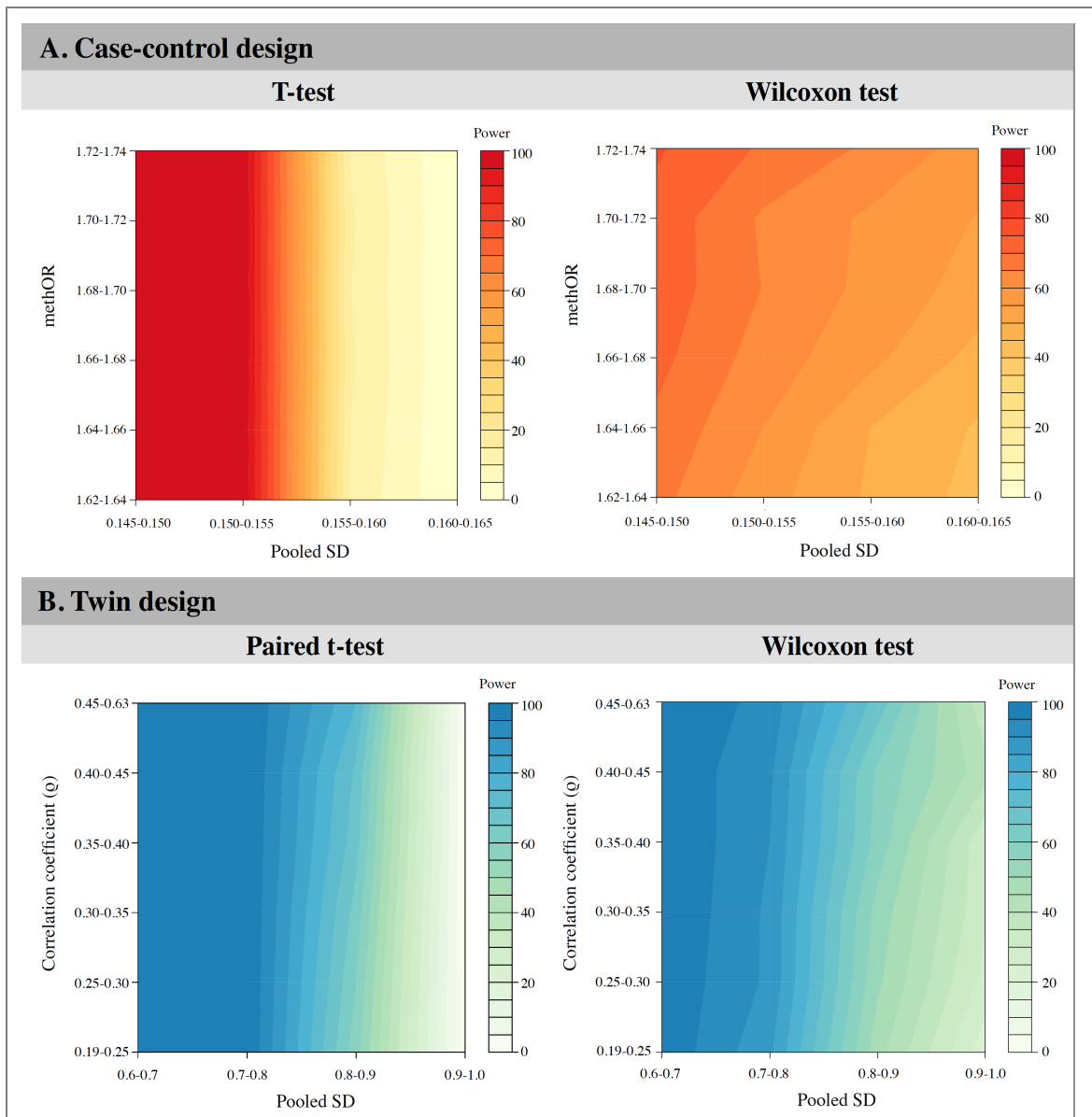
### 2.3.5 Methylation variance and methOR affect power under the same mean difference in the case-control design

For the case-control design, changes in the pooled SD and methOR can influence power under the same mean difference. In general, the smaller pooled SD and bigger methOR gives more significant P values under both the two-sample t-test and Wilcoxon rank sum test. In addition, permutations with smaller pooled SD tend to have higher methOR.

To find how these two factors influenced power, permutations with 10% methylation mean difference and equal variances were selected and the power was detected at  $P = 0.05$ . I categorized the pooled SD into 4 groups: 0.145-0.150, 0.150-0.155, 0.155-0.160, and 0.160-0.165, and the methOR was categorized into 6 groups: 1.62-1.64, 1.64-1.66, 1.66-1.68, 1.68-1.70, 1.70-1.72, and 1.72-1.74. Figure 2-6 shows the relationship between power, pooled SD, and methOR for a set mean difference of 10% as calculated by the t-test (Figure 2-6A, left) and Wilcoxon test (Figure 2-6A, right). Under the t-test, the pooled SD immensely influences power such that greater pooled SD will lead to lower power despite methOR differences. In comparison, power of the t-test can be dramatically affected from the pooled SD (power range: 0%-100%), while under the same parameters the Wilcoxon test gives more similar power (power range: 42%-76%). Both the pooled SD and methOR still do have an influence on power estimated using the Wilcoxon test, such that greatest power can be achieved with smaller pooled SD and at highest methOR.

To explore the influence of methylation variance on power, I selected permutations with the same 20 cases and 20 controls at a 10% methylation mean difference, but only using simulations where the variance of cases was not equal to that of the controls. The major difference between the equal and unequal variance t-test is in the denominator of the t-statistic and the degrees of freedom. In the unequal variance test, the variance between groups was calculated by:

$$SD_{\text{Case-Control}} = \sqrt{\frac{SD_{\text{Case}}^2}{N_{\text{Case}}} + \frac{SD_{\text{Control}}^2}{N_{\text{Control}}}}$$

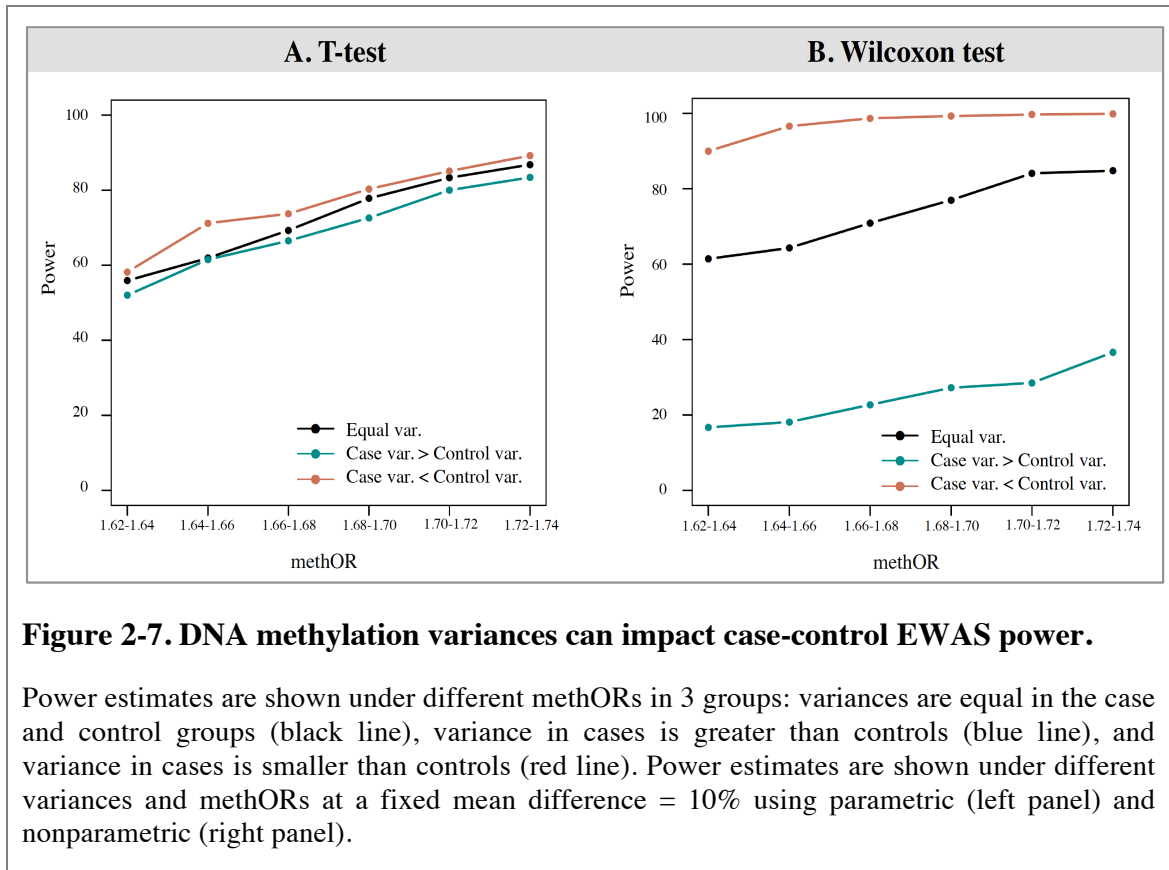


**Figure 2-6. DNA methylation variance and correlation can impact EWAS power**

Case-control power estimates (upper panel) are shown under different pooled SD and methORs at a fixed mean difference= 10% using parametric (left panel) and nonparametric (right panel) test statistics. MZ twin power estimates (lower panel) are shown under different pooled SD and correlation coefficients at a fixed mean difference= 9% using parametric (left panel) and nonparametric (right panel) test statistics.

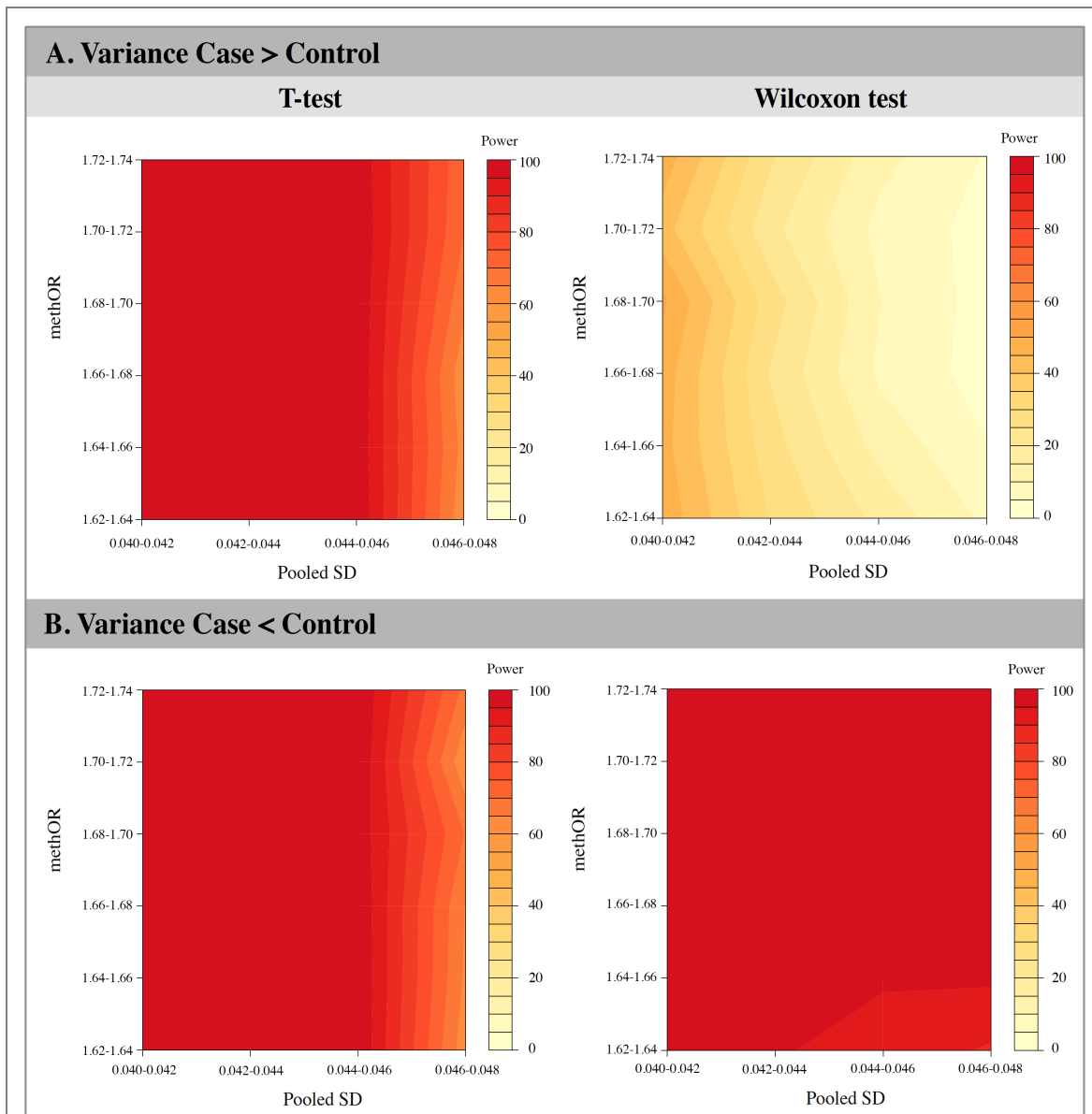
Power estimates in the unequal variance case-control simulations were categorized using this pooled standard deviation into 4 groups (0.040-0.042, 0.042-0.044, 0.044-0.046, 0.046-0.048) and using methOR into 6 groups (1.62-1.64, 1.64-1.66, 1.66-1.68, 1.68-1.70, 1.70-1.72, and 1.72-1.74). Furthermore, we also considered which group (cases or controls) had the greater variance. That is, either the variance in cases was greater than that in controls, or the variance in cases was smaller than that in controls.

Compared to the simulations with equal variances between the groups, the power estimations from the unequal variance results were quite similar for the t-test (Figure 2-7A). It is easier to reach greater power when the variance in the cases is smaller than that in controls, and a more distinct pattern is found using the Wilcoxon test under the same parameter settings (Figure 2-7B).



Similar to the equal variance results, the methOR and pooled variance impact power (Figure 2-8). This result also highlights the importance of choosing the appropriate analytical method across the equal variance t-test, unequal variance t-test, and the Wilcoxon test.





**Figure 2-8. Unequal DNA methylation variances in cases and controls can impact EWAS power**

Power estimates are shown under different variances and methORs at a fixed mean difference = 10% using parametric (left panel) and nonparametric (right panel) test statistics. Power estimates were categorized depends on the variances are greater in cases (upper panel) or controls (lower panel).

### **2.3.6 DNA methylation variance and twin correlation can influence power in the EWAS twin design**

In twin design, the pooled SD, methOR and the correlation between the co-twins, also influenced P values. Smaller pooled SD and greater intra-pair correlation can result in greater power, for a set mean difference. Both the methOR and Spearman's correlation coefficients are negatively correlated with pooled SD.

To better demonstrate the relationship between pooled SD, within-twin correlation and power, only permutations with a set methylation difference of 9% and methOR range between 1.65 and 1.70 were selected, and power was calculated at the significance level of 0.05. The pooled SD was categorized into 4 groups: 0.6-0.7, 0.7-0.8, 0.8-0.9, and 0.9-1.0, and the correlation was categorized into 6 groups: 0.19-0.25, 0.25-0.30, 0.30-0.35, 0.35-0.40, 0.40-0.45, and 0.45-0.62. Figure 2-6B shows the EWAS power under different combinations of pooled SD and correlation coefficients. Under the t-test (Figure 2-6B, left), the smallest pooled SD gives the greatest power, and under the same-pooled SD, permutations with higher correlation give greater power. Similar effects can be seen under the Wilcoxon test (Figure 2-6B, right). Compared to the t-test, the Wilcoxon test gives slightly lower power with moderate pooled SD, however, the t-test cannot provide sufficient power under the bigger pooled SD whereas Wilcoxon test can outperform it.

## 2.4 Discussion

I have estimated EWAS power under different sample sizes and effect sizes for the case-control design and disease discordant MZ twin designs. I found that compared to the t-test, the Wilcoxon rank sum test showed improved power if the sample size was big or the effect size was small. However, under large effect sizes (mean difference and sample size to exceed 8% and 25 pairs), there was minimal difference between the two statistical methods. When comparing designs, a MZ twins were slightly more powered, because a lower sample size was required in discordant MZ twin design to achieve 80% power compared to case-controls.

Currently, power estimation relies on the magnitude of the effect size between the comparison groups, for example the methylation difference between cases and controls at one CpG position. Some EWAS studies remove probes with < 5% methylation difference as the first step of the selection criteria, as potential background noise. The problem is that this step potentially removes informative probes. One should not only consider the methylation difference between cases and controls, but also the methylation variance between groups, and the methylation odds ratio. Furthermore, a great majority of probes on the Illumina 450k array have small variance, which suggests that large differences cannot be easily observed, and in MZ twins, who share similar methylation levels, these differences would be even smaller.

I think that our estimates of permutation-based power give more conservative values than power estimated using the traditional formula:

$$N = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)}{(\mu_1 - \mu_2)^2}$$

With a sample size of 50 pairs and a 14% mean difference our simulation results achieved 70.6% power, whereas 90.2% was estimated using the formula in the twin design. Similarly a 17% mean difference in 50 case-control pairs resulted in 25.8% power in our study, and 42.3% power from the formula. This suggests that the power estimates are too optimistic when estimated by mean difference and SD alone. When I compared our simulation results with those from Rakyan et al. (Rakyan, Down, *et al.*, 2011), using similar methOR of 1.49 and md = 7.2% with 200 pairs of cases and

controls, our estimates show 61% and 67% power using the t-test and Wilcoxon test, but only 16% power was reached using the Wald's test in their study. This divergence could be explained by the different composition of the distributions used, as we use a single beta distribution and Rakyan *et al.* (Rakyan, Down, *et al.*, 2011) used two combined beta distributions. Compared to the previous two studies that proposed that the methylation distribution at one locus is composed either by a Uniform-Normal mixture structure (S. Wang, 2011) or single or combined Beta distributions (Rakyan, Down, *et al.*, 2011), we assumed that both cases and controls at one CpG locus followed a single Beta distribution and the controls remained unmethylated. This assumption was based on the findings of a current methylation dataset of 172 healthy female subjects (J. T. Bell *et al.*, 2012) measured by the Illumina 27K array, where 69% (N = 24641) of the autosomal CpGs were unmethylated and the majority of distributions on each locus followed single beta distribution with small standard deviation (85% of probes with SD < 0.05). Therefore, the power estimation based on a single beta distribution is perhaps more appropriate for most DMPs measured on the Illumina 27k array.

Two types of effect size were explored: the mean difference and the methOR, which are often used as a measure of methylation effect size in published studies. However, because hemi-methylated probes tended towards larger standard deviation that could result in power bias, the methOR is considered a better indicator (Rakyan, Down, *et al.*, 2011). In our simulations, I found that both measures are useful: the mean difference is suitable if differences between groups are minor, whereas the methOR can better illustrate the association between cases and controls at one locus with large methylation difference. Ideally, if one could take into account both of these effect sizes that would result in a more precise estimate of effect. For example, if I draw identical sample sizes of 178 cases from Case 1 (Mean = 0.21, SD = 0.14) and Case 2 (Mean = 0.25, SD = 0.14) and select permutations with a fixed methOR = 1.2 compared to 178 controls, the results show that the two permutation groups have the same methOR, mean difference, and sample size for 5000 permutations, however I found that 94.4% of permutations selected from Case 2 distributions were significant at P value = 0.05 using t-test, whereas only 80.5% were significant if selected from Case 1 distribution. Similar results are shown in other simulation results, with the same methOR and mean difference criteria; a smaller sample size was required when drawn from Case 3

distribution compared to Case 2 distribution. This difference is likely due to the variance in methylation: mean differences from Case 3 distribution have a narrower pooled SD range, but higher values than those selected from Case 2. Therefore, I further examined the effect of pooled SD on power under the two designs and two statistical methods, and found that statistical power is highly influenced by the pooled SD under all scenarios. In addition, I found that the methOR and Spearman's correlation coefficients also affect power under different circumstances. For example, higher between-group correlations that may indicate a heritable methylation locus, and at such loci there will be greater power to detect differential methylation effects under the discordant twin design. I found that the methylation difference and methOR alone might not be sufficient for power estimation, because the pooled SD can differ greatly with the same mean difference and methOR, and the results give diverging power estimates. Furthermore, power seemed to shift dramatically under the t-test, but was relatively more stable under the Wilcoxon test with smaller sample sizes ( $N = 20$  in the simulations).

Heritability has been considered as a factor for power estimation in the study, that is, the methylation correlation between the MZ twin pair. However, there was no evidence of a significant pattern between Spearman's correlation coefficient within the co-twins and the significance level. To test whether the range of correlation coefficients affect our simulation results using the same sample size, power estimations were performed on simulations with a much narrower correlation coefficient of 0.29 to 0.31. The co-twins with narrower correlation coefficients required slightly smaller sample sizes to reach 80% power for a given mean difference. For example, with the same 11% mean difference, 81 pairs of twins were required to detect power at 80% under the wider correlation range, whereas a slightly lower 78 pairs were sufficient for twins with the narrower range using t-test.

There are limitations from the study assumptions. Firstly, we assumed the methylation changes are causal to disease and longitudinally stable. Current studies have found that methylation could be dynamic and may require a longitudinal study to characterize the temporal methylation pattern. Secondly, we assumed the case-controls are closely matched as MZ twins, and this will bias the case-control sample towards homogeneity and advantages power estimates for the case-control. Finally, we assumed there is no

genetic effect on methylation levels in the study, but current studies have identified underlying genetic effect on the methylation levels, so this will also impact our results in case-control design.

The major application of this power study is to help design an EWAS study. For example, one can get the effect size from the previous references or the pilot study, then estimate the required sample size for a study. In addition, our results can assist in interpreting the impact of EWAS findings. Our power estimates are potentially applicable to other methylation or gene expression data under the same assumption of data distribution.

## **2.5 Conclusions**

In summary, I provide power and sample size estimation for both case-control and disease discordant MZ twin studies under various effect sizes. More complex simulations are needed to incorporate co-methylation patterns and factors such as age and environment. Furthermore, our results are also relevant to the power and sample size estimation for other case-control or twin epidemiology studies of finite quantitative genomic data.

# Materials and Methods: An Overview of the Methylation Datasets and Quality Control Procedure

---

This chapter provides an overview of the methylation datasets that I have used, a brief description of the Illumina methylation arrays, and a description of the standard quality control procedures that I adopted using an example dataset. Lastly, because I used twins, I also investigated DNA methylation heritability in the example dataset.

---

## 3.1 Methylation datasets

I have used five methylation datasets in this thesis: three derived from blood (Dataset 1-3), one from adipose tissue (Dataset 4), and one from skin tissue (Dataset 5, Table 3-1). The first dataset from blood was generated on the Illumina 27k array (Bibikova *et al.*, 2009) and was previously published (Rakyan *et al.*, 2010; J. T. Bell *et al.*, 2012) and the other two blood datasets were based on the Illumina 450k array (Bibikova *et al.*, 2011; Dedeurwaerder *et al.*, 2011) and have not yet been published. The skin and adipose data were generated in a subset of individuals from the MuTHER study (Grundberg *et al.*, 2012). The adipose methylation dataset (Dataset 4) has been published (Grundberg *et al.*, 2013) while the skin methylation dataset (Dataset 5) has not yet been published. All subjects were twins from the TwinsUK cohort (Moayyeri *et al.*, 2013), and in some cases data were only available for one twin per pair.

**Table 3-1. Summary of the five methylation datasets**

Dataset	Array	Tissue	Subjects <sup>1</sup>	Mean age (range)	Chapters <sup>2</sup>
1	27k	Blood	172	57 (32, 80)	4
2	450k	Blood	449	55 (28, 78)	4, 5, 6
3	450k	Blood	50	55 (39, 72)	5
4	450k	Adipose	648	59 (39, 85)	4, 6
5	450k	Skin	469	59 (39, 85)	4

<sup>1</sup>Total subjects, including some with cancer or other diseases (such as type 2 diabetes);

<sup>2</sup>Chapters that have used these datasets: Chapter 4 (methylation and age); Chapter 5 (methylation and birth weight); Chapter 6 (methylation and smoking).

Some datasets are used in the multiple chapters either using the full dataset or a subset. Further details about the samples will be given in each chapter where they are used (Table 3-1).

## 3.2 Illumina Infinium HumanMethylation assays

In this section, I briefly discuss the two Illumina Infinium HumanMethylation arrays (Illumina Infinium HumanMethylation27k and Illumina Infinium HumanMethylation 450k; Illumina Inc, San Diego, CA) used in this study. Prior to running the arrays, DNA samples should be first bisulfite converted. This changes the unmethylated cytosines into uracils while the methylated cytosines remain unchanged.

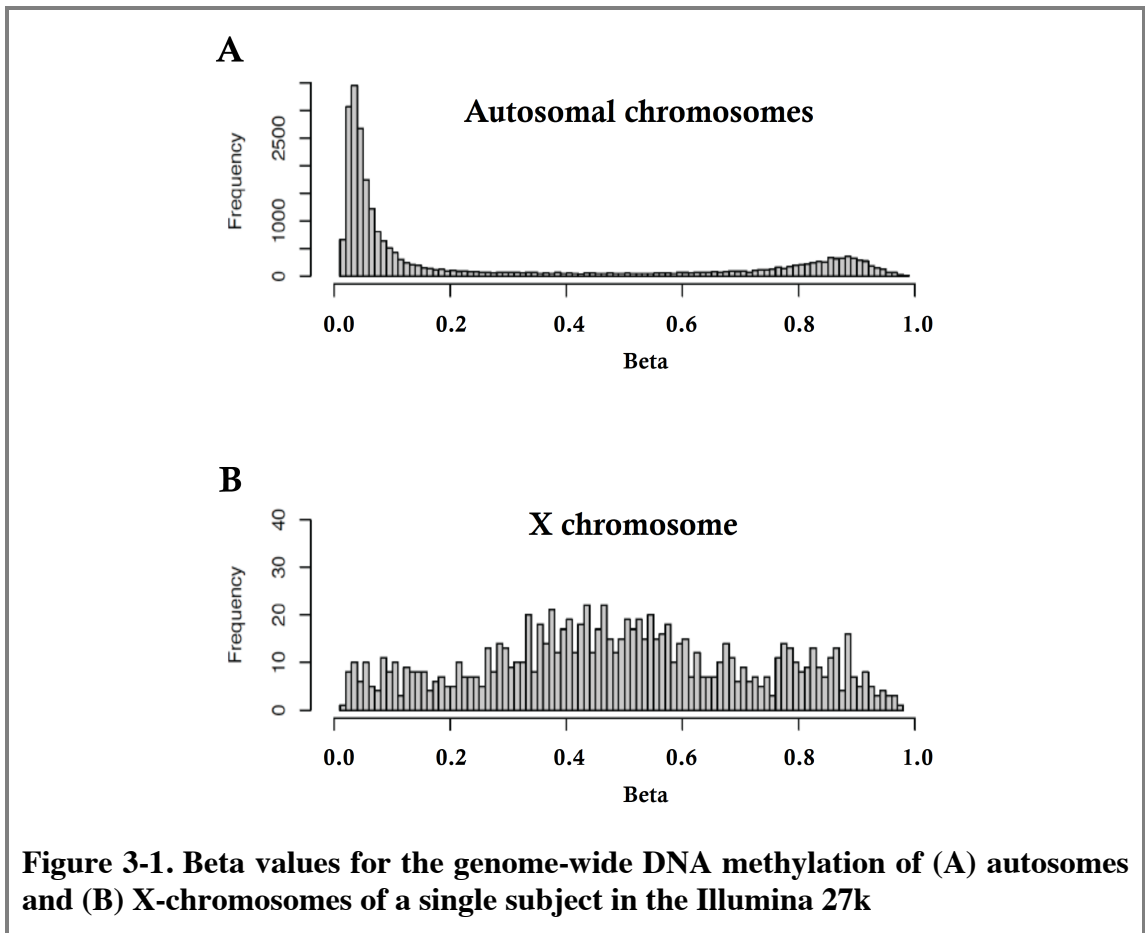
### 3.2.1 Illumina 27k array

On the Illumina 27k array each CpG locus is represented by two probes of length 50 base pairs (bp), representing unmethylated and methylated bead types. The unmethylated probe perfectly matches the unmethylated version of the CpG, while the methylated probe matches the methylated version of the CpG. The DNA sample is first bisulfite converted and denatured into single strands, and hybridized to the array by annealing to the bead probes. Only the perfectly matched one will continue to base extension with hapten labelled dideoxynucleotides, and only ddCTP is labelled with biotin. The labelled ddNTPs (2',3' dideoxynucleotides, including ddGTP, ddATP, ddTTP, and ddCTP) will be fluorescence stained multiple times to distinguish the two bead types, and the chip will be scanned for intensities. More details will be discussed in the following Illumina 450k section.



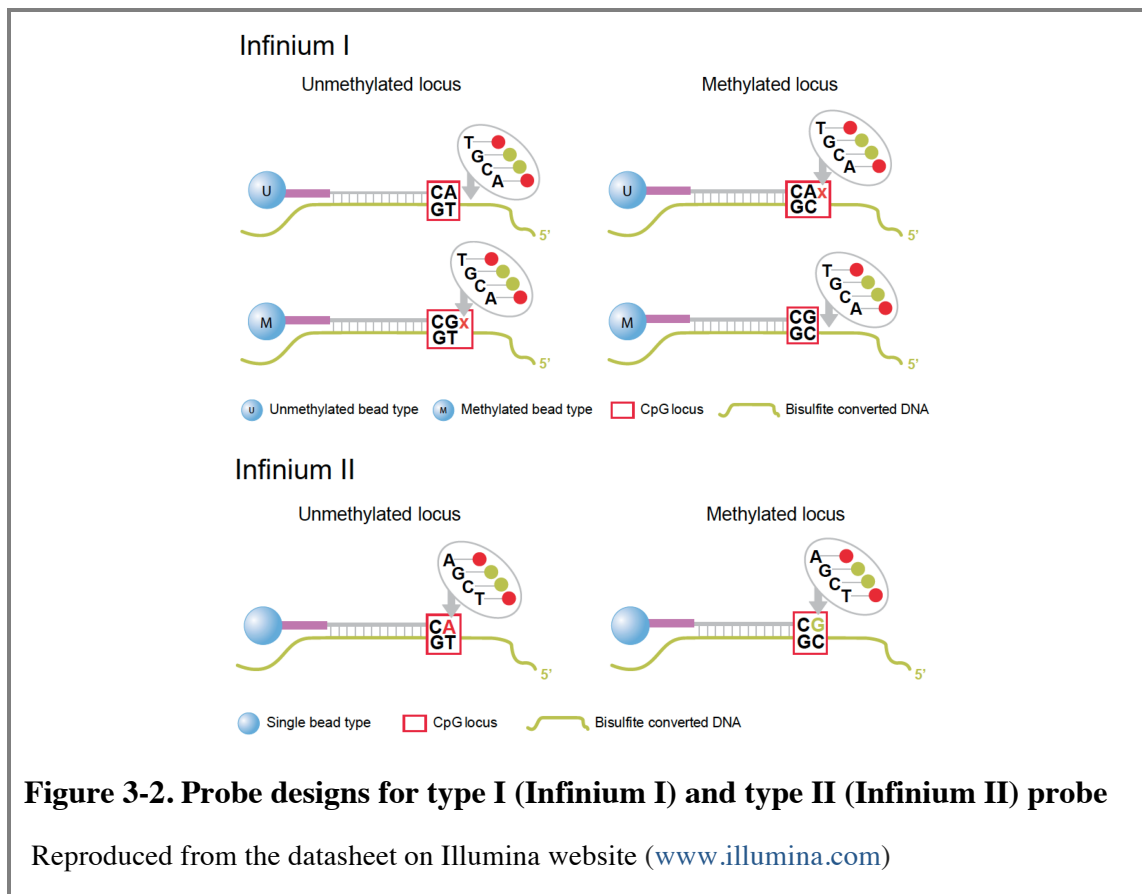
The intensity of the unmethylated and methylated beads is detected. The degree of methylation at a single locus is measured as a beta value, which is calculated as the ratio of intensity of methylated beads over the total intensity at the locus (sum of the methylated and unmethylated intensities). The range of beta is between 0 (unmethylated) and 1 (methylated).

A single Beadchip covers 12 samples and assays 27,578 CpG dinucleotides in the promoters of 14,495 genes. I used 26,690 probes that mapped to the genome within 2 mismatches and in the promoters of ~13,000 genes using Ensembl annotations (J. T. Bell *et al.*, 2011). Of the 26,690 probes 25,690 are located on autosomes and 1,000 probes on the X-chromosome. For a single subject in dataset 1, the distribution of beta values on the autosomes and X-chromosome shows distinct shapes (Table 3-1). The distribution for autosomes is concentrated around methylated and unmethylated signals (approximately 70% of probes were unmethylated). The pattern on the X chromosome should be hemi-methylated due to the random inactivation of one X-chromosome in females, which is methylated. The observed X-chromosome distribution indeed has a peak at hemi-methylated probes (beta of 0.5), along with some unmethylated and methylated probe signal in the tails, that might be due to certain stretches of the X-chromosome containing genes that have similar inheritance mechanisms as the autosomal genes, known as pseudoautosomal regions (PAR). In my analysis using Illumina 27k, I excluded sex chromosomes and missing probes. In the end, I used 24,641 probes for downstream analysis.



### 3.2.2 Illumina 450k array

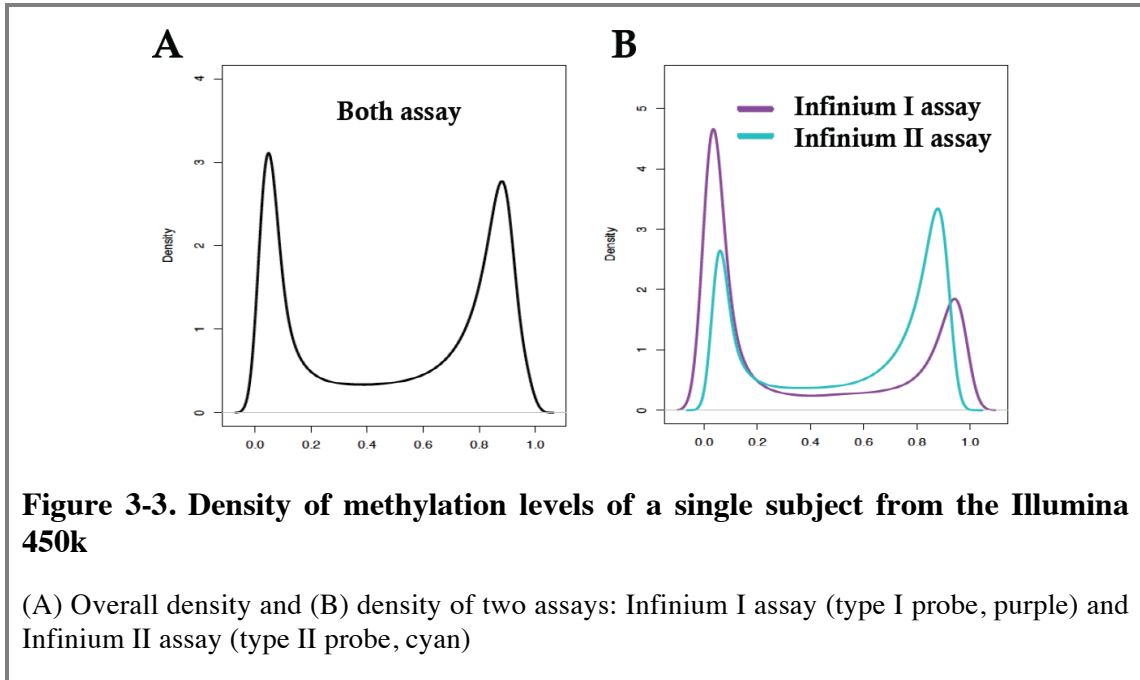
The Illumina 450k array offers a wider coverage of genome-wide methylation. The array consists of two types of probes: the Illumina 27k Infinium I probes (type I probes), and the new Illumina 450k Infinium II probes (type II probes). Unlike the two-bead design for type I probes, only one bead per locus is required to detect the methylation levels in the type II probes. Here, if the target CpG site is unmethylated (base A or T), the red fluorescent labels will be detected, and a green fluorescent label will only be detected at a methylated locus. On type I probes the fluorescent labels are extended on the next bp past the CpG site (51<sup>st</sup> bp), and on type II probes the fluorescent labels are extended on the last bp (50<sup>th</sup> bp, the actual C/G position). The design of the two probe types and the detection process is shown in the Figure 3-2.



The Illumina 450k array assays 485,836 sites and 27.9% of these are detected by type I probes. There should be a 98% concordance rate of methylation levels detected between two assays, according to the Illumina technical report (Bibikova *et al.*, 2011; Dedeurwaerder *et al.*, 2011). In a single subject, the methylation distribution is composed by two asymmetric beta distributions (Figure 3-3A). It is complicated by the fact that type I probes have a wider methylation distribution compared to type II probes (Figure 3-3B) suggesting that type II probes might not be as sensitive as type I probes for extreme values (Dedeurwaerder *et al.*, 2011). It has been suggested that the two probe types distributions should be made comparable, prior to between subject normalization, because without such adjustment, there would be an enrichment bias on type I probes to have a more extreme rankings than type II probes (Teschendorff *et al.*, 2013).

To correct for this bias, two methods are currently commonly used, implemented in the Subset-quantile Within Array Normalization method (SWAN; (Maksimovic *et al.*, 2012)) and Beta Mixture Quantile dilation (BMIQ; (Teschendorff *et al.*, 2013)). The SWAN method is based on a quantile-normalization of a subset of type I and II probes

and adjusting the intensities of the remaining probes based on the subset probes. The BMIQ method is based on the interpolation of the distribution of type II probes to type I probe.



### 3.3 Quality control of the Illumina array data

Many approaches have been proposed in analysis of Illumina array methylation data (Bock, 2012; Warden *et al.*, 2013; Morris *et al.*, 2014). The approach I used involves the following steps: (1) the identification of probes that map incorrectly or to multiple locations in the reference sequence; (2) the identification of individuals who are outliers, that is, their methylation profiles are not consistent with the methylation density in the remainder of the sample; (3) the identification of batch effects and covariates that affect methylation levels in the sample; and (4) the application of data normalization and adjustment for covariates. I explain these in detail below, with an example using Dataset 1 and 2.

#### 3.3.1 Identification of probes mapping to multiple locations

There were two things to consider when deciding whether probe should be excluded prior to even exploring the actual beta values. First, all probes were designed to anneal

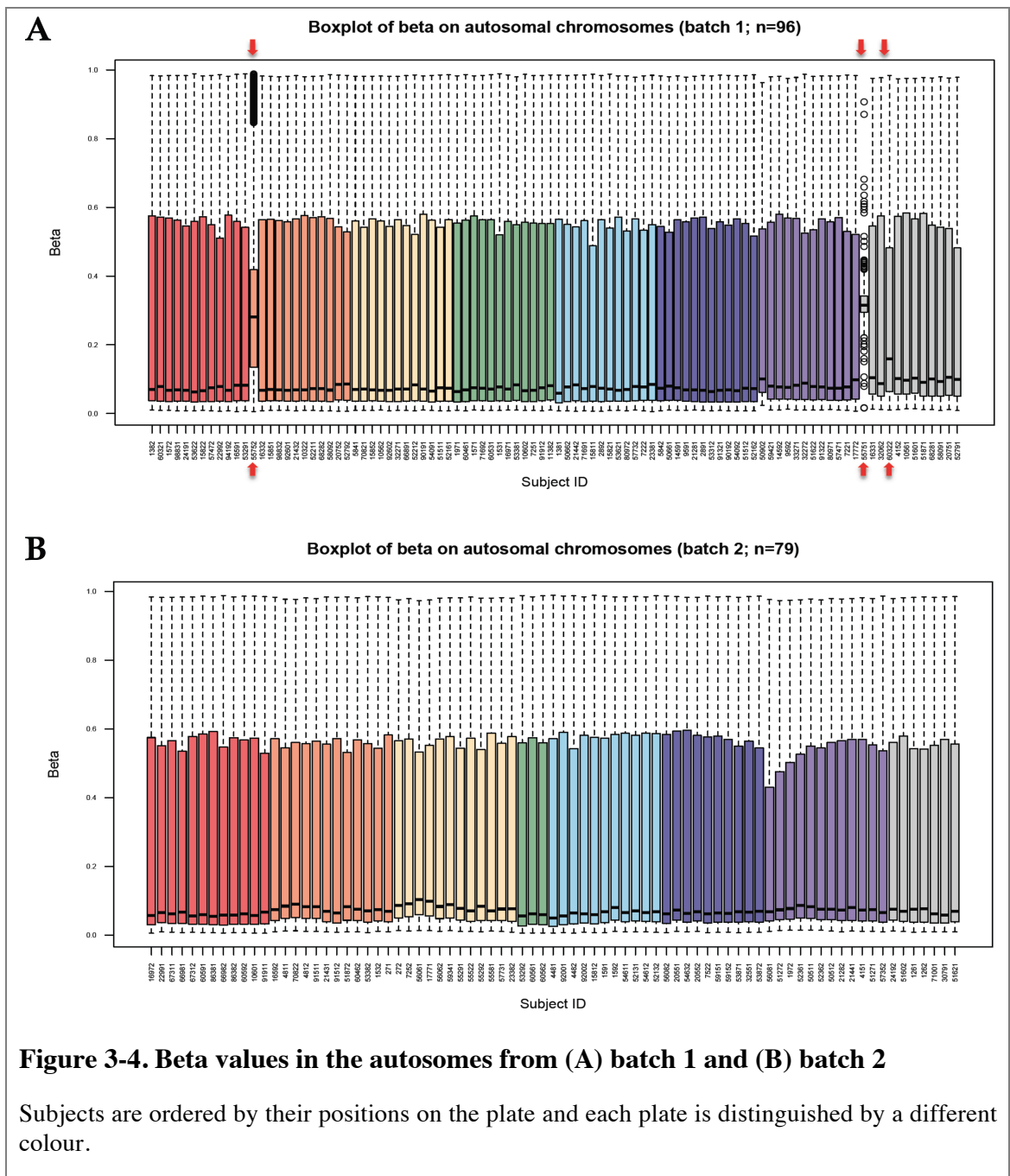
to 50 base pairs at a single location in the genome. We checked for probes that mapped to multiple locations within 2 mismatches (based on the hg18 for the Illumina 27k and hg18 and hg19 for the Illumina 450k). Due to this reason, 888 probes were removed in the Illumina 27k (J. T. Bell *et al.*, 2011), and 17,651 probes were excluded from the Illumina 450k array (these works were done by Idil Yet and Dr. Wei Yuan in the department).

Secondly, I carefully considered probes that may contain genetic variants (SNPs or CNVs) that might impact hybridization due to incomplete annealing. Two studies have suggested that probes with SNPs located on the CpG site affect the quality of hybridization and influence methylation levels (Price *et al.*, 2013; Naeem *et al.*, 2014). This might be problematic since  $\sim 1/3$  of all probes on the 450k array overlap known SNPs. However, the SNP in the probe should only impact DNA methylation if the individual has a “non-reference allele” (the SNPs designed on the Illumina probe). In this thesis, all these probes were kept from the analyses, and only probes in the top results were assessed to see if they contained a SNP in the probe or at the CpG site.

### **3.3.2 Identification of outliers**

In the following 3 sections, I use the Illumina 27k data as an example for the outlier identification and batch effect identification. This step involved visual inspection of plots to identify outliers, which show extremely skewed or abnormal distribution of beta values, or have high rates of missing data. First, a boxplot and density plot of genome-wide beta values within a subject could identify subjects with different methylation patterns relative to the remainder of the sample. Second, the pair-wise methylation correlation matrix was computed and visualized using a heatmap and a dendrogram, to identify outliers that were dissimilar from the other individuals. Figure 3-4 shows the beta values of the autosomes from Dataset 1. The methylation levels of subjects were detected at two time points, which I defined as batch 1 data (N = 96, including outliers) and batch 2 data (N = 79, no outliers). By ordering the subjects according to their position on the plate, differences and batch effects could be observed. For example, from batch 1, subject 55751 had much missing data while subjects 55752 and 60322 showed different means and ranges compared to other subjects (red arrows, Figure 3-

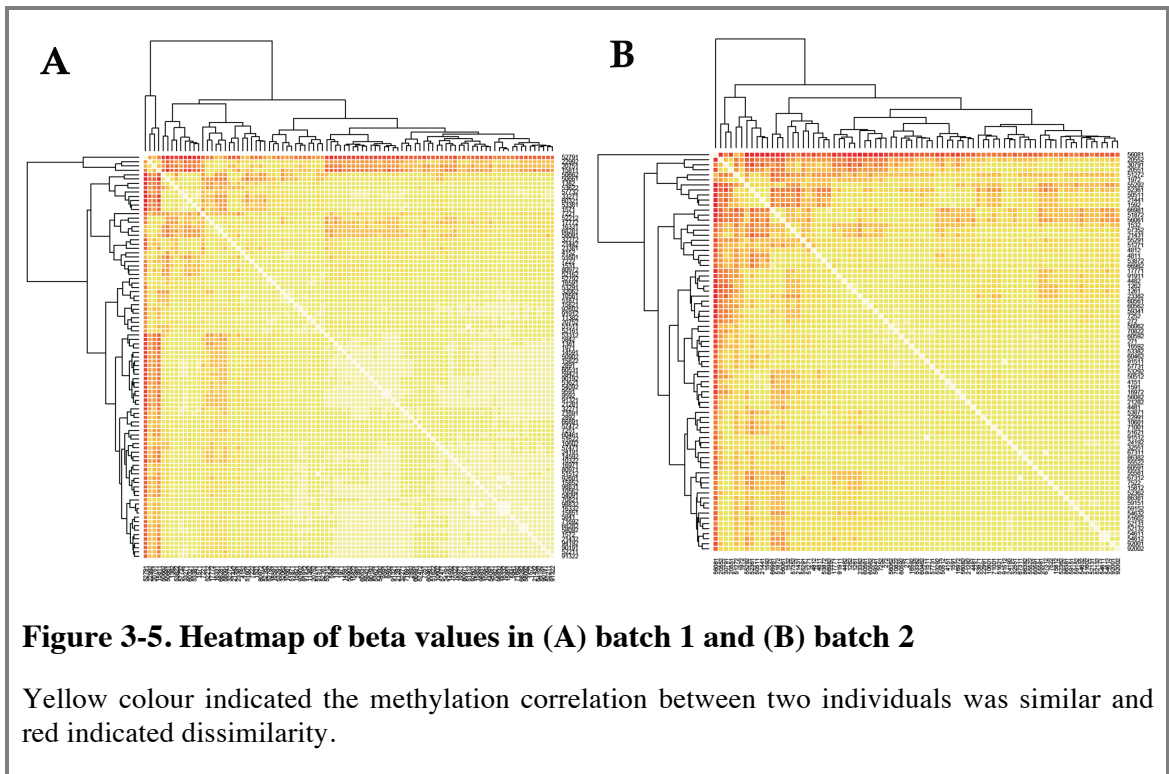
4A). I considered these three subjects as outliers. Batch 2 showed a more similar distribution for each subject (Figure 3-4B).



**Figure 3-4. Beta values in the autosomes from (A) batch 1 and (B) batch 2**

Subjects are ordered by their positions on the plate and each plate is distinguished by a different colour.

I then computed the pairwise correlations in genome-wide beta values using Pearson's correlation. Figure 3-5 shows the heatmap of the correlation matrix for all pairs of subjects after excluding the three outliers in batch 1, which shows minimal structure in the data.

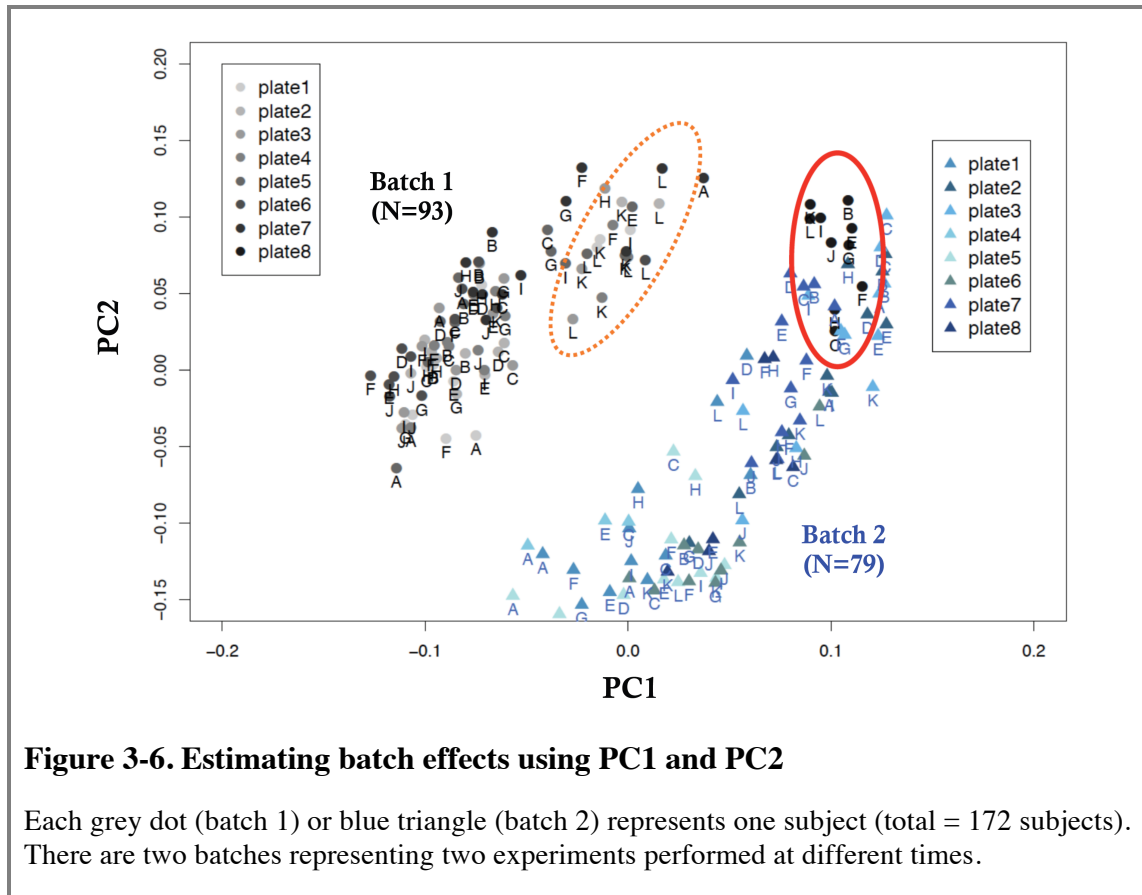


### 3.3.3 Identification of the batch effects and covariates

I applied a quantile-quantile normalization (across subjects) to the probe-level data to ensure that the data followed the Normal distribution. Subsequently, I used principal component analysis (PCA) to find potential batch effects. The PCA is a method of reducing multidimensional data by transforming the variables into a smaller number of uncorrelated (orthogonal) variables or Principal Components (PCs). The first PC captures the majority of the variance in the data. To determine potential covariates (biological, such as age, zygosity; or systematic, such as batch effect), I correlate each potential covariate with the first several PCs and assessed the significance of the correlation. I then selected the nominally significant variables for inclusion as covariates in downstream analyses.

An example of this approach in Dataset 1 is shown in Figure 3-6. Here, PC1 and PC2 explained 19.31% and 16.56% of the genome-wide methylation variance. Each plate has 12 positions (lettered A-L) on the Illumina 27k and within batch 1, the L position from several plates clustered together to indicate a ‘position on the plate’ effect (orange circle), since the assumption was that the positions on plate should not adhere to any pattern. There is also a ‘plate’ effect (red circle) as a plate from batch 1 clustered with

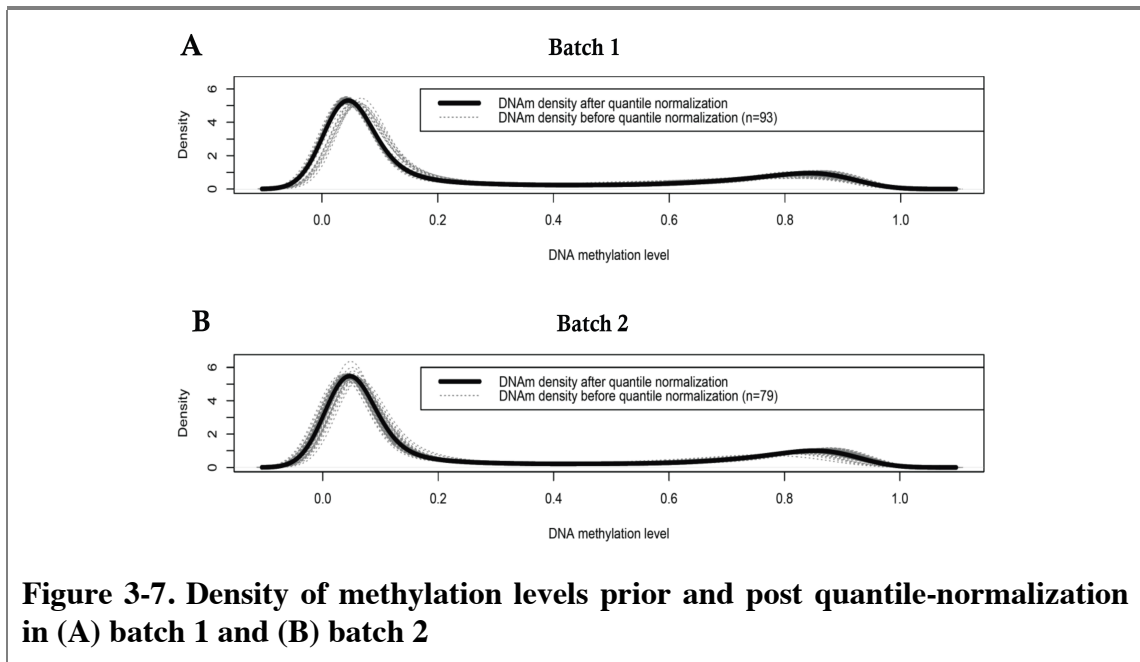
batch 2. These two effects are due to systematic technical noise introduced during the experiment. Therefore, we considered batch, plate effect, and position on the plate effect as covariates in all analyses. Similar findings were observed for the Illumina 450k data therefore these technical covariates were adjusted for in all downstream analyses for both the Illumina 27k and on the Illumina 450k array datasets.



### 3.3.4 Data normalization and adjusting for batch effects

A normalization step was necessary to make subjects comparable because each subject may have a different methylation distribution. Therefore, I applied quantile-normalization (within subjects across probes) to the raw data. Figure 3-7 shows the methylation density of subjects from the two batches, prior (grey dashes) and post quantile-normalization (black lines). The quantile-normalization has been widely applied in gene expression array data analysis (Bolstad *et al.*, 2003). Consequently, subjects in the same array would have similar distribution of methylation density.





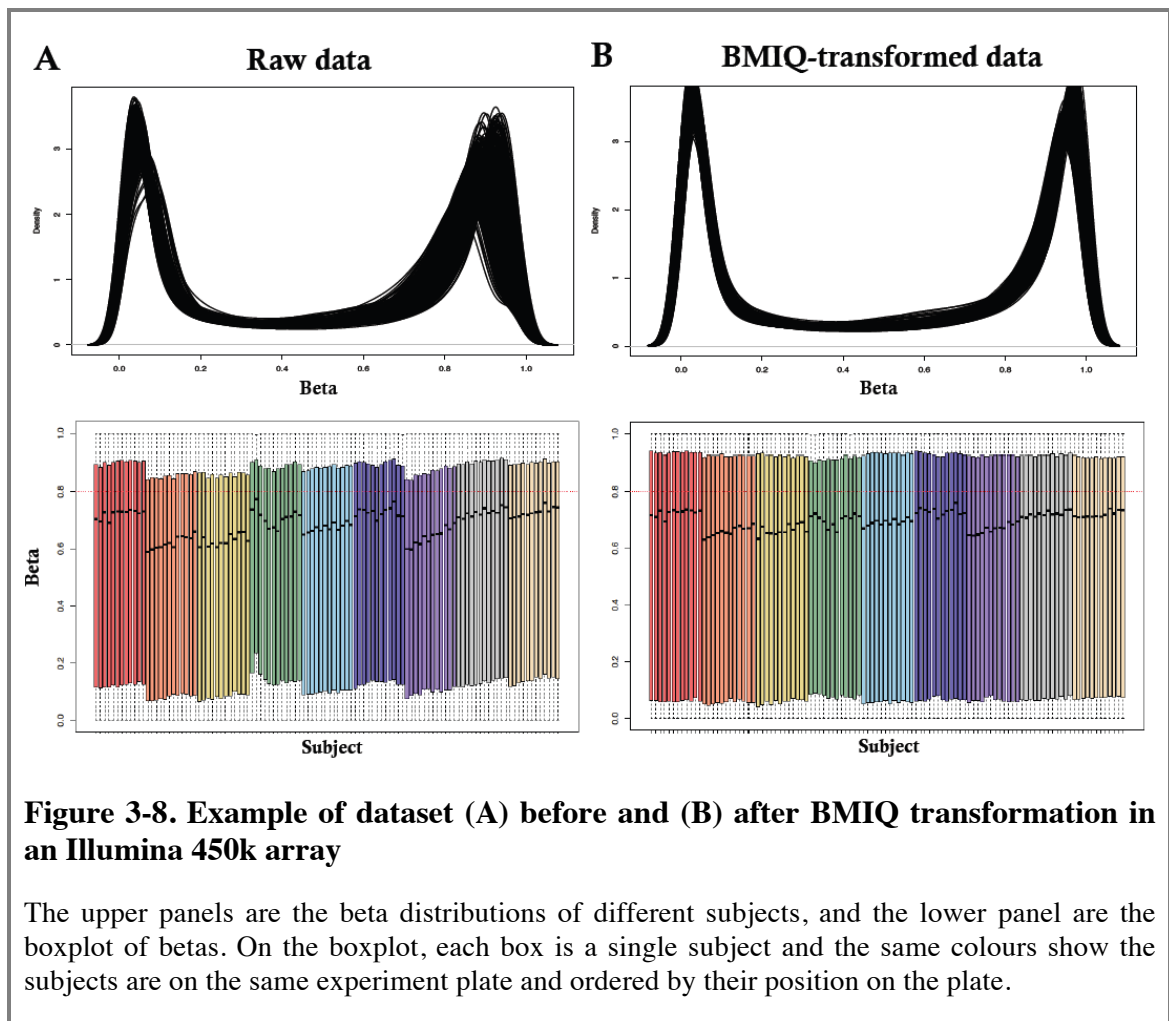
**Figure 3-7. Density of methylation levels prior and post quantile-normalization in (A) batch 1 and (B) batch 2**

However, the direct quantile-normalization used in Illumina 27k data would be less suited to normalize the two probe types with different distributions on the Illumina 450k data. To overcome the issue with probe types and their distribution, several R packages are available for array data normalization, such as the BMIQ (Teschendorff *et al.*, 2013), SWAN (in R “minfi” package) (Maksimovic *et al.*, 2012), and DASEN (in R “watermelon” package) (Pidsley *et al.*, 2013).

To make type II probes more comparable to the type I probes, I used BMIQ (Teschendorff *et al.*, 2013) to first normalize the methylation betas. The purpose of BMIQ was to transform the beta distribution of type II probes to fit that of type I probes. Both types of probes were categorized into unmethylated, hemi-methylated, and methylated, and the type II probes were transformed to fit the quantiles of type I probes using the inverse of the cumulative beta distributions in each category.

Figure 3- 8 shows the data distribution before and after the BMIQ normalization. Before normalization, the methylation distribution showed variation between subjects, and the median differed across plates (Figure 3-8A), and by position on the plate (subjects at the beginning of the plate have lower methylation levels compared to those at the end of the plate). After BMIQ transformation, the type II probe distribution was transformed, and the methylation range between subjects became more similar but their medians were relatively unchanged (Figure 3-8B).

For downstream analysis using linear regression, the methylation levels on each probe were further quantile-quantile normalized to fit a standard normal distribution.

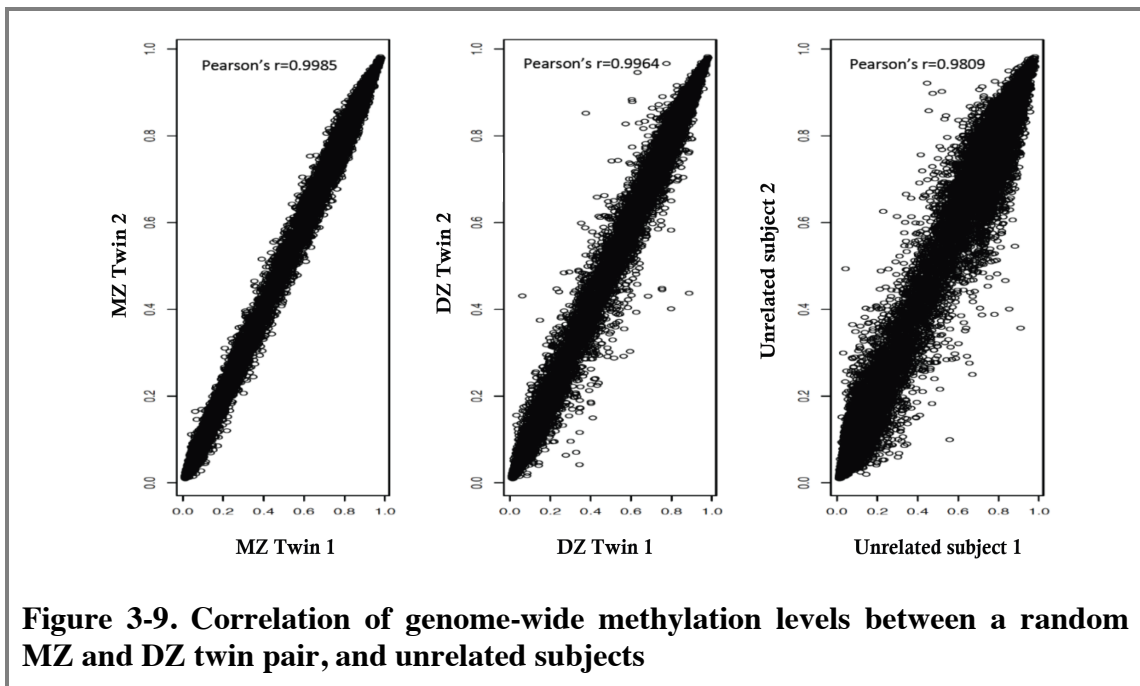


### 3.4 Methylation heritability and patterns in twins

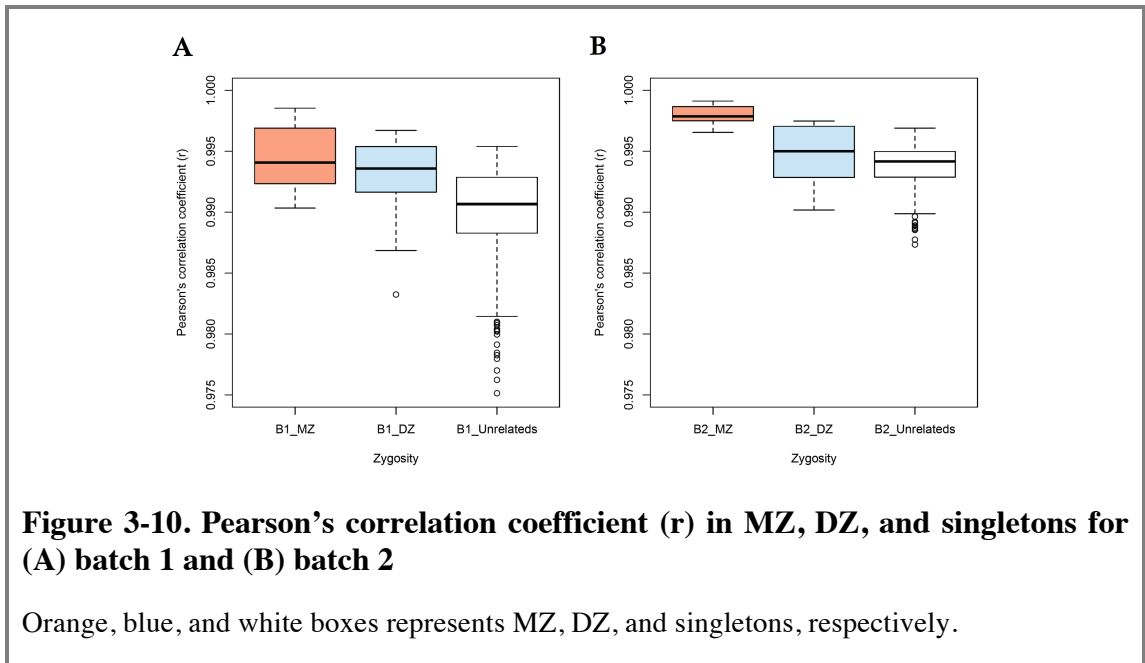
Because the datasets contained twins, I wanted to explore structure in the data and evidence for heritability in DNA methylation itself. Previous studies have shown evidence for methylation heritability and genetic influences at a subset of probes (Z. A. Kaminsky *et al.*, 2009; J. T. Bell *et al.*, 2012; Grundberg *et al.*, 2013). To study this, I categorized the subjects in the Illumina 27k dataset (Dataset 1) into two batches based on the date that their methylation levels were measured. Batch 1 included 12 MZ pairs and 17 DZ pairs and batch 2 included 9 MZ pairs and 14 DZ pairs.

### 3.4.1 DNA Methylation Patterns in Twins

I found that the methylation patterns of MZ twins were more correlated than those of DZ twins, and that twins shared more similar patterns compared to unrelated subjects in both batches. This result was consistent with a previous study (Z. A. Kaminsky *et al.*, 2009). Figure 3-9 shows the correlation between a random pair of MZs, DZs, and two unrelated subjects. Despite the fact that correlation coefficients were high between unrelated subjects, they were still lower than those within twin pairs.



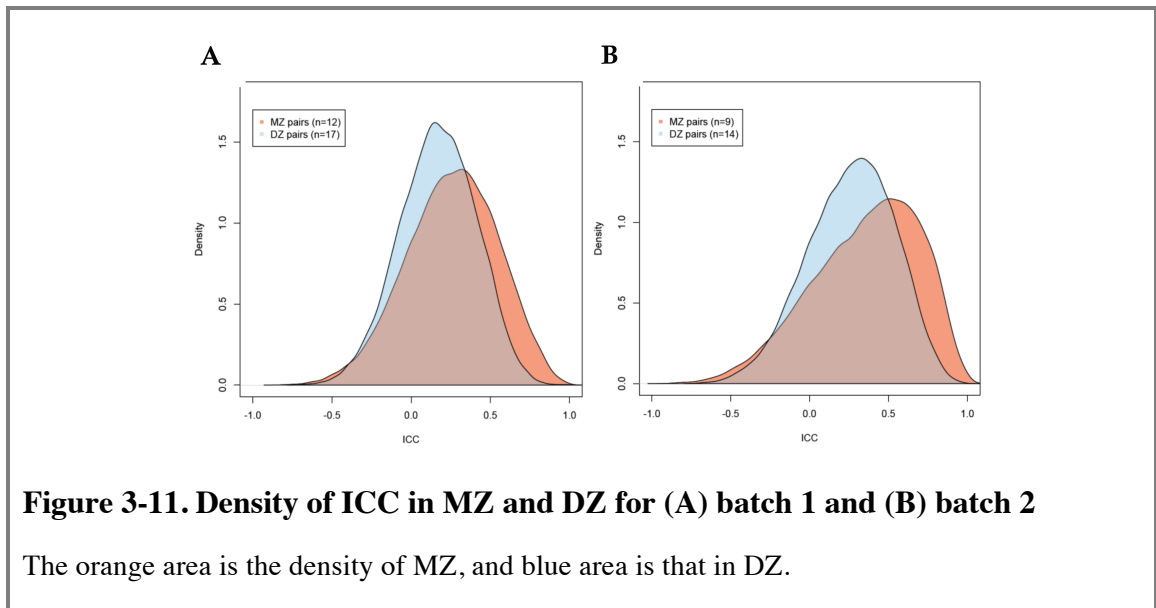
I then calculated the Pearson's correlation coefficients for all twins and between unrelated singletons from the two batches. Again, the methylation levels within MZ pairs were more similar than those within DZ twins, and twins were more correlated than pairs of unrelated individuals (Figure 3-10). The correlation patterns are similar in both batches (Figure 3-10A: batch1; Figure 3-10B: batch2). This trend indicated that certain methylation regions could be heritable, and therefore family and zygosity should be included as covariates in the study of twins.



### 3.4.2 Heritability of DNA methylation

I calculated the intra-class correlation coefficients (ICC) to compare the genome-wide methylation patterns within and between MZ and DZ twins. The intra-class correlation is used for data with paired structure and can account for the resemblance of units in the same group (Fisher, 1954). The heritability of methylation is calculated as twice the difference of ICC in MZ and DZ.

The genome-wide heritability was estimated using the ICC for both batches (Figure 3-11). The MZ twins had a higher ICC than DZ twins, and the average heritability estimated over the 24,641 CpG sites was 0.176 in batch 1 and 0.188 in batch 2.



In summary, the methylation patterns in twins showed more similarity than those in unrelated individuals, as previously observed (Z. A. Kaminsky *et al.*, 2009). I also found that methylation was more similar in MZs than DZs, and twins had more similar methylation levels than unrelated subjects. On average, my estimate of genome-wide heritability of methylation was 0.18 using Illumina 27k data. This value is higher than the Kaminsky *et al.* estimate of 0.014 using Human 12K CpG island microarray in whole blood (Z. A. Kaminsky *et al.*, 2009). In a recent study using 344,092 probes in Illumina 450k array, Grundberg *et al.* used 97 MZ and 162 DZ twin pairs in adipose tissue, and reported that methylation heritability was 0.19 (Grundberg *et al.*, 2013). An alternative method of calculating the heritability is to use the OpenMX tool (Boker *et al.*, 2011). Using a subset of 349,237 probes in 94 MZ and 25 DZ pairs, a mean heritability rate of 0.22 was found in blood sample (Table 1, Dataset 2; work done by Juan Edgar Castillo-Fernandez in the department). Only focused on a subset probes from the Illumina 450k array that overlapped with the Illumina 27k, there was a slightly lower heritability rate of 0.2, which was approaching my value of 0.18. It could be that the genome-wide heritability in the Illumina 27k array is lower than that estimated on the Illumina 450k array because more probes are located on CpG islands, which are generally unmethylated and have low levels of methylation variability.

# DNA Methylation Associates with Age and Age-Related Phenotypes

---

The aim of this chapter is to identify the DNA methylation changes that associate with age and age-related phenotypes. This chapter is divided into two sections based on the array platforms used. In the first section, I will present my analysis of age-related differential methylation sites of Illumina 27k data in whole blood. The second section extends my results to the Illumina 450k platform in whole blood, and two other tissues (adipose and skin), to identify tissue-shared effects.

Part of this work has been published as a research article in PLoS Genetics (J. T. Bell *et al.*, 2012). To gain a comprehensive overview on this subject, I prepared a review of the published epigenome-wide association scans (EWAS) of age and age-related complex human traits, and this has been published as a review article in Epigenomics (Tsai *et al.*, 2012).

---

## 4.1 Introduction

Age-related methylation changes have not been fully characterized. An initial study observed that older monozygotic (MZ) twins aged 50 years old have more methylation differences than younger twins aged 3 years old (Fraga *et al.*, 2005). On the other hand, methylation might not dramatically change over time, as a study looking at methylation levels in two age groups (26 and 68 years old) across three chromosomes of different tissues reported only a 0.275% difference between the groups (Eckhardt *et al.*, 2006). Since the same subjects were not traced longitudinally, both findings remain somewhat

inconclusive. One longitudinal study sampled the same subjects, at ages 5 and 10 years old, and showed that longitudinal changes in DNA methylation exist, as methylation levels of three genes (*DRD4*, *SERT*, *MAOA*) varied among subjects and were unstable over time (Wong *et al.*, 2010).

Moving away from candidate gene studies, recent studies have increasingly focused on EWAS, and continue to report strong correlations between methylation and age - firstly, using the Illumina GoldenGate methylation (Boks *et al.*, 2009), and from 2010 onwards, using the Illumina 27k platform (Gibbs *et al.*, 2010). I have compared age-related findings from the Illumina 27k platform across several studies (see Table 1 from (Tsai *et al.*, 2012)). Some of the first studies that used the Illumina 27k to investigate a-DMPs include two genome-wide studies published in 2010 (Rakyan *et al.*, 2010; Teschendorff *et al.*, 2010). One of these found 231 CpG sites hyper-methylated with age and 147 CpG sites hypo-methylated with age in 31 twin pairs and 31 singletons in whole blood samples (Rakyan *et al.*, 2010). The authors replicated the a-DMPs in CD4+ and CD14+ cells, suggesting that changes had occurred in the precursor hematopoietic stem cells, prior to their divergence into the myeloid and lymphoid lineages. They also reported that a-DMPs were enriched at bivalent chromatin domain promoters at the precursor stage. The second study assessed methylation patterns of 491 subjects at varying stages of ovarian cancer and identified 69 CpGs hyper-methylated with ageing using different cell lines (Teschendorff *et al.*, 2010), suggesting that the ageing process might silence genes that were suppressed in stem cells and contribute to carcinogenesis. A hypothesis was proposed that age-related methylation changes might induce cells into a 'stem-like' state and predisposes an individual towards carcinogenesis (Rakyan *et al.*, 2010; Teschendorff *et al.*, 2010).

Subsequent studies have explored age effects on the Illumina 450k array and more CpG sites differentially methylated with age have been identified (Horvath, 2013; Martino *et al.*, 2013; Florath *et al.*, 2014; Tserel *et al.*, 2014; Zykovich *et al.*, 2014). However, most array-based EWAS continue to be cross-sectional studies. One of a few notable longitudinal studies followed the methylation changes of 67 individuals free of age-related diseases during 8 years (Florath *et al.*, 2014). The authors identified 155 a-DMPs from the observatory and confirmatory cross-sectional datasets, in a total of nearly a thousand subjects. They found that methylation levels at these a-DMPs

persistently changed after 8 years, suggesting that methylation changes on certain CpG sites could be good ageing markers.

Recent studies have also explored the methylation-age effects across different tissues related to ageing, such as whole blood/leukocytes (B. C. Christensen *et al.*, 2009; Rakyan *et al.*, 2010; Teschendorff *et al.*, 2010; Adkins *et al.*, 2011; Alisch *et al.*, 2012; J. T. Bell *et al.*, 2012), brain (Hernandez *et al.*, 2011; Numata *et al.*, 2012), skin (Gronniger *et al.*, 2010; Koch *et al.*, 2011), saliva (Bocklandt *et al.*, 2011), skeletal muscle (Zykovich *et al.*, 2014) and others (Koch & Wagner, 2011). Several a-DMPs overlapped across studies, indicating that the age effect is not only highly replicable, but can also be shared across tissues.

The methylation changes at a-DMPs might reflect an individual's true biological age. Several recent studies have successfully used different sets of a-DMPs to construct models for estimating the DNA methylation age (Hannan *et al.*, 2009; Bocklandt *et al.*, 2011; Hannum *et al.*, 2013; Horvath, 2013; Florath *et al.*, 2014). In two studies, researchers have found that the methylation age was accelerated in the disease-related tissues, exclusively in cancerous ones, suggesting that methylation age might be associated with biological processes (Hannum *et al.*, 2013; Horvath, 2013).

In this chapter, I will first present my results identifying a-DMPs using Illumina 27k, and compare these with the results of Rakyan et al (Rakyan *et al.*, 2010), because one of the datasets in my chapter (Dataset 1, batch 1) were previously studied by Rakyan et al. Secondly, I extend a-DMP identification to the Illumina 450k, and identify shared a-DMPs across tissues. Lastly, to test the hypothesis that methylation age could be associated with biological ageing, I examine the correlation between age-related phenotypes and methylation age acceleration changes in multiple tissues.



## 4.2 Materials and methods

### 4.2.1 Illumina 27k dataset

Twin volunteers were recruited from the TwinsUK cohort and their methylation levels were measured by Illumina 27k using DNA from whole blood samples. The subjects were Caucasian female twins aged from 32 to 82 years old. There were 172 subjects who passed quality control. Table 4-1 shows the 172 subjects who were further divided into 93 (batch 1) and 79 (batch 2). For batch 1, a subset of 64 subjects as unrelated, which comprised of a single individual from each twin pair and all the singletons, and similarly for batch 2 there were 56 selected unrelated individuals. These data have previously been published (Rakyan *et al.*, 2010; J. T. Bell *et al.*, 2012)

**Table 4-1. Summary of Illumina 27k datasets**

<b>Dataset</b>	<b>Total samples (N)</b>	<b>MZ pairs</b>	<b>DZ pairs</b>	<b>Unrelated</b>
Batch 1	93	12	17	64
Batch 2	79	9	14	56
Combined	172	43*	33*	20*

\*Some of co-twins are selected into batch 1 and batch 2 as internal controls

### 4.2.2 Illumina 450k dataset

DNA methylation levels were obtained in three tissue datasets using the Illumina 450k, to investigate cross-tissue a-DMPs and predict the DNA methylation age. Table 4-2 shows the total samples from datasets that were filtered, for example, for the blood dataset, 306/449 and 383/449 subjects were used in the a-DMPs and age acceleration analysis. The filtering was due to missing covariate data, such as white cell count, which was only available for 306 individuals with blood for a-DMPs analysis. For the age acceleration analysis I did not require these covariate data and so all the 449 subjects could be used, however, I selected 383 subjects with no age-related or severe disease.

**Table 4-2. Summary of Illumina 450k datasets**

<b>Dataset</b>	<b>Total sample (N)</b>	<b>a-DMP analysis (N)</b>	<b>Mean age (range)</b>	<b>Age acceleration Analysis (N)</b>	<b>Mean age (range)</b>
Blood	449	306	53 (33, 78)	266	58 (37, 82)
Adipose	648	551	59 (39, 85)	542	59 (39, 85)
Skin	469	469	59 (39, 85)	469	59 (39, 85)

The adipose methylation dataset has previously been published (Grundberg *et al.*, 2013) and the individuals in the skin methylation dataset are a subset of the 648 individuals in the adipose methylation dataset.

The DNA methylation datasets passed the quality control procedure as described in Chapter 3. For the samples used in this chapter, the following covariates were used in all analyses: age, BMI, family, zygosity, methylation chip, order of the sample on the chip, and bisulfite conversion levels. Additional covariates included blood cell counts (whole blood samples), and bisulfite efficiency (adipose tissue).

### **4.2.3 Ageing-related clinical measurements**

A total of 36 quantitative indicators of ageing were included in the analysis, Table 4-3 list the mean and standard deviation of these age-related phenotypes in three datasets. Some of the phenotypes have been normalised and therefore the mean value is close to 0. I tested for the correlation between age-related DNA methylation variables and these age-related phenotypes, as well as age-related disease status, such as type 2 diabetes, obesity, and high blood pressure. The age-related phenotype data were obtained from biochemical measures from blood and anthropometric and physical measurements during clinical twin visits (Moayyeri *et al.*, 2013). The findings that passed nominal statistical significance are discussed in the results of this chapter.

**Table 4-3. List of ageing-related indicators**

<b>Phenotypes</b>	<b>Blood (Mean ± SD)</b>	<b>Adipose (Mean ± SD)</b>	<b>Skin (Mean ± SD)</b>
<b>Haematological values</b>			
Haemoglobin (Hgb) <sup>1</sup>	-0.01 ± 1.05	-0.01 ± 1.01	-0.04 ± 0.99
Mean corpuscular volume (MCV) <sup>1</sup>	-0.01 ± 1.02	-0.04 ± 0.98	-0.01 ± 0.99
Packed cell volume (PCV) <sup>1</sup>	0 ± 1.06	0 ± 1.00	-0.05 ± 0.98
Platelet count (PLT) <sup>1</sup>	-0.05 ± 1.03	0.05 ± 0.92	0.04 ± 0.94
Red blood cell (RBC)	4.31 ± 0.35	4.32 ± 0.33	4.30 ± 0.34
White blood cell (WBC) <sup>1</sup>	-0.07 ± 1.00	0.01 ± 1.01	-0.01 ± 1.00
<b>Heart function test</b>			
Heart rate (HR)	68.19 ± 11.61	67.26 ± 10.71	67.36 ± 10.83
RR interval	906.29 ± 159.45	914.8 ± 146.33	914.1 ± 148.52
QT interval	405.54 ± 30.28	408.41 ± 29.06	407.06 ± 28.69
<b>Liver function test</b>			
Albumin	41.29 ± 2.81	41.08 ± 2.90	41.2 ± 3.02
Total Bilirubin	8.24 ± 4.75	8.8 ± 3.87	8.68 ± 3.82
Apolipoprotein A-1 (ApoA1)	1.67 ± 0.26	1.66 ± 0.253	1.66 ± 0.26
Apolipoprotein B (ApoB) <sup>1</sup>	-0.14 ± 0.26	-0.12 ± 0.26	-0.13 ± 0.26
Gamma glutamyl transferase (GGT)	27.82 ± 26.08	25.26 ± 21.45	24.97 ± 18.81
<b>Type II diabetes-related markers</b>			
HOMA-insulin resistance <sup>1</sup>	0.31 ± 0.76	0.25 ± 0.69	0.25 ± 0.65
HOMA-beta cell <sup>1</sup>	3.00 ± 0.79	2.97 ± 0.73	2.97 ± 0.70
Glucose <sup>1</sup>	5.00 ± 0.50	4.95 ± 0.47	4.93 ± 0.50
Insulin	3.75 ± 0.70	3.70 ± 0.64	3.71 ± 0.61
<b>Morphological measurements</b>			
Height	161.3 ± 6.23	161.56 ± 5.92	161.52 ± 5.94
Weight	69.08 ± 14.10	69.99 ± 14.05	69.45 ± 13.49
Body mass index (BMI)	26.56 ± 4.75	26.77 ± 4.91	26.58 ± 4.74
Waist	79.7 ± 9.95	79.94 ± 10.53	80.07 ± 10.37
Hip	101.16 ± 9.03	101.39 ± 9.83	101.4 ± 9.35
Waist-Hip-ratio (WHR)	0.79 ± 0.057	0.79 ± 0.05	0.79 ± 0.06
<b>Blood lipid profile</b>			
Total cholesterol <sup>1</sup>	5.50 ± 1.02	5.60 ± 1.06	5.61 ± 1.07
Triglycerides <sup>1</sup>	0.06 ± 0.98	-0.02 ± 1.01	-0.05 ± 1.04
Low density lipoproteins (LDL) <sup>1</sup>	4.20 ± 1.21	4.27 ± 1.20	4.31 ± 1.22
High density lipoproteins (HDL) <sup>1</sup>	-0.04 ± 1.04	-0.04 ± 0.98	-0.01 ± 0.98
Leptin	2.59 ± 0.59	2.54 ± 0.66	2.59 ± 0.68
Adiponectin	1.95 ± 0.51	1.94 ± 0.49	1.95 ± 0.51
<b>Other biochemistry</b>			
C-reactive protein (CRP)	2.95 ± 4.71	3.13 ± 5.41	3.31 ± 7.16
Uric acid <sup>1</sup>	0.05 ± 1.03	0.03 ± 1.00	0.07 ± 0.98
Bicarbonate	24.81 ± 2.47	24.87 ± 2.59	25.03 ± 2.69
Creatinine <sup>1</sup>	4.28 ± 0.156	4.28 ± 0.16	4.29 ± 0.16
Urea <sup>1</sup>	1.14 ± 0.150	1.12 ± 0.15	1.12 ± 0.14
<b>Other physical measurements</b>			
Blood pressure (systolic)	126.34 ± 14.68	128.2 ± 15.67	128.5 ± 15.42
Blood pressure (Diastolic)	77.42 ± 9.26	77.54 ± 9.58	77.81 ± 9.42
Lung function (FVC)	3.18 ± 0.58	3.22 ± 0.61	3.23 ± 0.60
Lung function (FEV)	2.50 ± 0.51	2.54 ± 0.56	2.56 ± 0.54

<sup>1</sup>Normalized phenotypes

## 4.2.4 Statistical analyses

### 4.2.4.1 Age-differential methylation sites analyses

Based on the previous findings and my results, the age-related differential methylation could occur at single position and in a region. I will then use a-DMPs for the whole chapter. Two methods were used to help identify the a-DMPs: permutation-based and linear mixed effect regression (LMER) model.

#### 4.2.4.1.1 Permutation-based analysis (used on Illumina 27k)

The raw beta values were quantile-normalized within each subject, and then normalized to normal distribution on each probe. The normalized data were fitted to a linear mixed effect model to correct for batch effect and covariates. The residuals of each probe were correlated with age using the Spearman's correlation coefficient rho ( $\rho$ ). The true  $\rho$  (observed  $\rho$ ) between methylation levels and age was calculated. Age was shuffled to calculate the  $\rho$  for each permutation (permuted  $\rho$ ). Permutations were performed 1000 times. Permutations with more extreme values than the true coefficient  $\rho$  observed were counted then divided by 1000.

$$P \text{ value of each probe} = \frac{\text{Number of } (|\text{permuted } \rho| > |\text{observed } \rho|)}{1000}$$

Probes with P value  $\leq 0.01$  were considered as a-DMPs. The analysis was performed on the Illumina 27k dataset only.

#### 4.2.4.1.2 Linear mixed effect regression (LMER) model (used on Illumina 27k and 450k)

A linear mixed effect regression model adjusted for the family structure and twin structure as the data contained MZ and DZ twins. The covariates, such as fixed-effect terms (age, plate, position on the plate) and random-effect terms (family structure and zygosity, i.e. MZ, DZ and singleton) were included. For each probe, a full model that regressed the raw beta values on all of the covariates was compared to a null model that excluded age. The models were compared using the ANOVA F statistic in R. An a-DMP was accepted if the P value passed the Bonferroni correction or false discovery

rate. In the dataset of unrelated subjects, I replaced the LMER model with a simple linear regression model where the random-effects were not considered.

#### ***4.2.4.2 Analysis of Illumina 27k data***

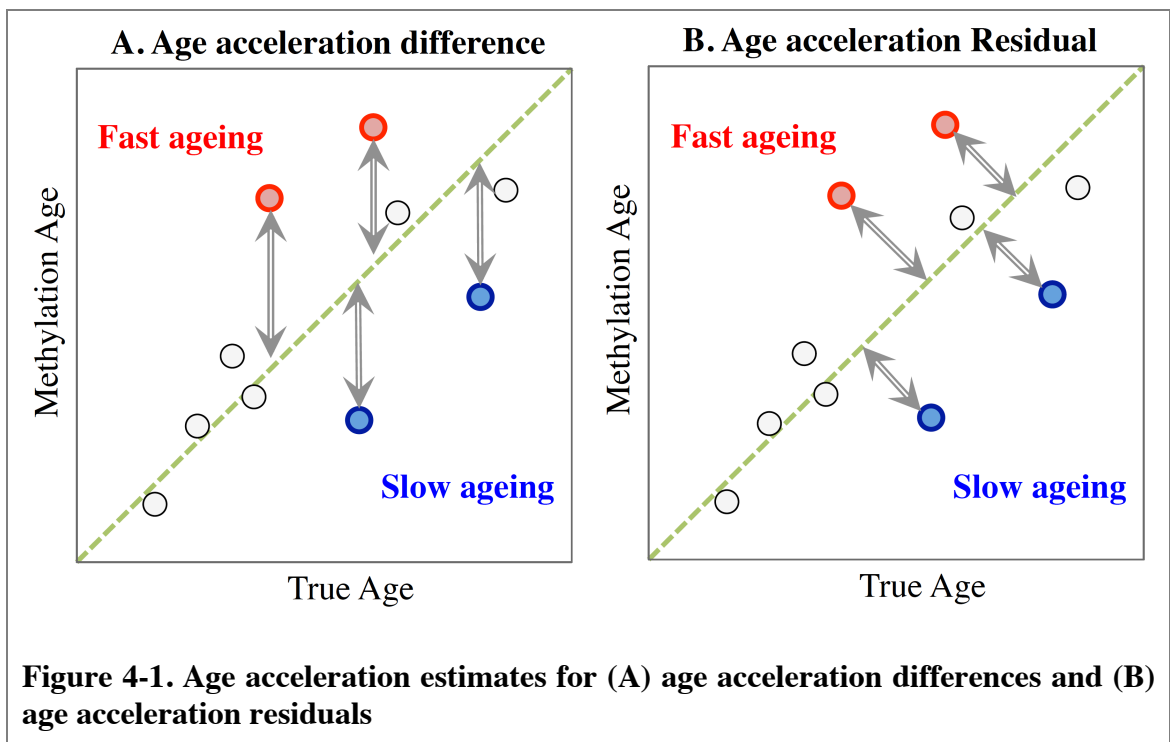
Three blood datasets were used to identify a-DMPs. Specifically, (1) 93 subjects from batch 1 (the same subjects from Rakyan et al. (Rakyan *et al.*, 2010)), (2) 64 unrelateds from batch 1, and (3) 172 subjects, combined from batch 1 and batch 2. The methylation levels at each probe were compared to age across all individuals using permutations and LMER. I obtained a more precise measure of age at DNA extraction of the sample, rather than chronological age at present, as the timing of the sample collection varied.

#### ***4.2.4.3 Analysis of Illumina 450k data***

The first Illumina 450k analysis focused on the identification of tissue-shared a-DMPs. The significant a-DMPs were determined using LMER (including two significance criteria: Bonferroni adjusted P value and false discovery rate). Because the comparison of a-DMPs by the Bonferroni adjusted P value could be too conservative and influenced by the sample size, tissue-sharing was then assessed using the ‘proportion of true positives from the P value distribution’ (from the ANOVA P values), a method first introduced by Storey and Tibshirani (Storey & Tibshirani, 2003) and recently applied to assess tissue-sharing in gene expression data (Nica *et al.*, 2011). There were two values,  $\pi_0$  and  $\pi_1$  ( $\pi_1 = 1 - \pi_0$ ), which represented the proportion of false positive and true positive associations in the P value distribution of the dataset. These two values could be calculated in the R package ‘qvalue’ (Dabney *et al.*). The idea was to check whether the proportion of significant hits from one dataset was also significant in the other dataset. The following is an example of how this analysis was performed: I took the significant a-DMPs found in the blood dataset and obtained the P values of these exact probes in the adipose dataset. Using the qvalue package I calculated  $\pi_1$  based on the P values in the adipose subset. If the  $\pi_1$  was 0.5, this indicated there was 50% tissue sharing of a-DMPs in blood with a-DMPs in adipose tissue.

#### 4.2.4.4 Age acceleration analysis in Illumina 450k data

The second Illumina 450k analysis focused on age acceleration. Firstly, the predicted DNA methylation age was calculated using R code kindly provided by Dr. Steve Horvath (Horvath, 2013). For each of the Illumina 450k datasets, I extracted 21,369 of CpGs then BMIQ transformed them following his pipeline. The 353 ‘clock’ CpGs (Horvath, 2013) were extracted and used to the predict methylation age using a penalized regression model, which was built from publicly available datasets. After the methylation age was estimated, two age accelerations were calculated as (1) ‘Age acceleration difference’, which is defined as the difference between the DNA methylation age and chronological age; and (2) ‘Age acceleration residual’, which is defined as residual from regressing DNA methylation age on chronological age (Figure 4-1). The recommended age acceleration estimate is the age acceleration residual, because it adjusts for the effect of the age contribution.



Positive age acceleration indicates greater methylation age than true age, suggesting a faster ageing process in the individual. The age acceleration was compared against a number of clinical phenotypes and quantitative traits (Table 4-3). Ideally, subjects with faster ageing were expected to have less good healthy ageing indicators. The

comparisons were performed using Pearson's correlation and the results were reported at a significance level of  $P < 0.05$ .

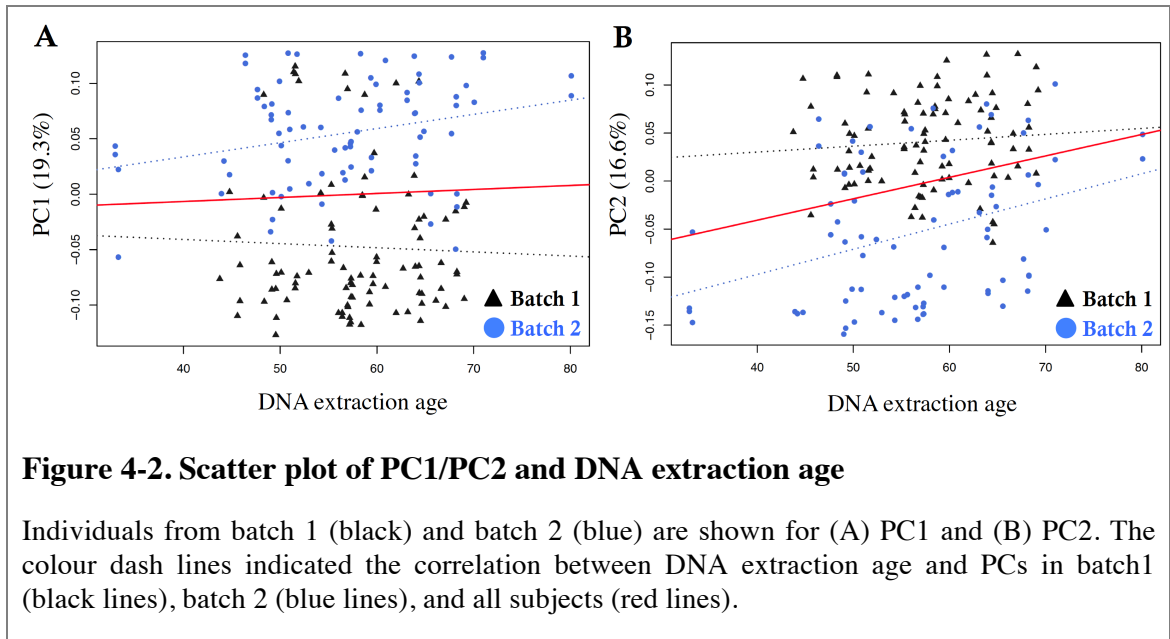
## **4.3 Results**

In following paragraphs, I will separate the results into two big sections. The first section I present age the differential methylation results using Illumina 27k platform on whole blood samples. The second section, are results from all age analysis done for the three Illumina 450k datasets, including age differential methylation results and age acceleration results.

### **4.3 Section A: Illumina 27k data (blood)**

#### **4.3.1 Overall methylation patterns with age**

The overall genome-wide methylation patterns on the Illumina 27k array were compared to age, first using principal component analysis. The first PC (PC1) in the autosomal DNA methylation data captured 19.3% of the overall variance, but did not correlate significantly with age at DNA extraction of the sample ( $r = 0.039$  and  $P = 0.613$ , Figure 4-2A). PC2 to PC4 can explain 16.6%, 7.7% and 5.6% of the overall methylation variance, respectively, and only PC2 was significantly correlated with age ( $r = 0.239$  and  $P = 0.002$ , Figure 4-2B). The combined PC1 to PC4 can explain 49.2% of the overall variance. This indicated that age might induce variability in a subset of genome-wide DNA methylation levels.



### 4.3.2 Age-related differential methylation site: Permutation-based

The objective here was to repeat the analysis of Rakyan et al. (Rakyan *et al.*, 2010) and confirm their results using the same samples (93 subjects), approach (permutation of P values), and significance criteria. However, Rakyan et al used chronological age in 2009, but there was some variability in the date of blood sample collection across these individuals. Therefore, I obtained a more precise estimate of age using the age at which DNA was extracted for the DNA methylation assay.

#### 4.3.2.1 Using chronological age to estimate a-DMPs

Table 4-4 shows the 213 hyper-a-DMPs found in the previous study (131 hyper-a-DMPs were replicated across multiple cell types, such as white blood cells, CD4+, and CD8+ cells) using permutations of chronological age (Rakyan *et al.*, 2010). Similarly, I used chronological age and found more a-DMPs: 419 hyper-a-DMPs and 374 hypo-a-DMPs, and also confirmed the 131 hyper-a-DMPs. All of the 213 previously reported hyper-a-DMPs were in my list of 419 hyper-a-DMPs. The increase in a-DMP likely occurred because I removed a number of probes in the quality control (e.g. probes incorrectly mapped to the genome), which relaxed the multiple testing criteria and the resulting significance thresholds.



**Table 4-4. Age-related differential methylation identified with chronological and DNA extraction age**

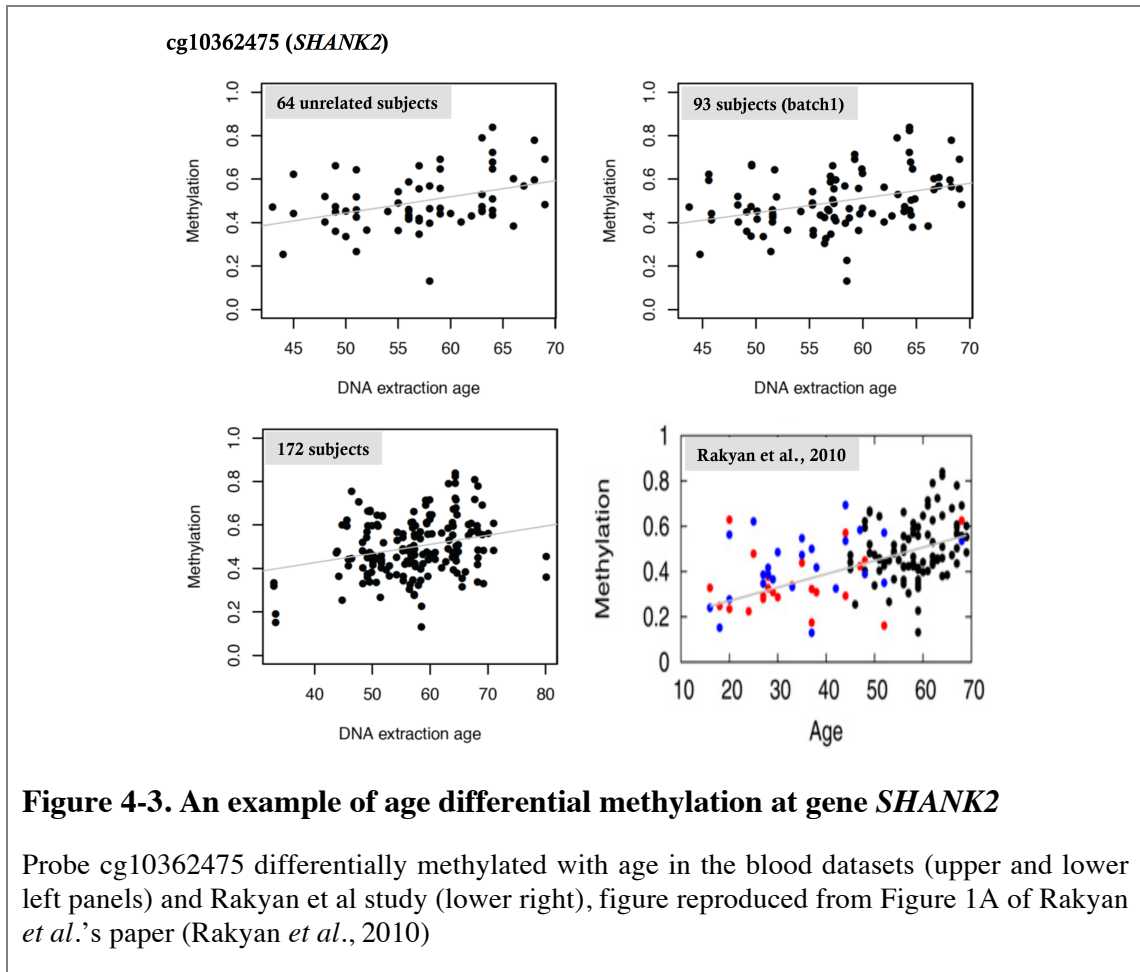
Study/dataset	Hyper-a-DMPs*	Hypo-a-DMPs	Total
<b>Chronological age</b>			
93 subjects [Rakyan et al, 2010]	213 (131)	147	360
93 subjects [blood 27k data]	419 (131)	374	793
<b>DNA extraction age</b>			
93 subjects [blood 27k data]	570 (127)	530	1100
64 subjects [blood 27k data]	397 (87)	401	798
172 subjects [blood 27k data]	1274 (114)	1236	2510

\*Parenthesis is number of validated a-DMPs (out of 131) from Rakyan et al study. A significance level of 0.01 was set for all permutation results.

#### **4.3.2.2 Using DNA extraction age to estimate a-DMPs**

Using the extraction age, there were 570 hyper-a-DMPs and 530 hypo-a-DMPs (Table 4-4). Of the previously identified 131 hyper-a-DMPs, 127 were identified at  $P < 0.01$  and 4 were borderline significant ( $P$  value of 0.019 to 0.029). The same analysis was performed for 64 unrelated subjects and for the entire dataset of 172 subjects. A large proportion of the 131 previously reported a-DMPs were confirmed to have strong hyper-a-DMP effects across all data subsets. The total a-DMPs found in the 172 subjects were 2,510.

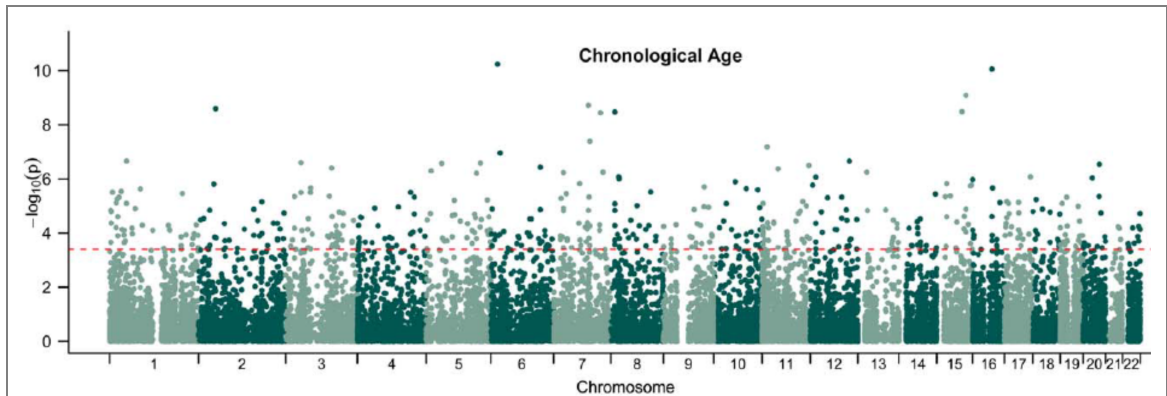
Figure 4-3 shows the consistent correlation between methylation levels and age for one a-DMP: cg10362475 (chromosome 11, in gene body of *SHANK2*). For the 93 subjects, the Spearman's coefficient between methylation of the chronological age ( $\rho = 0.285$ ;  $P = 0.004$ ) and DNA extraction age ( $\rho = 0.298$ ;  $P = 0.001$ ) were similar. This probe was hyper-methylated with age and showed a significant age effect across all three datasets and in the previous study (Rakyan *et al.*, 2010).



The permutation-based method was used solely to confirm the results from Rakyan *et al* using a more precise measure of chronological age. The LMER method to find a-DMPs (described below) should be more accurate in these data owing to the incorporation of family and zygosity structure in the model.

### 4.3.3 Age-related differential methylation: LMER model

A linear mixed effect regression (LMER) model and the permutation-based method was applied to the 172 subjects and identified 490 a-DMPs that passed 5% FDR threshold (J. T. Bell *et al.*, 2012). All a-DMPs were positively correlated with extraction age (Figure 4-4,  $P = 3.96 \times 10^{-4}$ ), and 75 hyper-a-DMPs overlapped between this study and the 213 hyper-a-DMPs from the previous study (Rakyan *et al.*, 2010). Most a-DMPs demonstrated the same effect direction in both studies. Furthermore, these a-DMPs were concordant with other studies, such as 36/88 a-DMPs in saliva (Bocklandt *et al.*, 2011) and 3/10 a-DMPs in brain tissues (Hernandez *et al.*, 2011).

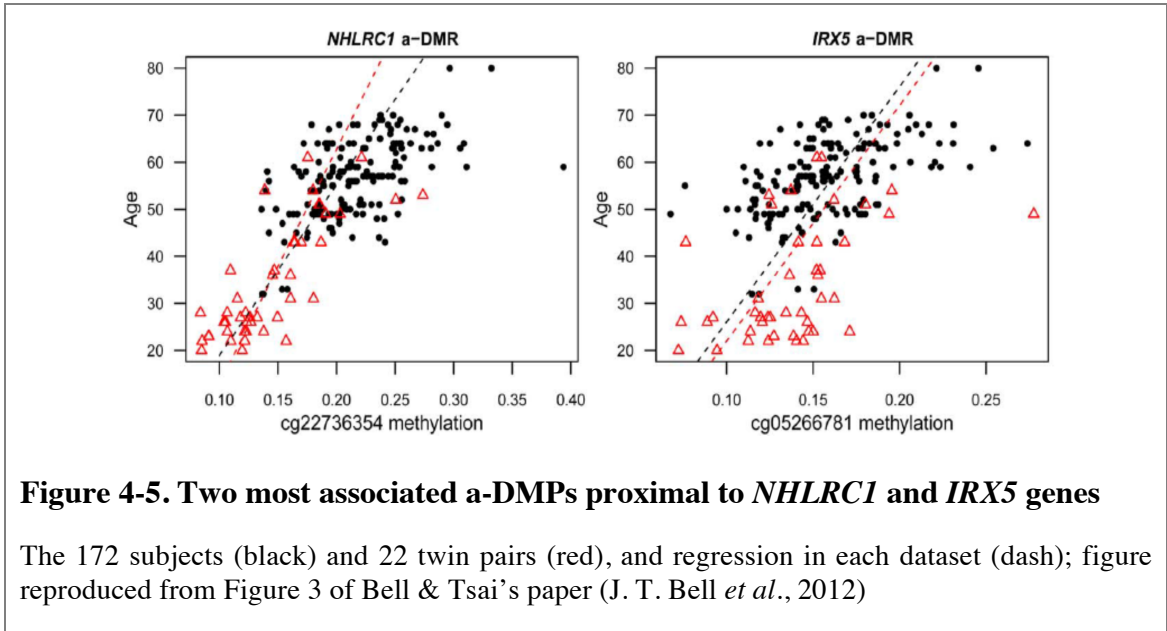


**Figure 4-4. Manhattan plot of EWAS using chronological age at 5% FDR**

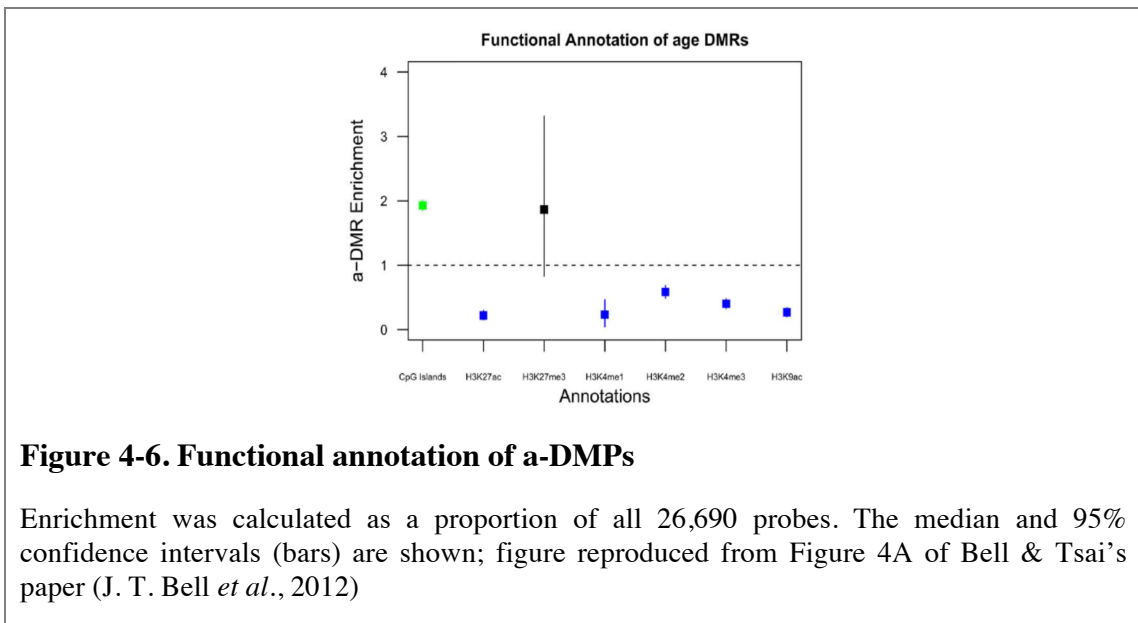
A total of 490 significant a-DMPs were found pass 5% FDR (dots above red dash line) from 172 subjects. Figure reproduced from Figure 2 of Bell & Tsai's paper (J. T. Bell *et al.*, 2012)

The 490 a-DMPs were followed up in a replication cohort of 22 MZ twin pairs profiled on the Illumina 27k in whole blood from a previous study (E. L. Dempster *et al.*, 2011). The 22 twin pairs were younger (age range 20-61, median age 28). I tested the correlation between methylation levels and age in the 22 twin pairs, and in a subset of 22 unaffected unrelated individuals (using the healthy co-twin of the 22 pairs). The beta values of the 22 twin pairs were normalized to normal distribution) then a linear mixed effect model was applied along with the random effect (family) and fixed effect (plate, gender, and age) terms. For the 22 unrelateds, the Spearman's correlation coefficients between raw methylation levels and age were calculated. In summary, for the twin pairs and unrelated datasets, there were 184/490 (38%) and 69/490 (14%) a-DMPs, respectively. These a-DMPs had the same effect direction at a nominally significant level ( $P = 0.05$ ), and 404/490 (82%) and 369/490 (77%) a-DMPs had the same effect direction but no significance.

Figure 4-5 shows the two most significant a-DMPs (cgg22736354 in *NHLRC1* and cg05266781 in *IRX5*) in the discovery (black) and the replication (red) datasets. The correlation patterns were similar in both datasets.



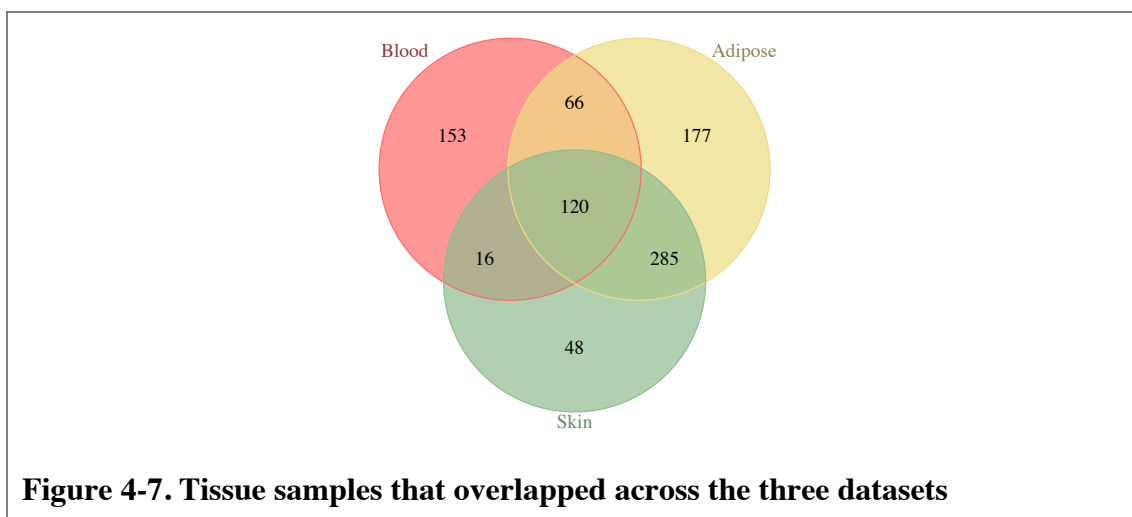
To gain insights into functional characteristics of these a-DMPs, the methylation probes were categorized according to whether they fell into CpG islands or histone modification marks (H3K9ac, H3K27ac, H3K27me3, H3K4me1, H3K4me, and H3K4me3, (Rosenbloom *et al.*, 2013)) of the human lymphoblastoid GM12878 cell line. Gene ontology enrichment was performed using the Gorilla tool (Eden *et al.*, 2009) based on a ranked list of a-DMPs genes. The a-DMPs revealed a two-fold enrichment for CpG islands as compared to the total 26,690 probes (Figure 4-6). These a-DMPs were hyper-methylated (98%) and some were involved in the regulation of developmental processes and regulation of transcription.



### 4.3 Section B: Illumina 450k datasets (blood, skin, adipose tissue)

#### 4.3.4 a-DMPs analysis in three tissues

The DNA methylation profiles of three tissue samples (blood, adipose, skin) were profiled using the Illumina 450k array (Table 4-3). The skin and adipose tissues were collected from the same biopsy, therefore shared the same DNA extraction age. The blood samples were collected at different times with different extraction age. Figure 4-7 shows 120 subjects who contributed all three samples, and samples contributed two samples, such as in adipose and skin (N = 405), blood and adipose (N = 186), and blood and skin (N = 136).



The three tissue datasets consisted of twin pairs. The proportion of MZ and DZ twin pairs varied across datasets, and in blood dataset, there were markedly more MZ than the DZ twins (Table 4-5). This was due to the selection criteria for those pairs with no severe diseases and availability of covariate information.

**Table 4-5. Summary of twin pairs included in the three datasets**

Datasets (Tissue)	All	Zygoty (MZ, DZ)	a-DMPs analysis	Zygoty (MZ, DZ)	Age acceleration analysis	Zygoty (MZ, DZ)
Blood	449	332, 70	306	160, 42	355	202, 54
Adipose	648	194, 324	551	102, 250	648	194, 324
Skin	469	102, 160	469	102, 160	469	102, 160

To find the tissue-shared a-DMPs from the three datasets, the Bonferroni-adjusted P value was used at  $1.08 \times 10^{-7}$ ,  $1.35 \times 10^{-7}$ ,  $1.43 \times 10^{-7}$  in blood, adipose, skin tissues, respectively. The highest number of a-DMPs ( $N = 7,183$ ) was found in adipose tissue as well as the most significant a-DMP (cg16867657, *EVOLV2*). The skin and blood tissues showed a smaller number 3,142 and 1,256 of significant a-DMPs, respectively (Table 4-6).

Using a single P value to call a-DMPs across tissues might be misleading because the results would depend on the actual number of probes tested and on the sample size of the study, as a larger sample would have more power to identify DMPs. Therefore, I also used the false discovery rate to identify the a-DMPs (See Table 4-6). At an FDR = 1% and 5% as the selection criteria for a-DMPs, there were many a-DMPs observed across the three sample Table 4-6), and blood and adipose tissue consistently had the greatest number of a-DMPs.

**Table 4-6. List of significant a-DMPs found in three datasets**

	<b>Blood</b>	<b>Adipose</b>	<b>Skin</b>
Total CpG sites (N)	461,039	370,960	350,463
Sites pass Bonferroni correction (N)	1,256	7,183	3,142
Sites pass FDR 5% (N)	58,439	51,965	30,478
Sites pass FDR 1% (N)	26,416	30,623	16,271
P value of most significant hit	$3.95 \times 10^{-34}$	$8.65 \times 10^{-92}$	$3.20 \times 10^{-38}$

\*Number of CpG site that passed Bonferroni correction of P value at  $1.08 \times 10^{-7}$ ,  $1.35 \times 10^{-7}$ ,  $1.43 \times 10^{-7}$  in blood, adipose, skin tissues, respectively.

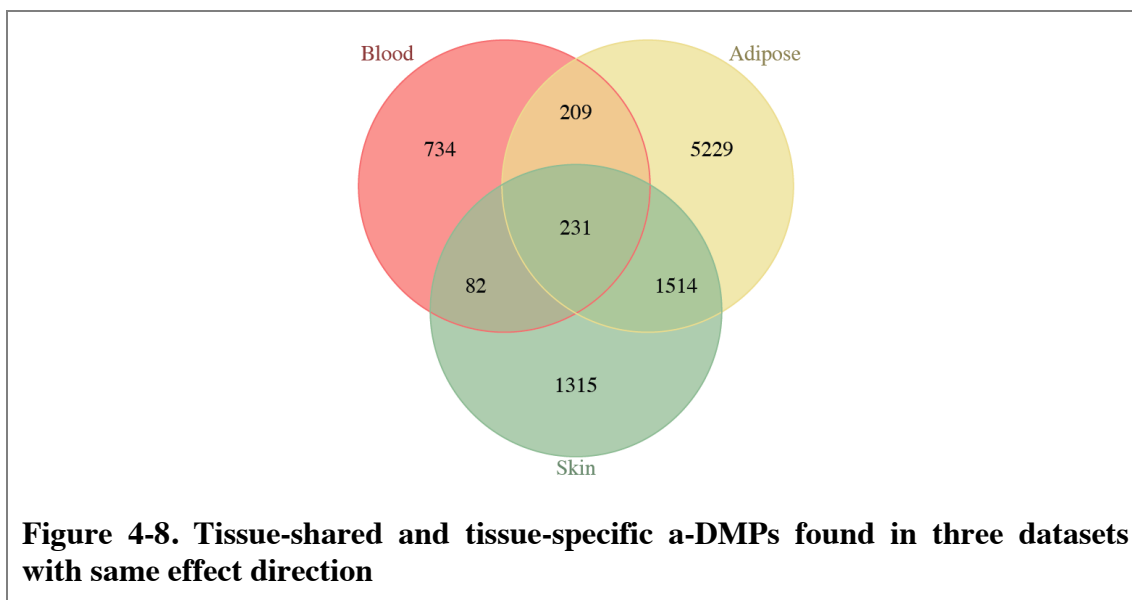
### 4.3.5 a-DMPs analysis across multiple tissues

#### 4.3.5.1 Using a single Bonferroni adjusted P value as the significance threshold

There were 480,504 probes across the three datasets, and 108,755 a-DMPs (22.6%; 68,515 hyper-methylated and 40,240 hypo-methylated) had the same effect directions across all three tissues, including the 2,665 CpG sites from the X chromosome.

Using a single multiple testing threshold of  $P = 1 \times 10^{-7}$  (Bonferroni correction adjusting for 500,000 tests), there were overlapping a-DMPs (Figure 4-8) between adipose and skin (440 a-DMPs; 431 hyper-methylated and 9 hypo-methylated), between blood and

skin (313 a-DMPs; 287 hyper-methylated and 26 hypo-methylated), and between adipose and skin (1,745 a-DMPs; 1,726 hyper-methylated and 19 hypo-methylated). There were 231 a-DMPs that overlapped across all three tissues.



#### 4.3.5.2 Using FDR as significance criteria

Using the a-DMP results across the three samples at FDR = 1% (Table 4-6), I then applied the proportion of true positive test to estimate the tissue-shared effect. At FDR 1%, 7,263 a-DMPs overlapped between blood and adipose, 5,398 overlapped between adipose and skin, and 5,398 between skin and blood.

Table 4-7 shows the result of the true positive analysis. Generally, the sharing between any two tissues was high (> 60% in all comparisons) suggesting that the age effect on methylation was more tissue-shared than tissue-specific. The highest sharing was between adipose and skin, at about 72-77%.

**Table 4-7. Pairwise tissue-shared a-DMPs**

Reference (FDR 1%) <sup>1</sup>	Blood <sup>2</sup>	Adipose <sup>2</sup>	Skin <sup>2</sup>
Blood (N = 264,16)	-	30,020; 62.82%	15,935; 76.45%
Adipose (N = 30,623)	20,912; 64.34%	-	13,889; 77.08%
Skin (N = 16,271)	20,028; 66.46%	25,027; 72.16%	-

<sup>1</sup>Reference tissue; parenthesis were a-DMPs that passed FDR 1%

<sup>2</sup>Number of target a-DMPs overlapped with reference tissue (left) and the proportion of tissue-shared (right).

In total, there were 3,441 a-DMPs identified across the three tissues that mapped to 1,892 unique genes. To understand the biological significance of these a-DMPs, a gene ontology analysis was performed using the WEB-based GENE SeT AnaLysis Toolkit. Many of the 1,892 a-DMP genes were involved in the developmental process (707 genes, adjusted  $P = 1.67 \times 10^{-47}$ ), nervous system development (394 genes, adjusted  $P = 1.26 \times 10^{-62}$ ), DNA binding in the regulatory regions (79 genes, adjusted  $P = 1.07 \times 10^{-12}$ ) and others. Moreover, there were genes that associated with diseases, predominately mental disorders (113 genes, adjusted  $P = 1.28 \times 10^{-39}$ ), such as schizophrenia (82 genes, adjusted  $P = 1.21 \times 10^{-32}$ ), bipolar disorder (73 genes, adjusted  $P = 3.56 \times 10^{-27}$ ), and anxiety disorders (48 genes, adjusted  $P = 1.66 \times 10^{-25}$ ).

## **4.3.6 Age acceleration analysis**

### ***4.3.6.1 DNA methylation age and age acceleration across three tissues***

For each subject, the DNA methylation age was estimated based on the 353 ‘clock’ CpGs (Horvath, 2013). Age acceleration was calculated by subtracting chronological age from DNA methylation age. The distributions of age acceleration differences and age acceleration residuals in the three tissues are prone to be normally distributed (Figure 4-9A, Figure 4-9B). Age acceleration differences in blood showed slightly more variability than the other two tissues, but the methylation age seems to be a better age predictor than the other two tissues (Figure 4-9C). The DNA methylation age was highly correlated with the chronological age of each tissue sample: Pearson’s correlation coefficient were  $r = 0.82$  in blood,  $r = 0.79$  in adipose, and  $r = 0.79$  in skin (Figure 4-9C). The methylation age of the extreme age groups (youngest and oldest) deviated more from the prediction lines in skin and adipose tissue. There was a negative correlation between the chronological age and age acceleration differences (Figure 4-9D), therefore use age acceleration residuals could yield a more unbiased results (Figure 4-9E).



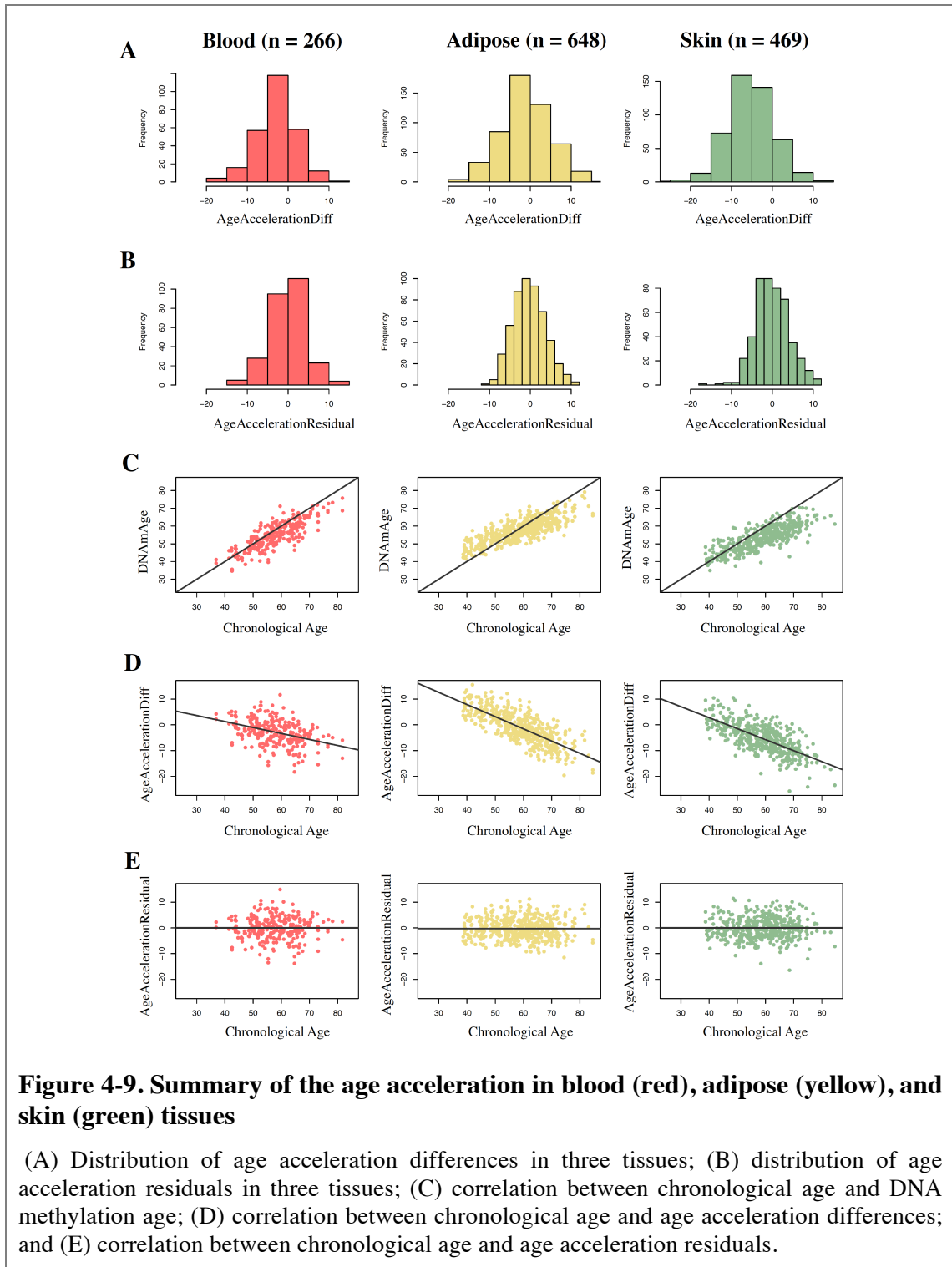
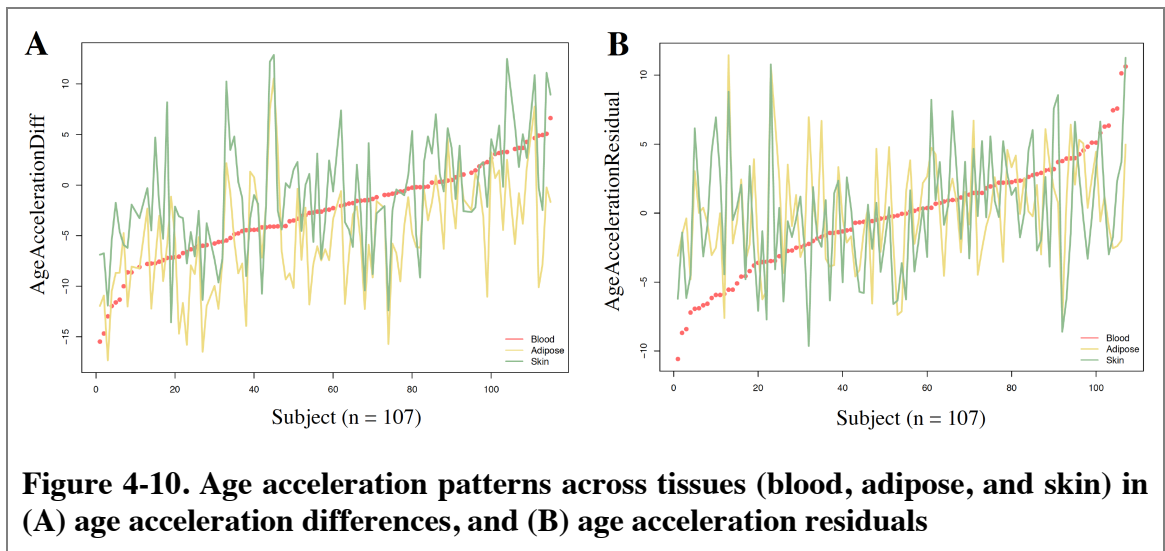


Figure 4-10 shows the age acceleration changes of 107 subjects who provided all three tissues samples. Here, blood age acceleration was made the reference (red), while the relative age acceleration of adipose (yellow) and skin (green) were compared for each subjects. Age acceleration was more similar between adipose and skin for both the age acceleration estimates, and quite different when compared to blood samples.



I correlated the age acceleration residuals with the age-related phenotypes in all three tissues. Table 4-8 shows the phenotypes associated with age acceleration at nominal significance of  $P < 0.05$ . In addition to the age-related phenotypes, I have tested the association between age acceleration and environmental effects, such as alcohol consumption and smoking status. I have also compared the age acceleration residuals with different disease statuses, for example, subjects with type 2 diabetes and those free from the disease. Most of correlations were observed in blood and the least was in skin. In blood samples, the age acceleration residuals seemed to be a good biomarker for age-related phenotypes, and the correlations were as expected. For example, the subjects who were ageing faster (a higher value of age acceleration residual) had higher LDL, triglyceride, uric acid, and higher blood pressure, also showed lower HDL and lung function. Interestingly, higher alcohol consumption is significantly associated with faster ageing, and in current smokers there was slightly higher age acceleration compared to non-smokers and ex-smokers. However, I did not find any significant differences between the subjects with and without diseases (i.e. type 2 diabetes, and cardiovascular disease). This might be due to having fewer subjects in this category than the other two datasets, and a further analysis should be performed in a bigger sample size for cross-validation of these results.

**Table 4-8. List of significant age accelerated phenotypes in three tissues**

<b>Phenotypes</b>	<b>Blood r (P value)</b>	<b>Adipose r (P value)</b>	<b>Skin r (P value)</b>
<b>Haematological values</b>			
White blood cell (WBC) <sup>1</sup>	0.22 ( $4 \times 10^{-4}$ )	-	-
QT interval	-	-0.10 (0.032)	-
<b>Liver function test</b>			
Apolipoprotein B (ApoB) <sup>1</sup>	0.14 (0.019)	-	-
HOMA-insulin resistance <sup>1</sup>	0.16 (0.011)	-	-
HOMA-beta cell <sup>1</sup>	0.12 (0.049)	-	-
Insulin	0.14 (0.007)	-	-
<b>Morphological measurements</b>			
Height	-	-	0.11 (0.016)
Weight	0.12 (0.049)	-	-
Body mass index (BMI)	0.17 (0.007)	-	-
Waist	0.25 (0.001)	-	-
Hip	0.18 (0.021)	-	-
Waist-Hip-ratio (WHR)	0.21 (0.007)	-	-
<b>Blood lipid profile</b>			
Triglycerides <sup>1</sup>	0.20 ( $8 \times 10^{-4}$ )	-	-
Low density lipoproteins (LDL) <sup>1</sup>	0.16 (0.011)	-	-
High density lipoproteins (HDL) <sup>1</sup>	-0.14 (0.025)	-	-
Adiponectin	-	-0.15 (0.017)	-
<b>Other biochemistry</b>			
C-reactive protein (CRP)	0.19 (0.003)	-	-
Uric acid <sup>1</sup>	0.29 (0.001)	-	-
<b>Other physical measurements</b>			
Blood pressure (systolic)	0.25 ( $5 \times 10^{-5}$ )	-	-
Lung function (FVC)	-0.21 ( $6 \times 10^{-4}$ )	-	-
Lung function (FEV)	-0.18 ( $3 \times 10^{-3}$ )	-	-
<b>Environmental effect</b>			
Alcohol	0.16 (0.010)	-	-

<sup>1</sup>Normalized phenotypes

\* Pearson's correlation coefficients and P value

## 4.4 Discussion

The primary goal of this chapter was to identify the DNA methylation changes with age. A number of age-related methylation changes were identified on the Illumina 27k array dataset. Most were hyper-methylated with age, suggesting that methylation levels at these sites increased with age. Since a-DMPs occupied < 10% of the scanned regions from both arrays, the vast majority of CpG sites (90%) were not strongly influenced by age. Age might impact a further set of CpG sites in a more complex manner, because the PC2 was significantly associated with age. This relationship was not observed for the other PCs perhaps because the vast majority of the CpGs showed only small variability across subjects. To study age-effects on DNA methylation in more detail a future strategy could be to focus on the probes that capture more variable DNA methylation patterns.

One of the initial analyses that I performed in my project was to repeat the analyses of a previous study (Rakyan *et al.*, 2010), using the same dataset of 93 individuals. I refined the measure of chronological age by obtaining DNA extraction age for each sample. I was able to validate the a-DMPs from the previous study, and I identified additional a-DMPs in the set of 93 subjects and in the larger dataset of 172 subjects.

The LMER model, which adjusted for family and zygosity using DNA extraction age, identified 490 a-DMPs among 172 twins. The findings had a high agreement (17-58%) with another five a-DMPs studies also on the Illumina 27k array (Table 4-9), but in different sample types. There was a high proportion of hyper-methylated a-DMPs across studies, and most were located on CGIs (see Table 2 of Tsai *et al.*'s paper (Tsai *et al.*, 2012)). Together, the six studies revealed 1,093 unique a-DMPs, however, but none overlapped across all six studies. There were 2 a-DMPs (near *NTPX2* and *PDE4C* genes) found in five studies and 12 a-DMPs (near genes: *GLRA1*, *TMEM179*, *GCM2*, *TRIM58*, *PTGER3*, *ATP8A2*, *MYOD1*, *BRUNOL6*, *GRIA2*, *KCNK12*, *B3GALT6*) across four studies and 11/12 of these were located in CGIs, suggesting that tissue-shared a-DMPs were possibly enriched on CGIs. About 16% of the total unique genes overlapped in at least two studies. The low rate might be in part due to analysis methods, sample size, significance criteria, and different tissues.

**Table 4-9. Pairwise comparison of a-DMPs across six studies (Tsai *et al.*, 2012)**

Studies <sup>1</sup>	589 a-DMPs	490 a-DMPs <sup>3</sup>	131 a-DMPs <sup>4</sup>	88 a-DMPs	19 a-DMPs <sup>2,4</sup>	10 a-DMPs <sup>4</sup>
589 a-DMPs (Teschendorff <i>et al.</i> , 2010)	CGIs: 379 Non-CGIs: 210	81 (78, 3)	30 (30, 0)	42 (30,12)	7 (7, 0)	4 (4, 0)
490 a-DMPs (J. T. Bell <i>et al.</i> , 2012)	16.5% (92.3%, 7.7%)	CGIs: 484 Non-CGIs: 6	75 (75, 0)	36 (34, 2)	11 (11, 0)	3 (3, 0)
131 a-DMPs (Rakyan <i>et al.</i> , 2010)	22.9% (100%, 0%)	57.3% (100%, 0%)	CGIs: 126 Non-CGIs: 5	10 (10, 0)	4 (4, 0)	3 (3, 0)
88 a-DMPs (Bocklandt <i>et al.</i> , 2011)	47.7% (71.4%, 29.6%)	40.9% (94.4%, 5.6%)	11.4% (100%, 0%)	CGIs: 73 Non-CGIs: 15	9 (9, 0)	1 (1, 0)
19 a-DMPs (Koch <i>et al.</i> , 2011)	36.8% (100%, 0%)	57.9% (100%, 0%)	21.1% (100%, 0%)	47.4% (100%, 0%)	CGIs: 19 Non-CGIs: 0	0
10 a-DMPs (Hernandez <i>et al.</i> , 2011)	40% (100%, 0%)	30.0% (100%, 0%)	30.0% (100%, 0%)	10.0% (100%, 0%)	0%	CGIs: 9 Non-CGIs: 1

<sup>1</sup>Studies are compared pairwise, each box indicates the percentage of overlapping a-DMPs, in parenthesis are effect directions: hyper-methylated (left) and hypo-methylated (right), and the study number of CGI and non-CGI are indicated (grey diagonal boxes).

<sup>2</sup>This study contains subjects from Rakyan et al and Teschendorff et al

<sup>3</sup>This study contains subjects from Rakyan et al

<sup>4</sup>These studies provide hyper-a-DMPs only

I compared the platform difference between 27k and 450k at the Illumina 27k 490 a-DMPs detected in blood. The 490 a-DMPs were 97% concordant on the two platforms with the direction of effect, and 37% showed genome-wide significance, on both platforms.

Analysis of a-DMPs on the Illumina 450k array identified many more hits than the Illumina 27k array alone. In whole blood alone, there were now 1,256 significant a-DMPs using Bonferroni correction (compared to 490 on the Illumina 27k), and 26,416 a-DMPs at FDR 1% threshold. Since the Illumina 27k array focuses on the promoter regions, the new a-DMPs that were specific to the Illumina 450k array were predominantly located in gene body (31.3%). Apart from blood, I also extended the study of a-DMPs to two tissues (skin and adipose) to investigate the characteristics of tissue-shared a-DMPs. Using FDR 1% criteria across all tissues, 80% of 3,441 unique a-DMPs were located on CGIs, and more than 60% of the a-DMPs identified in each tissue were also significantly methylated with age in the other two tissues. Combining

the findings with the Illumina 27k array that has low coverage outside of promoter regions, this suggests that many of the tissue-shared a-DMPs are outside of promoters.

I also found evidence that some of the Illumina 450k a-DMPs identified in this chapter have been reported in other studies. For example, the most significant hit in this study was located in *ELOVL2* (cg16867657), and has been reported by other studies. In at least four other studies, the methylation levels on CpG sites of *ELOVL2* were highly associated with age, and this result was replicated in larger samples (Garagnani *et al.*, 2012; Hannum *et al.*, 2013; Florath *et al.*, 2014; Tserel *et al.*, 2014).

In a review of recent studies using the Illumina 450k array, more a-DMPs were identified at different age ranges (Table 4-10). All these studies provide the necessary evidence to show that these CpG site could be a predictor of age.

**Table 4-10. Recent EWAS using chronological age on the Illumina 450k array**

Age range (years)	Tissue/sample	a-DMPs analysis (significant threshold)	Major findings	Validation/Replication	Reference
0-18 (month)	Buccal swabs from 10 MZ and 5 DZ	Paired t-test, FDR<0.05 and delta beta> 0.2	99,198 a-DMPs with 3.1% methylation changes, 2,632 with >20% changes.	EpiTYPER	(Martino <i>et al.</i> , 2013)
0-100	82 datasets across multiple tissues	Penalized regression model	353 a-DMPs successfully defined DNA methylation age across multiple tissues.	NA	(Horvath, 2013)
50-75	WBC from 400 (observatory), 498 (replication), and 67 (8 years apart longitudinal) subjects	Rank correlation, mixed linear regression at Bonferroni-corrected P = $2.5 \times 10^{-4}$	162 a-DMPs in both cohorts, also more than 96% of these are to the same effect in the longitudinal cohort.	NA	(Florath <i>et al.</i> , 2014)
18-27; 68-89	Muscle tissue from 24 healthy younger and 24 older subjects	Modified t-test	2,114 genes with at least 1 a-DMP.	NA	(Zykovich <i>et al.</i> , 2014)
22-25; 77-78	CD14+ from 8 younger and 8 healthy elder subjects (8 males and 8 females)	FDR<0.05 and delta beta>0.2	368 a-DMPs identified, 26 are with difference > 0.2, and the majority are hypo-methylated.	EpiTyper in 10 younger and 10 elder subjects. 3 a-DMPs validated	(Tserel <i>et al.</i> , 2014)

One application of a-DMPs was to formulate a prediction model for chronological age. Several studies have developed such an age prediction model, and surprisingly with a few a-DMPs, the chronological age could be predicted to a high degree of accuracy. For example, using two CpG sites at the *EDARADD* and *NPTX2* genes, Bocklandt *et al* built a regression model that explained 73% of the variance in age with an error of 5.2 years (Bocklandt *et al.*, 2011). Another study proposed a model using 17 a-DMPs, which explained 78% of the variance in age with an error of 2.6 years for predicting age (Florath *et al.*, 2014). Two more studies built predictive models using larger training samples and a penalized multivariate regression method. Using blood methylation in 656 individuals on the Illumina 450 array, Hannum *et al* developed a predictive model using 71 a-DMPs with 96% accuracy rate and an error of 3.9 years. Horvath used 8000 samples from 82 public Illumina methylation arrays (Illumina 27k and 450k arrays) using training datasets of different tissues and age ranges (Horvath, 2013). Furthermore, the elastic net model selected 353 so-called ‘clock CpGs’ as predictors of chronological age. The model was applied to multiple test datasets for age prediction (DNA methylation age). The chronological age and predicted methylation ages were highly correlated ( $r = 0.96$ ) across a number of datasets. The average absolute difference between observed and predicted age was 3.6 years.

The two latter studies proposed measures based on the predicted methylation age (the apparent methylomic aging rate (AMAR) and the age acceleration) as markers of biological age or age-related phenotypes (Hannum *et al.*, 2013; Horvath, 2013). In both studies these methylation-based measures of biological age were higher in tumours, compared to control samples. These results together suggest that methylation-based measures of biological age could also be explored as markers of other age-related phenotypes or diseases.

Therefore, I have assessed the methylation age acceleration in all the three tissue samples, and correlated these with the age-related phenotypes and diseases. The majority of age acceleration correlations with phenotypes had occurred in blood (see below). In the adipose tissue, an interesting result was the negative association between the age acceleration and adiponectin ( $r = -0.15$ ,  $P$  value = 0.017). It is known that adiponectin is a hormone that plays an essential role in the control of type 2 diabetes and atherosclerosis, and there is increased risk if the adiponectin level is reduced.

Similarly, the subjects with lower QT level (cardiac function test -ECG) also have faster age acceleration

For age-related diseases, such as type 2 diabetes and hypertension, there were no acceleration differences, but the samples included in this study were not selected to have disease, so the sample size was very small. I further compared age acceleration with an example of environmental exposure, smoking status, and there was higher age acceleration in current smokers however it did not reach nominal significance.

It remains to be seen whether age acceleration truly serves any biological meaning as it was based on chronological age. Furthermore, there could be a problem using the raw methylation betas to predict the methylation age, without careful adjustment for batch effect, for example. Batch effects existed in the Illumina 450k datasets, particularly in the blood dataset, and likely contributed variation into the data (Figure 8). Future studies should address this problem and incorporate further covariates of blood (such as white blood cell types).

In conclusion, thousands of a-DMPs were identified in all datasets, and 231 a-DMPs are persistently associate with chronological age across different tissues, and an at least 60% of tissue-shared age effect is found between tissues. Using a subset of Illumina 450k probes, the methylation levels can successfully predict the true age of one individual. I also found the age acceleration estimates are associate with several age-related phenotypes, but the biological role in methylation age requires further research.



# Epigenome-Wide Association Scans And Longstanding DNA Methylation Changes Related to Birth Weight in Discordant Monozygotic Twin Pairs

---

In this chapter, I have investigated MZ twin pairs with a substantial weight difference at birth. The hypothesis is that birth weight likely reflects intrauterine growth restriction and should leave a footprint on DNA methylation changes *in utero* and persist into adulthood. To test this hypothesis, I performed a discovery birth weight EWAS (BW EWAS) on 20 MZ pairs discordant for birth weight on the Illumina 450k array. This was followed by a replication in 25 female MZ pairs and further verified in a third dataset of 310 unrelated subjects. In twin design, I performed two EWAS: considering the actual birth weight difference effect or categorize the twins as cases and controls. In general, none of the birth weight-related differential methylations were identified in both twin designs or in the replication of whole population.

I have performed a further analysis using a bigger sample size of 71 MZ BW discordant pairs recently and found a BW differential methylation site on *IGF1R* gene. I also validated this signal in another two independent cohorts. The results will be published in the Twin Research and Human Genetics journal in 2015 (Tsai *et al.*, 2015).

---

## 5.1 Introduction

Low birth weight (LBW) is defined as weight at birth lower than 2500 grams. It directly influences the outcome of childhood mortality (McCormick, 1985) and morbidity (Y. W. Wu *et al.*, 2011) and childhood asthma (Brooks *et al.*, 2001). It is further associated with disorders progressing into the adulthood, such as metabolic syndrome (Fagerberg *et al.*, 2004), type 2 diabetes (Johansson *et al.*, 2008), cardiovascular diseases (Leeson *et al.*, 2001), respiratory diseases (Walter *et al.*, 2009), and depression (Thompson *et al.*, 2001). The ‘foetal origins hypothesis’ postulates that nutritional exposures during pregnancy pre-program the foetus to develop specific diseases in adulthood (Barker, 1992). There is as much clinical evidence in support of this view as there are many questions asking whether there is in fact any casual relationship (K. Christensen *et al.*, 1995; Williams & Poulton, 1999; Phillips *et al.*, 2001; Rasmussen, 2001) since other studies report absent or weak association and a lack of replication (Skidmore *et al.*, 2004; Wojcik *et al.*, 2013; Yang *et al.*, 2013).

Two important factors that influence birth weight are the length of gestation and prenatal growth rate. Other modifiable factors also exist *in utero*, such as maternal smoking during pregnancy and maternal health (e.g. caloric intake). Several genetic variants have also been associated with birth weight (Freathy *et al.*, 2010; Horikoshi *et al.*, 2013). However, it is estimated that genetics can only explain a modest contribution to the total variance in birth weight (Battaglia & Lubchenco, 1967; McIntire *et al.*, 1999; Barker, 2004; Jarvelin *et al.*, 2004; Heijmans *et al.*, 2008; Freathy *et al.*, 2010; Horikoshi *et al.*, 2013). Given its relevance in predicting health in old age, understanding the molecular links between birth weight and age-related disease has attracted much attention recently.

Epigenetic changes may relate to low birth weight, and increasing evidence shows that maternal nutrition intake is a key factor. Researchers have investigated this question using animal models, and results suggested that nutritional changes could cause methylation changes and associate with the obesity and diabetes later in life (see (Seki *et al.*, 2012)). In humans, DNA methylation analysis of adults born during the Dutch Winter Famine identified significantly lower methylation at CpG sites of *IGF2* in subjects born during the Dutch Hunger Winter (1944) compared to same-sex siblings

who were not born during the famine (Heijmans *et al.*, 2008). However, a follow-up study that compared the methylation levels on six genes (including *IGF2*) between the high and low birth weight subjects found no significant differences between groups (Tobi *et al.*, 2009).

Other targeted gene studies have focused on birth weight related methylation changes in imprinted genes, such as *IGF2* and *H19* (Stegers-Theunissen *et al.*, 2009; Hoyo *et al.*, 2012) and glucocorticoid receptor *NR3C1* (Filiberto *et al.*, 2011; Mulligan *et al.*, 2012). Supporting evidence shows an association with *IGF2* methylation, but not with *H19*. In one study, an inverse relationship between *IGF2* methylation changes and birth weight was identified, and was independent from folic acid intake during pregnancy (Stegers-Theunissen *et al.*, 2009). Another study showed more indirect evidence that *IGF2* methylation was negatively correlated with IGF protein levels, and *IGF2* protein was positively associated with birth weight, suggesting that there might be a negative correlation between *IGF2* methylation and birth weight (Hoyo *et al.*, 2012). Two studies using cord blood (Mulligan *et al.*, 2012) and placenta samples (Filiberto *et al.*, 2011) found increased methylation of CpG sites in the *NR3C1* gene associated with low birth weight in infants. Another candidate was *WNT2* that also reported to be differentially methylated with birth weight (Ferreira *et al.*, 2011).

Recently, several EWAS studies on birth weight were performed. Most were conducted on cord blood or placenta of population-based unrelated subjects using the Illumina 27k (Banister *et al.*, 2011; Fryer *et al.*, 2011; Adkins *et al.*, 2012) and Illumina 450k (Engel *et al.*, 2014). There was no clear association between birth weight and imprinted genes *IGF2*, *H19* and other growth-related genes using the Illumina 27k and Illumina GoldenGate (Turan *et al.*, 2012). A study found 22 birth weight-differential methylation sites (BW-DMPs) after comparing SGA (small for gestational age) to normal size newborns (Banister *et al.*, 2011). Another found no strong association between birth weight and genome-wide methylation levels (Adkins *et al.*, 2012). In a large sample of 1,046 infants from the Norwegian Mother and Child Cohort Study using the Illumina 450k, researchers found 19 CpG sites significantly associated with birth weight, and after adjusting for the leukocyte cell-type proportion, 8 CpG sites remained significant (Engel *et al.*, 2014). In summary, these birth weight EWAS studies found few CpG sites associated with birth weight in newborns, and the magnitude of methylation changes in

these regions tended to be small. A potential reason was that these studies were mostly performed in unrelated subjects, and the maternal environment and exposures have not been appropriately adjusted. In this case, the methylation differences in MZ twins discordant for birth weight would be the ideal study design.

Weight differences in twin pairs are thought to originate from random differences in terms of foetal access to nutrition, which is affected by the position of the foetus *in utero* as well as the position of the umbilical cord. Generally, singletons and twins develop at a similar rate until the 30th week of gestation, after which uterine restrictions become a contributing factor (Cleary-Goldman & D'Alton, 2008). DZ twins almost always undergo development in two separate placentas and differences in access to nutrients due to foetal mass or placental lesions are usually moderate. On the other hand, MZ twins, who originate from the division of a single ovum post-fertilization, may have one or two chorions, and severe differences in nutritional intake due to improper positioning or umbilical cord insertion often lead to greater weight discrepancies. Thus, a lighter MZ twin has a greater likelihood of being genuinely growth restricted (Torche & Echevarria, 2011).

At the time of writing this thesis, only two birth weight discordant MZ EWAS studies have been conducted. One used the Illumina 27k and birth weight discordant 22 MZ and 12 DZ twin pairs (Gordon *et al.*, 2012). The study examined methylation levels of three tissues, and found 1 BW-DMP in 14 MZ pairs (human umbilical vascular endothelial cells) and 7 in 9 DZ pairs (cord blood mononuclear cells) but no significance in placental tissue. These 8 BW-DMPs reached genome-wide significance and associated with metabolic disease (Gordon *et al.*, 2012). The other study used Illumina 450k to compare the methylation levels of heavy and light co-twins in 17 monozygotic MZ adult twins in buccal samples. The study identified 3,153 BW-DMPs at  $P < 0.01$ , but none of these reached genome-wide significance (Souren *et al.*, 2013).

In this study, I have assumed that the causes of significant weight deviation in MZ twins are due to the competition for prenatal resources, such as nutrients. Nutrients, such as folate, Vitamin B are the dietary sources for methyl group. Diet rich in these methyl-donating nutrients can rapidly alter gene expression, especially during early development, when epigenome is being established. Therefore, the twin who loses the competition of nutrients would have lower birth weight, as well as different methylation

patterns compared to their co-twin. I assumed that variable methylation regions occur in the early development could pass on until their adulthood, and potentially link to risk of phenotype changes in later life.

## **5.2 Materials and Methods**

### **5.2.1 Datasets**

I used three datasets in this study: the discovery and replication datasets were used for the discordant MZ twin analysis and the verification dataset is used for population-based analysis.

#### ***5.2.1.1 Discovery dataset (BW discordant MZ twins)***

The discovery dataset included 20 female Caucasian MZ pairs discordant for birth weight. They were selected from a dataset of 355 subjects (Dataset 2, as described in Chapter 3). The discordancy was defined as a birth weight difference that exceeded the 70th percentile (0.45 kg) in the TwinsUK birth weight distribution, based on the birth weight differences of 3,010 MZ twin pairs. Twin pairs where both twins weighed less than 2 kg at birth were considered to be extremely low birth weight and were excluded from my study. All subjects were free from severe disease when their blood samples were collected.

#### ***5.2.1.2 Replication dataset (BW discordant MZ twins)***

The replication dataset included 25 female birth weight discordant MZ pairs, selected from 508 subjects in the TwinsUK cohort (Dataset 3, as described in Chapter 3). Similar to the observatory dataset, the MZs were selected with birth weight discordancy of 0.45 kg. These samples were more discordant for birth weight than the observatory twins.

### ***5.2.1.3 Verification dataset (unrelated female subjects)***

A verification dataset of 355 female subjects (some are twin pairs) was selected from the TwinsUK cohort to see if the top findings could be reproduced in a normal birth weight population. After excluding subjects with missing birth weight, BMI, and white blood cell (WBC) subtype, altogether 310 subjects were retained for analysis, and these included all individuals in the observatory dataset.

## **5.2.2 Phenotypes**

Questionnaire data included details of birth weight, medical history, height, weight, and BMI at visit when the subject's blood samples were collected. Due to the fact that DNA methylation levels might change over time, the age of DNA extraction has been carefully considered in my analysis as a covariate. Previous studies reported that the composition of white blood cells (WBC) might contribute to the DNA methylation levels (Houseman *et al.*, 2012). So I obtained WBC counts from fluorescence-activated cell sorting of peripheral blood, and the sub-types of specific cell counts were calculated by multiplying each cell type count by the total WBC cells. The four cell types obtained were: eosinophils, lymphocytes, monocytes, and neutrophils.

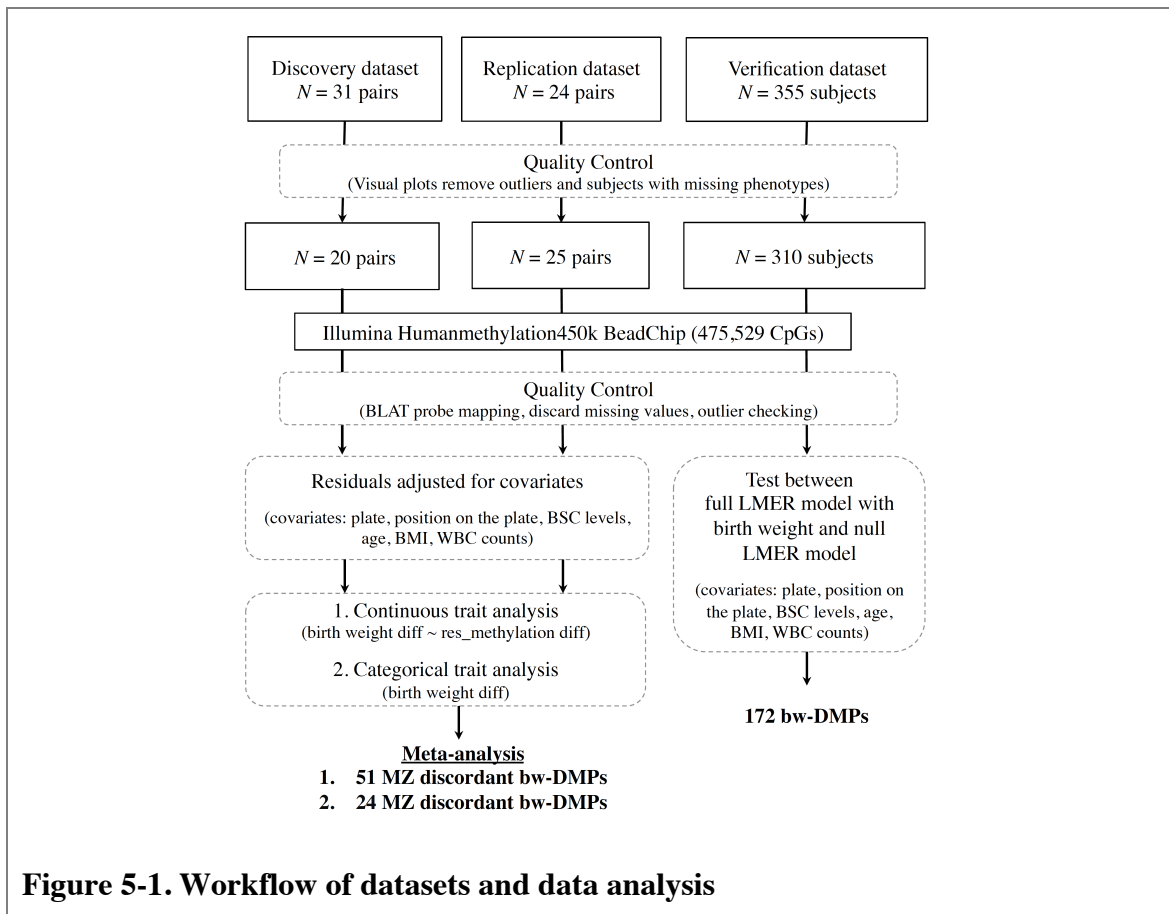
## **5.2.3 Methylation data**

The analysis was performed on 420,000 probes from Illumina 450K in all three datasets. To annotate each probe to a human gene, I mapped the probes to the gene body of all human genes (GRCh37) and extended to 30kb upstream and downstream of the gene start and end. The methylation levels were obtained from the Illumina 450k array as described in Chapter 3.

## 5.2.4 Statistical analyses

### 5.2.4.1 Quality Control for Illumina 450k data

An overview of my quality control analysis in this chapter is shown in Figure 5-1. Data from all subjects underwent quality control checks and normalization as described in Chapter 3, and briefly in Figure 5-1 below.



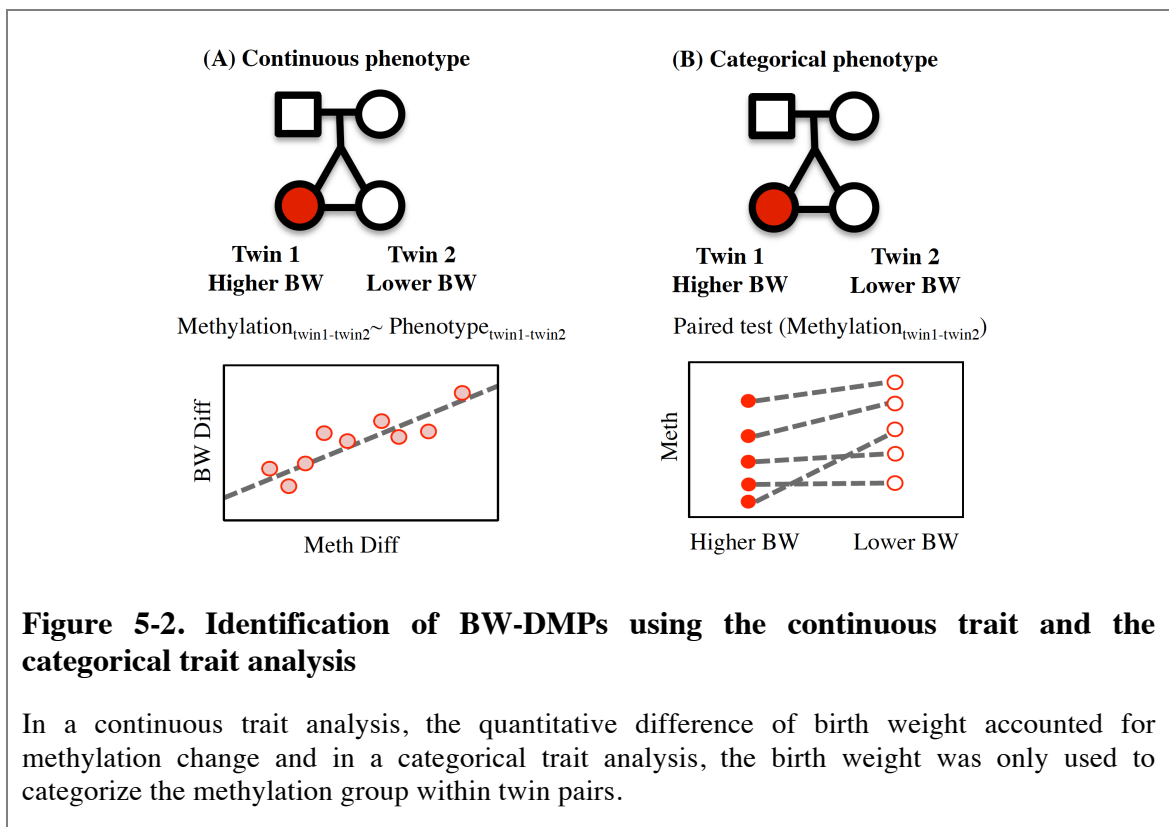
### 5.2.4.2 Birth weight differentially methylated positions (BW-DMPs) analysis

For both twin datasets (observatory and replication) methylation levels were normalized to follow normal distribution at each probe. The methylation residuals were then taken after adjusting for all covariates using the linear model that included plate, position on the plate, bisulfite conversion levels, age at DNA extraction, BMI, and WBC counts. The bisulfite conversion levels in the replication dataset were not included in the model due to missing data. The residuals of the discordant twin pairs from both twin datasets were extracted.

There were two ways of approaching the birth weight analysis using MZ twins: (A) treat birth weight as a continuous trait; or (B) treat birth weight as a categorical trait (and define the high and low phenotype co-twin based on the birth weight level). Figure 5-2 shows a simplified concept of these two types of analyses.

In the first analysis (A), comparing methylation differences with quantitative birth weight differences (or ‘continuous trait analysis’; Figure 5-2A), the difference of the residuals, calculated within twin pairs for all probes (methylation in the higher birth weight twin minus methylation in the lower birth weight twin), was correlated with the birth weight difference (higher birth weight minus lower birth weight) to identify BW-DMPs. The Pearson’s correlation coefficients ( $r$ ) and P values were reported.

In the second analysis (B), comparing methylation differences with qualitative birth weight differences (or ‘categorical trait analysis’; Figure 5-2B), the analysis overlooked the actual birth weight differences and considered methylation differences between the twins. The methylation residuals were taken between the twins (methylation of the higher birth weight minus the methylation of lower birth weight co-twin) and a paired t-test was performed on these differences.





Following the continuous trait analysis, I sought to observe the methylation levels of the BW-DMPs in the whole population. The 310 recruited subjects included the twin pairs from the observatory (N = 40 from 20 MZ twins) for this analysis. The raw methylation levels were first normalized then fitted to a full linear mixed effect model with methylation as an outcome and birth weight as the predictor. The full model was compared to the null model using a linear mixed effects model adjusting for both fixed and random effects (family structure, zygosity). The significant BW-DMPs were defined by comparing this model to the null model that excluded birth weight. The epigenome-wide significance was fixed at  $P < 1.05 \times 10^{-7}$  after Bonferroni correction.

#### ***5.2.4.3 Meta-analysis of twin datasets***

To compare the top hits from the twin datasets (observatory and replication), a meta-analysis was performed using METAL (Willer *et al.*, 2010), considering the sample size, effect direction, and P value. Among the 475,529 overlapping probes between the two datasets, only probes with the same direction of correlation from both datasets were considered as BW-DMP candidates.

#### ***5.2.4.4 Age differentially methylated positions (a-DMPs) analysis***

The hypothesis was that substantial differences in birth weight of MZ twins have lasting BW-DMPs into adulthood. So the methylation levels in these regions should be independent of age. To test this, I used the a-DMPs results (Chapter 4) to see whether BW-DMPs were also a-DMPs. The a-DMPs were defined by comparing the full model with age to the null model without age. Only the top hits from the BW-DMP meta-analysis were compared with a-DMPs.

#### ***5.2.4.5 Gene clustering analysis***

To find if the top genes were enriched for disease and related to biological pathways, I used the following online tools: gene ontology term enrichment analysis and disease-related enrichment analysis using WEB-based GEne SeT AnaLysis Toolkit (WebGestalt).

## 5.3 Results

### 5.3.1 The demographic characteristics of the twin datasets

Table 5-1 shows the sample size, age, and birth weight range of the two twin datasets. For the observatory dataset, the mean birth weight difference was  $0.67 \pm 0.29$  kg. The relative birth weight ranged from 16.6% to 51% (relative birth weight = ratio of absolute birth weight difference within pair over birth weight in heavier co-twin). For the replication dataset, more discordant twin pairs were selected, and these had a mean birth weight difference is  $0.83 \pm 0.40$  kg and the relative birth weight ranged from 12.5 to 58.2%. The subjects with a lighter birth weight were significantly shorter in height than their co-twins in both datasets, which is similarly seen in previous low birth weight studies. No significant difference was found for weight and BMI.

**Table 5-1. Characteristics of the two twin datasets**

Dataset	Discovery			Replication		
	Heavier	Lighter	P value*	Heavier	Lighter	P value*
N	20			25		
Age	50.70 (42, 64)			54.8 (39, 72)		
BW	2.32 (1.13, 3.49)			2.33 (1.13, 3.63)		
BW diff	0.67 (0.452, 1.475)			0.83 (0.454, 2.014)		
BW	$2.65 \pm 0.36$	$1.98 \pm 0.38$	$4.4 \times 10^{-9}$	$2.74 \pm 0.43$	$1.91 \pm 0.53$	$4.8 \times 10^{-10}$
Height	$163.8 \pm 5.7$	$160.8 \pm 6.3$	$8 \times 10^{-4}$	$160.5 \pm 6.3$	$158.7 \pm 7.3$	0.007
Weight	$70.1 \pm 15.5$	$69.5 \pm 12.5$	0.779	$75.8 \pm 17.7$	$71.9 \pm 15.5$	0.103
BMI	$26.3 \pm 4.2$	$26.7 \pm 4.6$	0.663	$29.5 \pm 6.9$	$28.9 \pm 5.7$	0.381

Abbrev: N, number of MZ twin pairs; Age: DNA extraction age; BW, birth weight (kg); BW diff: birth weight difference within twin pairs (kg); Height (cm); Weight (kg); BMI: body mass index. Numbers in parenthesis indicates the range of the phenotype. \*Result from paired t-test.

### 5.3.2 Birth weight differentially methylated positions (BW-DMPs)

#### 5.3.2.1 Identification of BW-DMPs using ‘continuous trait analysis’

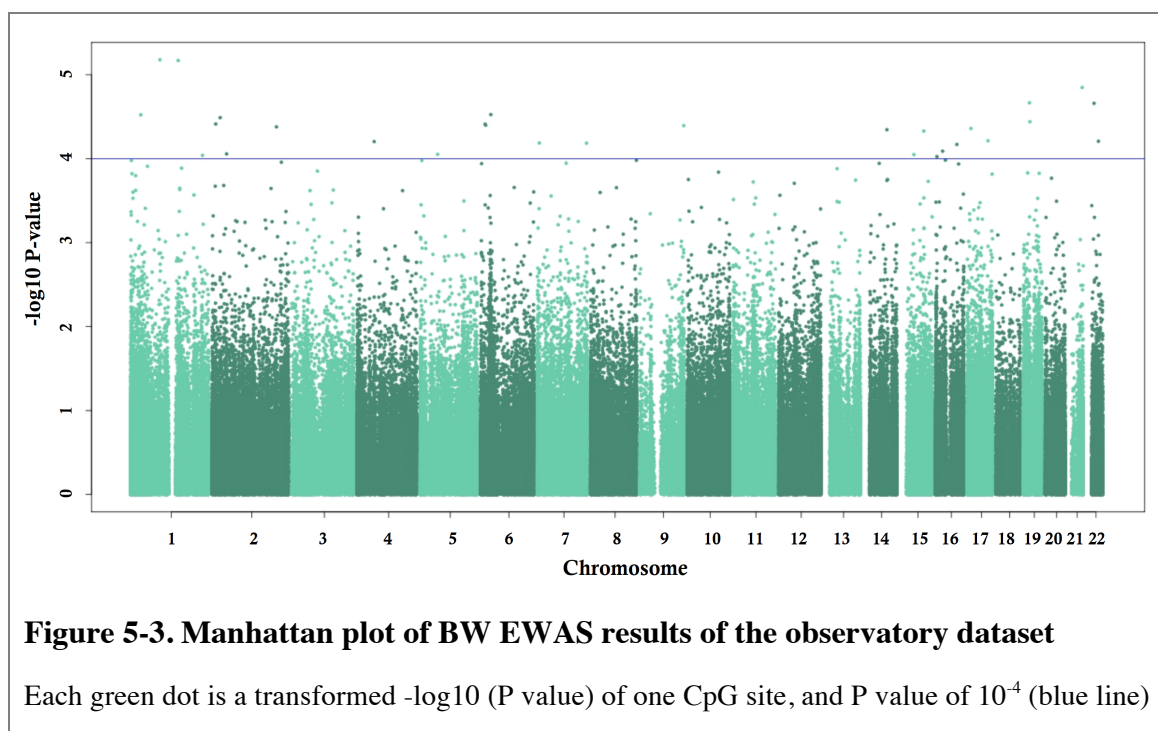
The methylation differences within pairs to absolute birth weight differences was compared using the Pearson’s correlation to test the hypothesis that prenatal conditions contribute to birth weight differences and changes to DNA methylation in twins. Table 5-2 shows a summary of the BW-DMPs at varying significance in the observatory, replication, and meta-analysis.

**Table 5-2. Summary of BW-DMPs discovered in the two twin datasets**

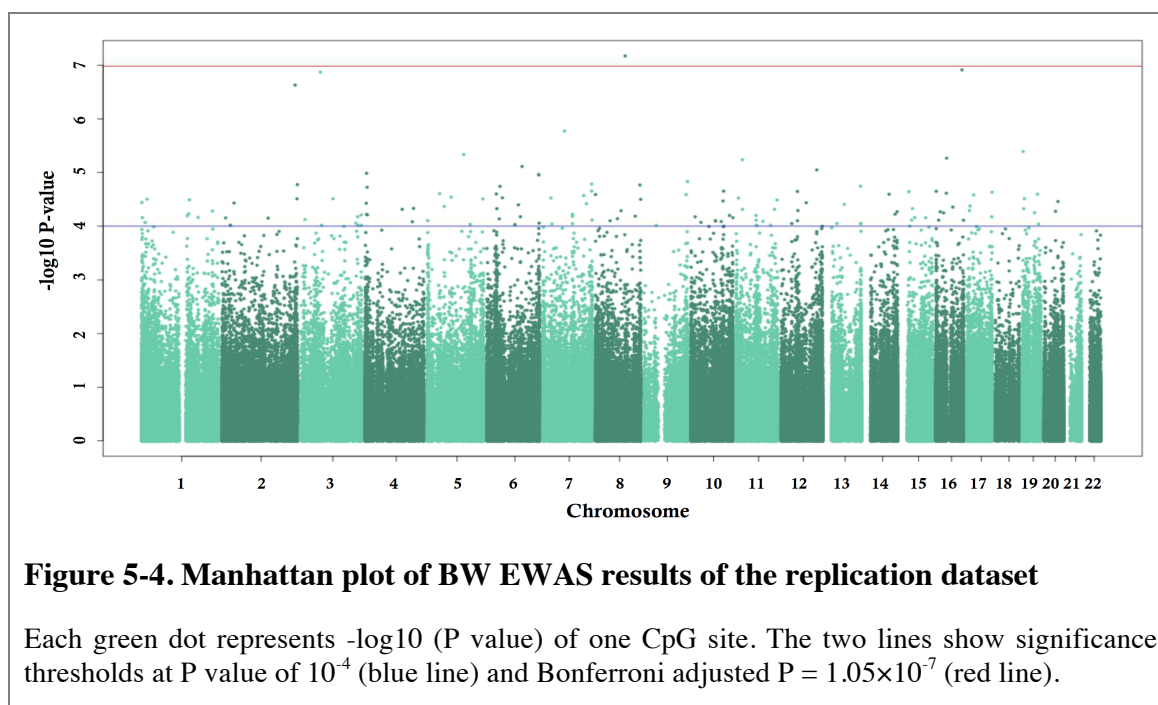
Datasets	Observatory (20 MZ)	Replication (25 MZ)	Meta-analysis
CpG sites	Total (hyper, hypo)	Total (hyper, hypo)	Total (hyper, hypo) <sup>2</sup>
Total	475,529 (51.9%, 48.1%)	475,529 (45.7%, 54.3%)	241,967 (47.6%, 52.4%)
Max P	$6.63 \times 10^{-6}$	$6.71 \times 10^{-8}$	$8.27 \times 10^{-7}$
P < $10^{-5}$	2 (0%, 100%)	13 (38.5%, 61.5%)	6 (16.7%, 83.3%)
P < $10^{-4}$	29 (51.7%, 48.3%)	150 (34.7%, 65.3%)	51 (39.2%, 60.8%)
P < $10^{-3}$	234 (47.9%, 52.1%)	1117 (36.6%, 63.4%)	511 (39.7%, 60.3%)
P < $10^{-2}$	2847 (51.4%, 49.6%)	7513 (37.4%, 62.6%)	4709 (40.8%, 59.2%)
Bonf <sup>1</sup>	0	1	0
FDR 5%	0	4	0

Each cell contains the number of probes that passed the significance criteria. Numbers in parenthesis are the percentages of hyper-methylated (left) and hypo-methylated (right) probe. In meta-analysis, only probes with same effect directions in both datasets were considered. <sup>1</sup>Number of significant probes reached Bonferroni-adjusted P value at  $1.05 \times 10^{-7}$ ; <sup>2</sup>Numbers in parenthesis are probes that were consistently hyper- or hypo-methylated in both datasets. Among the 475,529 overlapped probes, only 241,967 probes were with the same effect direction. Max P, means most significant P value.

In the observatory dataset, none of the probes satisfied the Bonferroni adjusted P value for multiple testing  $P = 1.05 \times 10^{-7}$  or the 5% FDR, since the most significant locus was associated observed at  $P = 6.62 \times 10^{-6}$  (Table 5-2). Overall, there were slightly more hyper-methylated probes (51.9%) in the observatory dataset than hypo-methylated probes (48.1%). Figure 5-3 shows the Manhattan plot of the EWAS in the observatory dataset.



The EWAS results in the replication dataset were similar to those in observatory dataset. The proportion of hyper- and hypo-methylated probes was roughly half-half. However, there were a higher proportion (> 60%) of hypo-methylated BW-DMPs at different significance levels, and nearly 5-fold increase BW-DMPs identified in the replication dataset, potentially due to the larger sample size and greater birth weight discordance on average. In total, there was one BW-DMP (cg06699564,  $\beta = -0.852$ ;  $P = 6.71 \times 10^{-8}$ ; Table 5-2) that surpassed the Bonferroni-adjusted P value threshold ( $P = 1.05 \times 10^{-6}$ ). This CpG site is located on chromosome 8 (cg06929843) in *LOC100288748*, and about 2 kb away from the transcription start site of the *ESRP1* gene. The *ESRP1* gene is a splicing regulator in epithelial cells and has been associated with carcinoma (Yae *et al.*, 2012). Although the effect direction at the CpG site was the same ( $\beta = -0.193$ ) in the observatory dataset, it was not significantly associated with birth weight ( $P = 0.415$ ). Four CpG-sites surpassed the genome-wide FDR 5% threshold, and these included sites in *LOC100288748* (cg06929843), *THOC7* (cg22134162), chromosome 5q11.2 (cg19673549; chr5:54916621; 5kb from the 3'UTR of *SLC38A9*), and chromosome 10q24.2 (cg07137429; chr10:100093476). Figure 4 shows the Manhattan plot of the BW EWAS in the replication dataset.

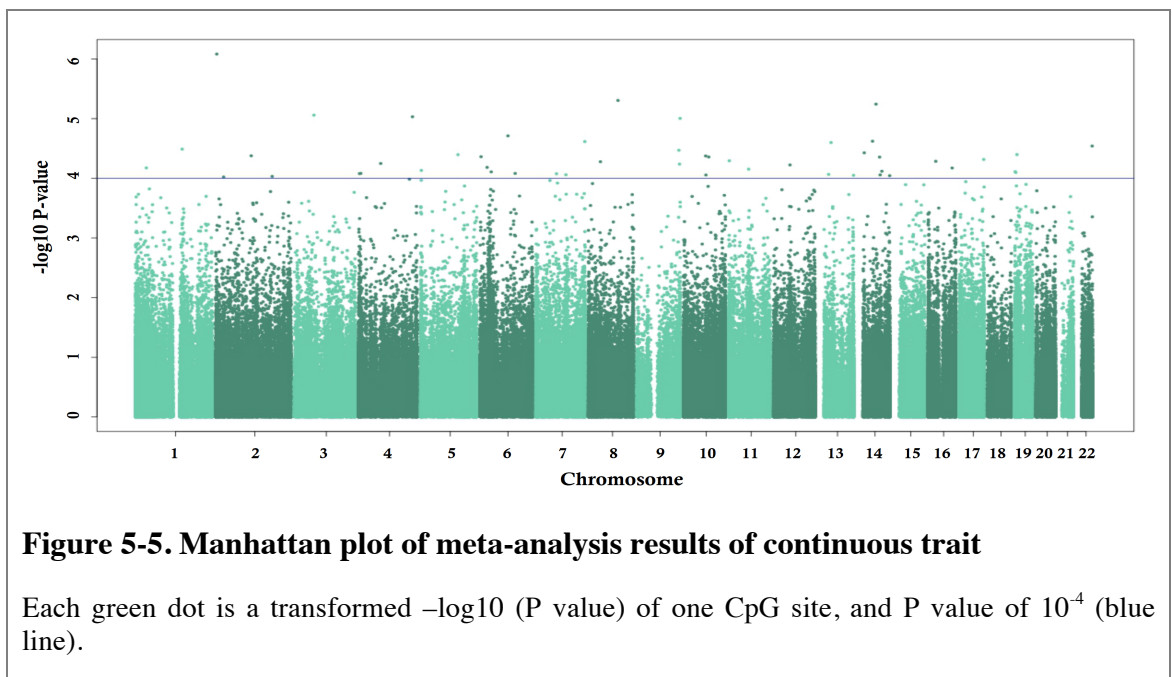


Since both datasets were small, a meta-analysis using the effects ( $\beta$ ) and P values of both BW EWASs was performed using METAL. In addition to the meta-analysis,

another test that combined both datasets (N = 45 pairs) was performed. In the second approach, the methylation residuals were taken separately from the two datasets, normalized to follow normal distribution separately then the methylation differences were taken within twin pairs. The overall methylation differences (N = 45 pairs) were then compared to their birth weight difference. Because the results were quite similar to the meta-analysis, only the meta-analysis results are reported here.

Of all CpG sites that passed quality control in both datasets, about 50.9% (N = 241,967) have the same effect directions in both datasets. In the meta-analysis results (Table 5-2) we observed that similar to the replication dataset, about 60% of the significant probes were hypo-methylated, and the P value of the most significant CpG site was  $8.27 \times 10^{-7}$  (Table 5-2). This CpG site is located on chromosome 2p25.3, roughly 5 kb and 10 kb away from the transcription end site of the *TRAPPC12* and *ADII* genes. This CpG site, however did not reach the Bonferroni adjusted P value or the 5% false discovery rate threshold.

Given this disappointing result, I next explored the potential birth weight-associated genes in specific biological pathways. To do so, I selected the top 51 probes (within 10 kb of the transcription start site or in the gene body of 39 unique genes) identified in the meta-analysis at a less stringent significance level of  $P < 10^{-4}$ . The details of the top 51 probes are shown in Table 5-3 and in the Manhattan plot in Figure 5-5.



**Table 5-3. Top 51 probes found in meta-analysis**

<b>IlmnID</b>	<b>CHR</b>	<b>Gene Name</b>	<b><math>\beta</math>_Discovery</b>	<b><math>\beta</math>_Replication</b>	<b>P value</b>
cg26174880	2	-	-0.708	-0.646	8.27E-07
cg06699564	8	-	-0.193	-0.852	4.95E-06
cg12165758	14	<i>PRKCH</i>	-0.574	-0.677	5.71E-06
cg22145181	3	-	-0.184	-0.842	8.73E-06
cg01324261	4	<i>SCRG1</i>	0.638	0.610	9.30E-06
cg14410072	9	<i>C8G;FBXW5</i>	-0.786	-0.450	9.90E-06
cg23366832	6	<i>BACH2</i>	0.459	0.702	1.94E-05
cg01510588	14	<i>CI4orf183</i>	0.756	0.443	2.38E-05
cg12415687	7	<i>PTPRN2</i>	-0.588	-0.608	2.42E-05
cg21258821	13	<i>KBTBD6</i>	0.722	0.481	2.52E-05
cg26621897	X	<i>TMSB15A</i>	0.363	0.739	2.74E-05
cg12961733	22	-	-0.515	-0.653	2.88E-05
cg05630111	1	<i>LASS2</i>	-0.555	-0.620	3.24E-05
cg06866628	9	<i>GTF3C5</i>	-0.350	-0.738	3.38E-05
cg16280098	14	<i>PABPN1</i>	-0.667	-0.519	3.73E-05
cg15222563	19	<i>TRAPPC5</i>	0.534	0.625	4.00E-05
cg06062821	5	-	0.561	0.606	4.01E-05
cg02120071	2	-	0.486	0.655	4.20E-05
cg06973667	10	<i>NEUROG3</i>	0.473	0.663	4.22E-05
cg15323253	6	-	0.756	0.410	4.35E-05
cg17045635	10	<i>ZMIZ1</i>	0.640	0.536	4.39E-05
cg07171024	14	<i>DPF3</i>	-0.595	-0.575	4.40E-05
cg12846139	17	<i>BAHCC1</i>	-0.747	-0.415	4.83E-05
cg27261733	11	<i>LSP1</i>	-0.540	-0.610	5.08E-05
cg04022912	16	<i>SLC5A11</i>	0.362	0.716	5.18E-05
cg05789704	8	<i>ADAM32</i>	-0.556	-0.596	5.27E-05
cg02971882	4	<i>UTP3</i>	-0.502	-0.632	5.65E-05
cg14333779	9	<i>OLFM1</i>	-0.472	-0.650	5.78E-05
cg19880852	12	<i>KRT71</i>	0.370	0.707	5.97E-05
cg10588720	6	-	-0.388	-0.693	6.56E-05
cg00416130	1	<i>GJA4</i>	-0.586	-0.562	6.68E-05
cg04260557	16	<i>VATIL</i>	0.464	0.648	6.71E-05
cg17059853	11	<i>NRXN2</i>	-0.345	-0.713	7.02E-05
cg15192932	X	<i>PLXNB3</i>	-0.478	-0.637	7.06E-05
cg17842821	5	-	-0.439	-0.660	7.37E-05
cg14975009	14	<i>DIO2</i>	-0.501	-0.619	7.61E-05
cg22088518	19	<i>ONECUT3</i>	-0.184	-0.781	7.73E-05
cg24569251	6	-	-0.437	-0.659	7.80E-05
cg18661237	19	<i>MRPL54;APBA3</i>	-0.318	-0.721	8.00E-05
cg20421058	4	<i>SORCS2</i>	-0.331	-0.714	8.26E-05
cg13526264	6	-	-0.213	-0.767	8.28E-05
cg02951274	4	<i>RGS12</i>	0.536	0.589	8.34E-05
cg06602498	7	<i>RABGEF1</i>	-0.699	-0.442	8.40E-05
cg26049501	13	<i>STARD13</i>	0.643	0.498	8.60E-05
cg21754854	7	-	0.654	0.487	8.72E-05
cg20718816	14	<i>LTBP2</i>	-0.742	-0.388	8.77E-05
cg10840389	10	<i>LRRC20</i>	-0.668	-0.472	8.79E-05
cg02738641	13	<i>ATP4B</i>	0.583	0.550	8.92E-05
cg14911242	14	<i>MTA1</i>	0.573	0.557	9.03E-05
cg16057262	2	<i>ITGA4</i>	-0.551	-0.573	9.28E-05

IlmnID: Illumina probe ID; CHR: chromosome;  $\beta$ \_Discovery: beta coefficient found in the discovery dataset;  $\beta$ \_Replication: beta coefficient found in the replication dataset, P value: p value of meta-analysis results.

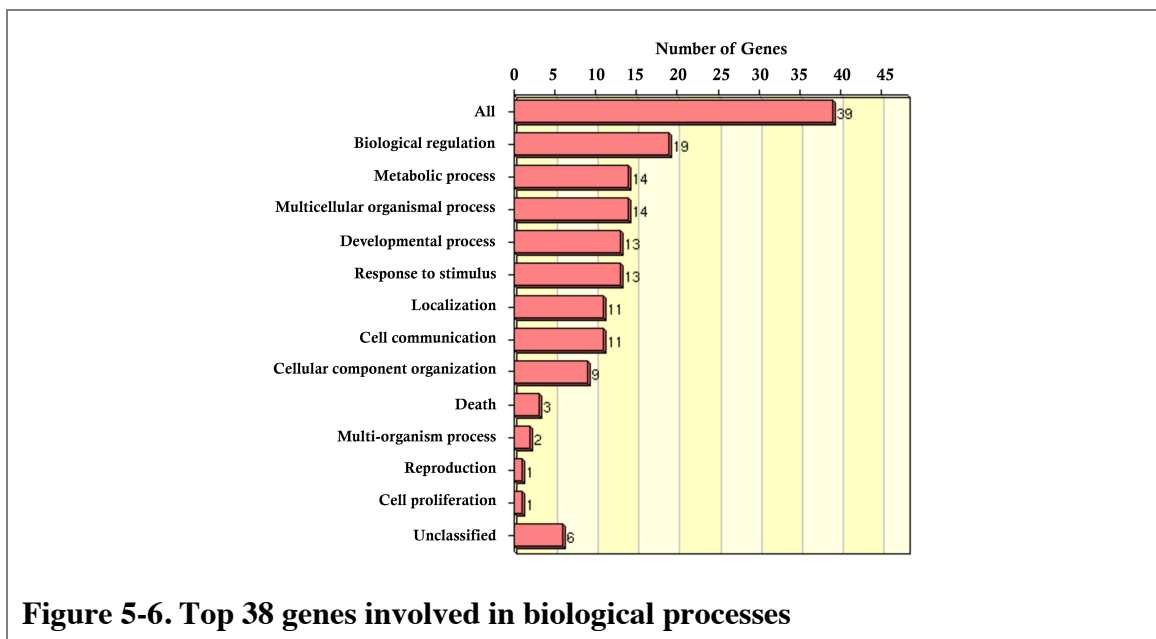
A disease enrichment analysis was performed using web-based gene set analysis toolkit by comparing the top 51 hits (in 38 genes) to the Entrez gene database. Several genes were associated with diseases. Table 5-4 lists the top diseases and genes (at least 3 genes were associated with each disease, adjusted P value < 0.01). Some of the genes were also associated with birth weight traits. For example, *GJA4* and *ITGA4* associated with infertility (adj. P value = 0.0085), and *NEUROG3* and *DIO2* associated with type 2 diabetes (adj. P value = 0.0363). In the enrichment analysis, the numbers of the observed genes were compared to the numbers of the genes in the gene set in each category (traits), and the ratio was calculated to assess enrichment. Here I have only considered the significant traits if they surpassed the BH adjusted P value (Benjamini & Hochberg, 1995) from the raw P value identified using the hyper-geometric test.

**Table 5-4. Diseases associated with top BW-DMPs**

Disease	Gene	Adj. P value*
Stroke;		0.0011
Stroke NOS;	<i>PRKCH, GJA4, LTBP2, SORCS2</i>	0.0011
Cerebral Infarction		0.0011
Subarachnoid Haemorrhage	<i>PTPRN2, RGS12, SORCS2</i>	0.0014
Diabetes Mellitus;		0.0022
Endocrine disturbance NOS;		0.0032
Endocrine system Diseases;	<i>PTPRN2, BACH2, NEUROG3, DIO2</i>	0.0032
Endocrine disorder NOS		0.0032
Autoimmune Disease	<i>PTPRN2, ZMIZ1, BACH2, ITGA4</i>	0.0032
Type I Diabetes Mellitus	<i>PTPRN2, BACH2, NEUROG3</i>	0.0032
Genetic predisposition to disease	<i>LSP1, GJA4, PTBP2, ZMIZ1, BACH2</i>	0.0039
Infarction	<i>PRKCH, GJA4, SORCS2</i>	0.005
Metabolic diseases	<i>PTPRN2, BACH2, NEUROG3, DIO2</i>	0.0063
Skin Diseases (genetic)	<i>LSP1, ZMIZ1, MTA1</i>	0.0076

\*Adj. P value: P value adjusted by the BH test

Figure 5-6 shows a summary of the biological categories that the 39 genes were involved in. Among these genes, 14 were involved in metabolic processes, and 13 were involved in developmental processes.



### 5.3.2.2 Identification of BW-DMPs using ‘categorical trait analysis’

Next, I proceeded to perform a categorical trait analysis. In this case, the methylation difference was calculated as the methylation in the higher birth weight twin minus the lower birth weight co-twin. There were no statistically significant CpG sites found in the observatory, replication, or meta-analysis results. In the observatory dataset, the hyper- and hypo-methylated sites were similar to the results of the continuous trait analysis. However, an increased proportion of hyper-methylated sites were seen in the replication dataset and in the meta-analysis results. A summary of the BW-DMPs discovered in all three analyses is listed in Table 5.

**Table 5-5. Summary of the BW-DMPs discovered in twin datasets**

Datasets	Discovery (20 MZ)	Replication (25 MZ)	Meta-analysis
CpG sites	Total (hyper, hypo)	Total (hyper, hypo)	Total (hyper, hypo)*
Total	475,529 (48.6%, 51.4%)	475,529 (56.8%, 43.2%)	237,081 (55.1%, 44.9%)
Max P	$1.24 \times 10^{-6}$	$1.15 \times 10^{-5}$	$5.72 \times 10^{-6}$
$P < 10^{-5}$	3 (33.3%, 66.9%)	0	1 (100%, 0%)
$P < 10^{-4}$	24 (54.2%, 45.8%)	29 (72.4%, 27.6%)	24 (62.5%, 37.5%)
$P < 10^{-3}$	366 (44.5%, 55.5%)	408 (77.9%, 22.1%)	373 (61.4%, 38.6%)
$P < 10^{-2}$	4337 (43.7%, 56.3%)	4946 (76.5%, 23.5%)	4520 (63.3%, 36.7%)

Each cell contains the total probes that passed the significance criteria. Numbers in bracket are the percentages of hyper-methylated (left) and hypo-methylated (right) probe. In meta-analysis, only probes with same effect directions in both datasets were considered. \*Numbers in bracket are probes that were consistently hyper- or hypo-methylated in both datasets. Among the 475,529 overlapped probes, only 237,081 probes were with the same effect direction. Max P, means most significant P value.

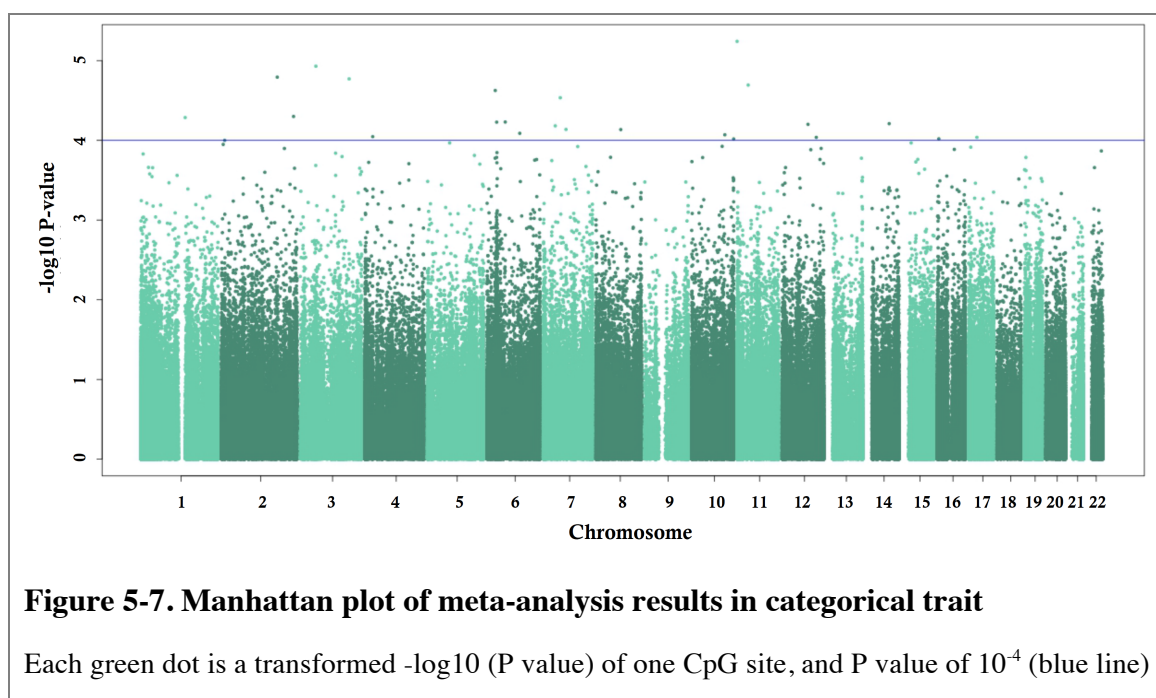


The details of the top 24 probes are shown in Table 5-6 and in the Manhattan plot in Figure 5-7. None of the top BW-DMPs had overlapped between the ‘continuous trait analysis’ and ‘categorical trait analysis’.

**Table 5-6. Top 24 probes found in meta-analysis**

IlmnID	CHR	Gene Name	$\beta$ _Discovery	$\beta$ _Replication	P value
cg14967066	11	<i>IFITM1</i>	2.452	4.948	5.72E-06
cg26384201	3	<i>HYAL3;NAT6</i>	-3.143	-3.887	1.17E-05
cg19916659	2	<i>MIR548N</i>	-3.925	-3.106	1.61E-05
cg21476666	3	<i>VEPH1</i>	3.751	3.219	1.69E-05
cg25841760	11	<i>LDLRAD3</i>	3.202	3.603	2.02E-05
cg14683235	6	<i>ZNF322A</i>	-4.267	-2.721	2.37E-05
cg23448850	7	-	-4.027	-2.806	2.92E-05
cg19184455	2	-	-2.454	-3.933	5.01E-05
cg21200085	1	-	2.422	3.952	5.17E-05
cg12723904	6	-	3.108	3.240	5.88E-05
cg13273236	6	-	2.756	3.558	5.91E-05
cg14986500	14	<i>LTBP2</i>	2.198	4.113	6.17E-05
cg12991976	12	-	-2.726	-3.559	6.29E-05
cg16039142	7	-	2.772	3.496	6.57E-05
cg03951180	7	<i>TRIM50;FKBP6</i>	-5.111	-1.816	7.29E-05
cg07290552	8	<i>FAM164A</i>	-2.666	-3.550	7.32E-05
cg17119521	6	<i>LIN28B</i>	2.558	3.607	8.16E-05
cg25083596	10	<i>SORCS3</i>	3.752	2.587	8.51E-05
cg14900814	4	<i>LG12</i>	-4.535	-2.060	8.97E-05
cg17582336	17	<i>SEBOX;VTN</i>	2.792	3.332	9.18E-05
cg02033206	12	<i>FOXP4</i>	3.698	2.596	9.18E-05
cg03684845	16	<i>LOC342346</i>	4.120	2.293	9.57E-05
cg19795151	10	-	2.835	3.274	9.60E-05
cg03467256	2	<i>HPCAL1</i>	2.234	3.847	9.99E-05

IlmnID: Illumina probe ID; CHR: chromosome;  $\beta$ \_Discovery: beta coefficient found in the discovery dataset;  $\beta$ \_Replication: beta coefficient found in the replication dataset, P value: p value of meta-analysis results.



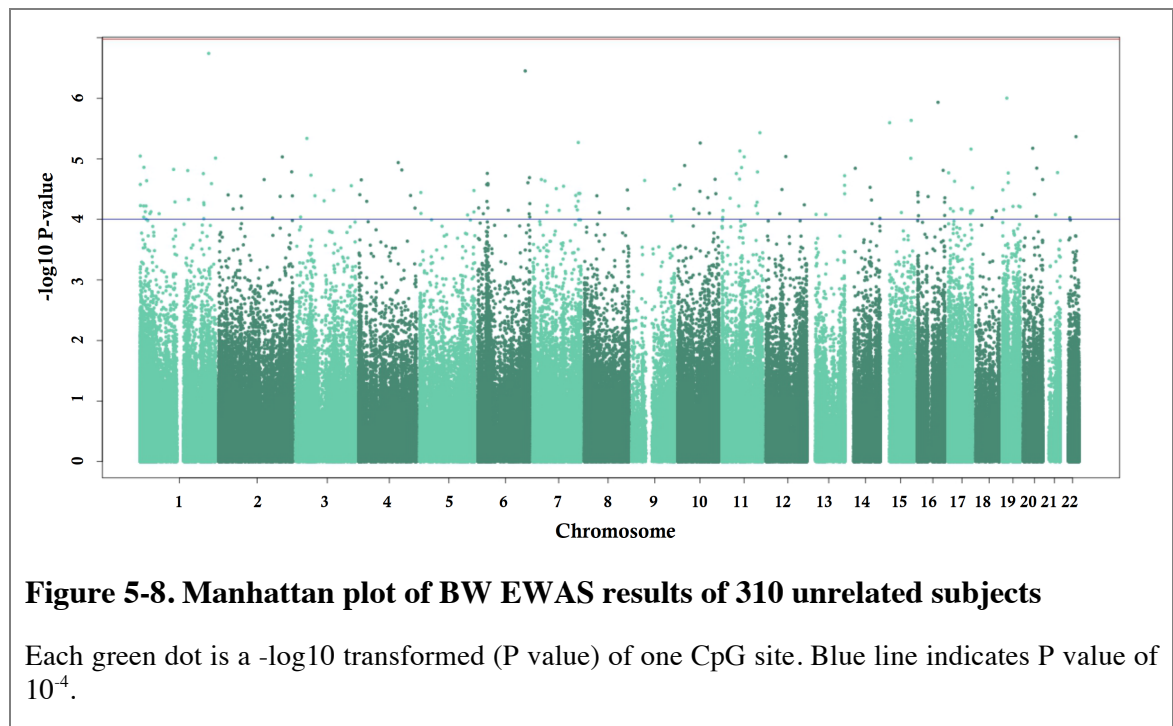
### 5.3.2.3 Identification of BW-DMPs in 310 unrelated subjects

The analysis was conducted in an expanded sample of 310 unrelated subjects to verify that the top BW-DMPs identified from the continuous trait analysis could be validated in the normal birth weight population. Compared to the discordant twin studies, there were more BW-DMPs in this larger sample size, but the top findings still did not reach the statistical significance threshold at Bonferroni adjusted P value =  $1.05 \times 10^{-7}$  (Table 5-7 and Figure 5-8). When I compared the observed effect directions here for the 51 top BW-DMPs from the continuous trait meta-analysis, 36 (64.3%) of the directions were the same, and a higher agreement (79.2%) with the BW-DMPs from the categorical analysis was found.

**Table 5-7. Significant BW-DMPs identified in 310 subjects**

CpG sites	Total (hyper, hypo)
Total	475,529 (53.3%, 46.7%)
Max P	$1.81 \times 10^{-7}$
$P < 10^{-5}$	20 (35.0%, 65.0%)
$P < 10^{-4}$	172 (56.4%, 43.6%)
$P < 10^{-3}$	631 (55.0%, 45.0%)
$P < 10^{-2}$	8166 (57.1%, 42.9%)

Each cell contains the total probes that passed the significance criteria. Numbers in bracket are the percentages of hyper-methylated (left) and hypo-methylated (right) probe. In meta-analysis, only probes with same effect directions in both datasets were considered. \*Numbers in bracket are probes that were consistently hyper- or hypo-methylated in both datasets. Among the 475,529 overlapped probes, only 237,081 probes were with the same effect direction. Max P, means most significant P value.



For the disease enrichment analysis, many genes at the top 172 CpG sites associated with previously identified birth weight related diseases, such as type 2 diabetes. Table 5-8 lists some examples of the diseases associated with the top BW-DMPs. Many of the BW-DMPs were associated with cancers and metabolic syndrome. Some of the genes were associated with neoplasms and nervous system diseases that occur during neonatal or childhood development.

**Table 5-8. Diseases associated with top BW-DMPs**

<b>Disease</b>	<b>Gene</b>	<b>Adj. P value</b>
Type 2 diabetes	<i>TNFRSF1B, APOC3, ARHGEF11, HSD11B1, MADD</i>	0.0231
Carcinoma	<i>SFRP1, MLH1, HEPACAM, BMI1, MTA1, CRTCI, HEPN1</i>	0.0231
Medulloblastoma	<i>SFRP1, CASP8, BMI1</i>	0.0231
Cancer or viral infections	<i>HEPACAM, MTA1, MYBL2, SFRP1, BCL2L1, MLH1, CASP8, IDH1, BMI1, CRTCI</i>	0.0231
Insulin resistance	<i>TNFRSF1B, APOC3, HSD11B1, MADD</i>	0.0293
Breast diseases	<i>SFRP1, CASP8, BCAS2, BMI1, MTA1</i>	0.0318

Adj. P value: P value adjusted by BH test

## 5.4 Discussion

The aim of this chapter was to identify the differential methylation sites related to birth weight in discordant identical twin pairs. The hypothesis was that there were longstanding prenatal-affected methylation changes that persist over time. While there was no birth weight related methylated regions at a genome-wide significance, there were still some meaningful associations from the top findings that link to metabolically related traits and other complex trait. These results suggested that intrauterine development might potentially impact the DNA methylation of certain regions and remain stable into later life, but larger samples are required to confirm this.

### 5.4.1 Evidence of neonatal and postnatal variations in DNA methylation

I did not detect BW-DMPs at genome-wide significance in the overall results from this chapter. This could be due to the small sample size, which was not powered enough to detect small effect sizes in this study. The methylation differences of older twins (Fraga *et al.*, 2005) and heritability studies (Z. A. Kaminsky *et al.*, 2009; J. T. Bell *et al.*, 2012; Gordon *et al.*, 2012) have indicated that MZ twins have very similar methylation patterns. This might be one of the reasons why extremely significant DMPs from the twin datasets were not found. Another could be that the observatory and replication datasets were small and so the current study was underpowered to detect modest effect sizes. Using the meta-analysis results from the categorical analysis as an example, I calculated the methylation difference between low birth weight MZ twin and their co-twins for the top 24 CpG sites. The methylation differences ranged from 0.6% to 5.8%. Based on the results from Chapter 2, if I took even the largest effect observed here (5.8%), given the samples sizes used in this study (N = 20, 25, 45 MZ twin pairs) the simulation-based power estimates were null (Chapter 2, Table 3). Given the small differences in methylation observed in this study more than 100 MZ twin pairs would be required to have reasonable power to achieve epigenome-wide statistical significance.

Variation in the neonatal methylome occurs in the intrauterine environment and during early development (Whitelaw *et al.*, 2010). Several imprinted genes, such as the *H19*

and *IGF2* were differentially methylated in newborns (Heijmans *et al.*, 2008; Hoyo *et al.*, 2012; Murphy *et al.*, 2012). Because these studies were undertaken in unrelated newborns, I compared my results from the 310 subjects to these findings. I observed association at nominal significance ( $P < 0.05$ ) at 17 (4.9% of total) BW-DMPs in the *IGF* family (that included 7 *IGF* genes) and 6 (10% of total) BW-DMPs in *H19*. A previous study identified 23 genes that explained 70 to 87% of the birth weight variance among which 6 genes (*ANGPT4*, *CPOE*, *CDK2*, *GRB10*, *OSBPL5*, and *REG1B*), associated with growth *in vivo/vitro* (Turan *et al.*, 2012). In my data, I found 2 DMPs on *CDK2*, 8 DMPs on *GRB10* (most significant with  $P = 2.8 \times 10^{-3}$ ), 2 DMP on *REG1B*, and 7 DMPs on *OSBPL5* (most significant with  $P = 2.3 \times 10^{-3}$ ) that were correlated to birth weight beyond nominal significance. Therefore, my findings are consistent with previous candidate gene methylation studies of birth weight.

I compared the BW-DMPs to other previous studies (Adkins *et al.*, 2012; Gordon *et al.*, 2012; Mulligan *et al.*, 2012). Several of the identified genes relate to early cell and embryonic development, growth, immune system, and inflammatory response, or are differentially methylated across tissues, such as cord blood and placenta. Many probes located in these genes were differentially methylated in adult blood samples. From the Dutch Famine study (Heijmans *et al.*, 2008; Tobi *et al.*, 2009), the methylation status of candidate genes for metabolic and cardiovascular diseases were examined and compared to non-exposed siblings. These results are consistent and suggest that regions differentially methylated with birth weight may persist and can act as a long lasting biomarker for metabolic syndrome.

#### **5.4.2 Replication of BW-DMPs in different twin datasets**

Most BW-DMP studies focus on finding DMPs that persistently change in all twins. Methylation differences were compared using a one-sample t-test or one-sample Wilcoxon test, and birth weight was used for categorizing low or high status in the twin pair. This is the first study to treat birth weight as a continuous trait in twins. The methylation differences in the less discordant twins might be modest, and the effects should be evident in an increased number of discordant twin pairs.

I compared my analysis with two published studies on birth weight discordant twins. All but one BW-DMP (cg02813863) that was reported in Gordon et al.'s study was also nominally significant in my analysis (Gordon *et al.*, 2012). One study (Souren *et al.*, 2013) was most similar to my study and also conducted a BW EWAS (Illumina 450k) in the adult methylome using 17 MZ birth weight discordant twin pairs. However, the authors used saliva samples to detect methylation levels and adjusted the methylation using a 'saliva specific' marker. In total 3,153 BW-DMPs were identified at  $P < 0.01$  between the heavy and light co-twins, and 45 BW-DMPs further showed moderate mean beta differences. Among these, 8 candidates were selected for deep bisulfite sequencing, but failed to validate. After comparing my categorical trait analysis to their 45 findings, one BW-DMP in particular, the *RUNX2* gene (cg22768222) was found at nominal significance ( $P = 0.02$  vs.  $P = 0.007$  in Souren et al.'s study). Closer inspection indicates that the locus was an a-DMP using my population-based samples and occurred in two tissues: blood ( $P = 10^{-6}$ ) and adipose tissue ( $P = 0.007$ ). In a re-evaluation of the 51 top probes from the meta-analysis, I found about 20% of BW-DMP were also a-DMPs at nominal  $P$  value ( $P < 0.05$ ). This suggests that some of the methylation changes related to birth weight might change with age, and hence the prenatal effects on these regions might be undermined over time. Therefore, these effects will be detected in methylation samples from newborns, but not in adults, as in the current study.

### **5.4.3 Disease-associated genes**

Out of the three types of analyses performed here (continuous trait analysis, categorical trait analysis, and population-based analysis) only 1 BW-DMP from the replication dataset in the continuous trait analysis reached genome-wide statistical significance within one of the datasets. The adult methylome changes during an individual's lifespan, so that the birth weight effect may be weakened overtime. It was interesting that disease enrichment analysis revealed that many of the top genes were enriched in birth weight associated diseases, for example, type 2 diabetes (Johansson *et al.*, 2008), metabolic diseases (Fagerberg *et al.*, 2004), stroke (Baker *et al.*, 2008), and multiple cancer types (Risnes *et al.*, 2011; A. H. Wu *et al.*, 2011). This suggests that quite a few BW-DMPs might have biological function in birth weight associated diseases. Further

work should focus on performing a similar analysis in an expanded sample of discordant MZ twins or population based dataset.

#### **5.4.4 Strengths and weaknesses of the current study**

The strengths of this study were using birth weight discordant MZ twin pairs, and EWAS of high coverage probes (Illumina 450k), and examination of birth weight as a continuous trait. Moreover, I have shown that some of the results also apply to the general population from which the twin pairs were derived having included a population-based dataset.

There were several limitations. Firstly, birth weight or weight in general is a complex phenotype, and exactly how much genetic and epigenetic variation associate with it, and the extent of its shared effects with many late-onset diseases remains unknown. The sample size was low and a valuable point would be to collect data on the same twins at different time points as in a longitudinal study. Also, we lack information on the chronicity of the twin pairs, and previous studies report that the MZ MC (monochorionic: twins share a single placenta) twins have more imbalanced nutrient supply than the MZ DC (dichorionic: twins share two separate placentas) twins (Derom *et al.*, 2006). A previous study (Z. A. Kaminsky *et al.*, 2009) suggests that some of the results may be diluted when I include more MZ DC twins.

In conclusion, using DNA methylation from birth weight discordant MZ twins in multiple datasets, I found some longstanding epigenetic markers that may associate with low birth weight. However, consistent replication at genome-wide significant thresholds was lacking. Severe intra-uterine growth differences might have caused the methylation changes and birth weight differences. Further studies with more samples and robust design are necessary to find the association between these markers and metabolic disease attributed to low birth weight.

# Tobacco Smoking Induces Coordinated DNA Methylation and Gene Expression Changes Across Multiple Tissues

---

Tobacco smoking is a major disease risk factor with well-known impacts on blood DNA methylation variation. Several studies have identified smoking-associated differential methylation regions, with replication and validation across populations. However, few studies identify the gene expression associated with smoking. It is also unclear whether smoking-induced DNA methylation changes are systemic effects, with functional impacts that can be influenced by underlying genetic variants. Here, I investigated the DNA methylation and gene expression profiles of 542 adipose tissues. In order to characterise the identified smoking effect in adipose tissues were tissue-specific or tissue-shared effect, I also performed the genome-wide scans for methylation and gene expression in two additional blood datasets.

---

## 6.1 Introduction

Smoking is a significant environmental risk factor that predisposes an individual to premature death and the development of chronic disease and several cancers (Ezzati & Lopez, 2003; Thun *et al.*, 2010). The effect of smoking is directed to the exposed regions of the lung. It also damages other organs of the body and causes DNA mutations linked to cancer (Pfeifer *et al.*, 2002). Cotinine is a widely used biomarker for smoking in serum or plasma and has a half-life of about 16 hours in the human body (Benowitz, 2008; Hannan *et al.*, 2009). It is a good clinical indicator, however, it is



largely limited to current smokers because cotinine levels in serum fall if a subject stops smoking for more than a day.

Compared to cotinine, persistent smoking might have longer lasting effects on DNA methylation. Cigarette smoking can change DNA methylation through various biological pathways. Firstly, more DNMTs are recruited (Mortusewicz *et al.*, 2005) and can methylate CpGs adjacent to the repaired nucleotides (Cuozzo *et al.*, 2007), which are mutated as a result of DNA damage from the carcinogens in cigarette smoke, e.g. arsenic, chromium, formaldehyde, polycyclic aromatic hydrocarbons, and nitrosamines (Smith and Hansch, 2000; Suter *et al.*, 2010). Secondly, nicotine from cigarette smoke has been shown to down-regulate DNMT1 mRNA and protein expression in mouse brain neurons (Satta *et al.*, 2008). Third, cigarette smoke increases Sp1 expression, a transcription factor that binds to GC-rich motifs in gene promoters (Kadonaga *et al.*, 1987) in lung epithelial cells (Mercer *et al.*, 2009; Di *et al.*, 2012). This might affect the methylation of CpGs during early embryogenesis (Han *et al.*, 2001). Lastly, hypoxia, which may result from competitive carbon monoxide (from cigarette smoke) binding to haemoglobin, might alter DNA methylation by HIF-1 $\alpha$ -dependent up-regulation of methionine adenosyltransferase 2A, an enzyme that synthesizes S-adenosylmethionine and thus donates methyl groups required for DNA methylation processes (Liu *et al.*, 2011).

Many EWAS studies have identified and replicated several smoking differentially methylated positions (smoking-DMPs) (Breitling *et al.*, 2011; Joubert *et al.*, 2012; Monick *et al.*, 2012; Wan *et al.*, 2012; Buro-Auriemma *et al.*, 2013; Philibert *et al.*, 2013; Shenker *et al.*, 2013; Sun *et al.*, 2013; Zeilinger *et al.*, 2013; Besingi & Johansson, 2014; Dogan *et al.*, 2014; H. R. Elliott *et al.*, 2014; Markunas *et al.*, 2014; Y. Zhang *et al.*, 2014). The smoking-induced methylation changes could occur in different tissues, ethnic groups, and throughout stages of development. For example, maternal smoking during pregnancy impacts methylation levels of newborns at genes such as *AHRR*, *CYP1A1* and *GFII* (Joubert *et al.*, 2012; M. A. Suter *et al.*, 2013; Markunas *et al.*, 2014). The majority of smoking-DMPs are reduced in current smokers compared to non-smokers. It is thought that hypo-methylated smoking-DMPs associate with the up-regulation of genes that are differentially expressed in smoking-related diseases.

The first smoking EWAS was performed using the Illumina 27k array in blood samples from 177 subjects (Breitling *et al.*, 2011), including 65 heavy smokers, 56 ex-smokers, and 56 non-smokers. They identified a single locus cg03636183 in the *F2RL3* gene as a highly significant smoking-DMP ( $P = 2.68 \times 10^{-31}$ ). In this region, methylation levels were lower in current smokers. The finding was replicated in 316 blood samples (95 smokers, 97 ex-smokers, 124 non-smokers) using mass spectrometry ( $P = 6.33 \times 10^{-34}$ ), and further replicated and validated in a number of follow-up studies (Breitling *et al.*, 2011; Joubert *et al.*, 2012; Wan *et al.*, 2012; Shenker *et al.*, 2013; Sun *et al.*, 2013; Zeilinger *et al.*, 2013; Besingi & Johansson, 2014; Dogan *et al.*, 2014; H. R. Elliott *et al.*, 2014; Harlid *et al.*, 2014; Y. Zhang *et al.*, 2014).

Subsequently, smoking EWAS have been increasingly performed on samples other than blood, such as placenta (M. Suter *et al.*, 2011) and airway epithelium (Selamat *et al.*, 2012; Buro-Auriemma *et al.*, 2013), and in individuals of different ethnicities (e.g. Africa American females (Dogan *et al.*, 2014; H. R. Elliott *et al.*, 2014)), and at different stages of development (e.g. newborns (Joubert *et al.*, 2012)), and on different platforms (e.g. Illumina GoldenGate (Siedlinski *et al.*, 2012), Illumina 27k (Breitling *et al.*, 2011; Selamat *et al.*, 2012), and Illumina 450k (Bibikova *et al.*, 2011; Dedeurwaerder *et al.*, 2011)). Successful replication has consistently identified a large number of loci (see Table 8 at the end of chapter) with highly reproducible effects at or near genes *AHRR*, *F2RL3*, *GFII*, and others. The largest number of smoking-DMPs found to date by a single study was 972 smoking-DMPs ( $P < 10^{-7}$ ) in 262 current smokers compared to 749 non-smokers using the Illumina 450k array in blood samples (Zeilinger *et al.*, 2013), where the effects at 187 CpG sites were replicated ( $P < 5 \times 10^{-5}$ ) in a further sample of 236 current smokers and 232 non-smokers, and at several sites methylation levels were also associated with smoking cessation in 782 former smokers.

Cessation of smoking could also lead to a change in methylation levels at particular genes, reaching methylation levels similar to those observed in non-smokers (Shenker *et al.*, 2013; Zeilinger *et al.*, 2013; Y. Zhang *et al.*, 2014). At multiple loci, ex-smokers show intermediate methylation levels, between those in non-smokers and in current-smokers. Methylation levels at multiple loci also positively correlate with the cumulative dose of smoking (years) and negatively associated with the time since smoking cessation (years) (Shenker *et al.*, 2013; Zeilinger *et al.*, 2013; Y. Zhang *et al.*,

2014). Two independent studies showed that the methylation levels at several smoking-DMPs gradually become more similar to those in non-smokers in the first 20 years after quitting smoking, and remain stable over time (Zeilinger *et al.*, 2013; Y. Zhang *et al.*, 2014). Because these results suggest that smoking induces long-term methylation changes, smoking-DMPs are good candidates for biomarkers of smoking.

Only a few studies have examined smoking effects on gene expression changes. Smoking leads to gene expression changes in multiple tissues, such as human airway epithelium (Woenckhaus *et al.*, 2006; Schembri *et al.*, 2009), lung tissue (McLemore *et al.*, 1990), and alveolar macrophages (Ito *et al.*, 2001). In a limited study, both the smoking induced methylation and gene expression changes were concurrently examined using the Illumina 27k array and GeneChip Human Exon 1.0 ST gene expression array, and 72 genes were both differentially methylated and differentially expressed at a significance level of  $P < 0.01$  (Philibert *et al.*, 2013). Among these 72 genes, 50 were negatively correlated. In another study, conducted in 39 subjects using the HELP (HpaII tiny fragment Enriched by Ligation-mediated PCR) assay for methylation and HG-U133 Plus 2.0 assay for gene expression, 35 out of 204 differentially methylated genes correlated with gene expression. Roughly half were inverse correlated (Buro-Auriemma *et al.*, 2013). However, the platforms used in these two studies had relatively low genome-wide coverage.

This chapter aims to identify the smoking-related genome-wide DNA methylation changes in twins using the Illumina 450k array. Most of these studies were conducted in blood sample, and only a few studies in other tissues. Here I performed the first smoking EWAS in the adipose tissue, which is considered to store the long-term exposure effect than that in other tissue. Additionally, previous evidence shows there is an association between adipose gene expression and metabolic diseases, which smoking is also a risk factor for the disease progression. The first aim was to identify smoking-DMPs in adipose tissue. The second aim was to explore whether smoking associates with both DNA methylation and gene expression changes. The third aim was to explore tissue-specific and tissue-shared effects of smoking by comparing adipose and blood samples.

## 6.2 Materials and methods

In this chapter, I analysed Illumina 450k methylation and RNA-seq expression datasets of two tissues (adipose and blood). All subjects are Caucasian females from the TwinsUK cohort, and ascertained to be free from severe disease when the samples were collected. The subjects were recruited as part of the MuTHER study (Multiple Tissue Human Expression Resource; <http://www.muther.ac.uk>) (Nica *et al.*, 2011; Grundberg *et al.*, 2012).

### 6.2.1 Datasets

#### 6.2.1.1 DNA methylation and RNA-seq datasets in adipose tissue

For adipose tissue, the same adipose biopsy was used to detect methylation and expression levels. In total, there are 349 subjects, including 32 MZ pairs, 49 DZ pairs, and 187 unrelated singletons. There were 186 non-smokers, 128 ex-smokers, and 35 current-smokers (Table 6-1). There were no significant differences in smoking status according to zygosity ( $\chi^2$  P value = 0.48).

All DNA methylation levels were profiled on the Illumina 450k array as previously described (Grundberg *et al.*, 2013). After removing probes with missing values and probes that mapped to multiple loci in the human genome (hg19) within 2 mismatches, a final 396,025 probes were used for analysis.

**Table 6-1. Zygosity and smoking status in the adipose dataset**

Zygosity	Non-smoker	Ex-smoker	Current-smoker	Total
MZ	28 (43.8%)	29 (45.3%)	7 (10.9%)	64
DZ	55 (56.1%)	35 (35.7%)	8 (8.2%)	98
Singleton	103 (55.1%)	64 (34.2%)	20 (10.7%)	187
Total	186 (53.30%)	128 (36.68%)	35 (10.03%)	349

#### 6.2.1.2 DNA methylation and RNA-seq datasets in whole blood samples

To explore the result from adipose tissue in a different cell sample, I also analysed the smoking with the methylation levels of 355 blood samples (Dataset 2, as described in

Chapter 3) (Tsaprouni *et al.*, 2014). Because white blood cell (WBC) counts have been previously reported to impact methylation levels, subjects without data available for the four WBC counts (eosinophils, lymphocytes, monocytes, and neutrophils) were excluded from the study. In total, 306 subjects (186 non-smokers, 94 ex-smokers, and 26 current-smokers) were selected for analysis, and 105 of them overlapped with the 349 subjects in adipose dataset. A total of 461,040 probes were included for analysis in the blood data.

The blood RNA-seq data were also obtained from the EuroBATS project. A smaller subset of 152 subjects (12 current-smokers, 55 ex-smokers, 85 non-smokers) from the 349 subjects was included with expression data in both adipose and whole blood to validate the adipose findings.

## **6.2.2 Phenotype collection**

During a subject's clinical visit, basic demographic information was collected, and other measurements such as height and weight were also measured onsite.

The cotinine levels were detected from serum available in a subset (N = 987) of the TwinsUK individuals. Cotinine and other metabolites in serum were detected by the non-targeted technology GC-MS (gas chromatography mass spectrometry) and LC-MS (liquid chromatography mass spectrometry) approach used by the Metabolon platform. The cotinine levels were cleaned and normalized by Idil Yet. Of the 349 subjects in the adipose dataset, 40 subjects had cotinine levels available. I further removed 11 subjects because their DNA extraction age for DNA methylation was at least 1 year away from the cotinine detection date. Of the remaining 29 subjects, 23 were current-smokers, 4 ex-smokers, and 2 non-smokers.

Preliminary analysis used cotinine levels to assess smoking status, however, there were many missing data and therefore the smoking phenotype was determined from a self-reported questionnaire. There was longitudinal self-reported data on the smoking status of each subject, since the twins regularly visit the clinic in our department. Current smokers were defined as those subjects who consistently smoked cigarettes (and have not stopped at any point) according to their longitudinal records up to the clinical visit closest to the DNA extraction date for DNA methylation. Ex-smokers were individuals

who have not smoked cigarettes for more than one year, and non-smokers were individuals that never smoked according to the longitudinal questionnaire records.

## **6.2.3 Statistical analyses**

### ***6.2.3.1 Quality Control for Illumina 450k and RNA-seq data***

Quality control was performed in adipose and blood methylation datasets following procedures as described in Chapter 3. The visual plots for detection of outliers and correlation tests between the PCs and potential covariates were performed.

For the adipose dataset, the covariates used were the same as those described by Grundberg *et al.* (Grundberg *et al.*, 2013), and included: DNA extraction age, BMI, plate, bisulfite conversion, bisulfite efficiency, family structure and zygosity. For the blood dataset, the covariates were: DNA extraction age, BMI, plate, position on the plate, WBC counts, family structure and zygosity.

The quality control and the identification of batch effect for expression datasets have been previously discussed (Brown *et al.*, 2014; Buil *et al.*, 2015). In brief, the sequenced paired-end reads (49 bp) were mapped to the human genome (GRCh37) by Burrows-Wheeler Aligner (BWA) software v0.5.9 (Li & Durbin, 2009), then genes were annotated as defined by protein coding in GENCODE v10 (Harrow *et al.*, 2012). Samples were excluded if they failed during library preparation or sequencing. Samples were only considered to have good quality if more than 10 million reads were sequenced and mapped to exons. The expression levels were presented as the RPKM values (reads per kilobase of transcript per million mapped reads) and was rank normal transformed prior to analysis. The genotype of each subject was used for identity check in case of accidental sample swapping.

### ***6.2.3.2 Smoking differential methylation sites analysis***

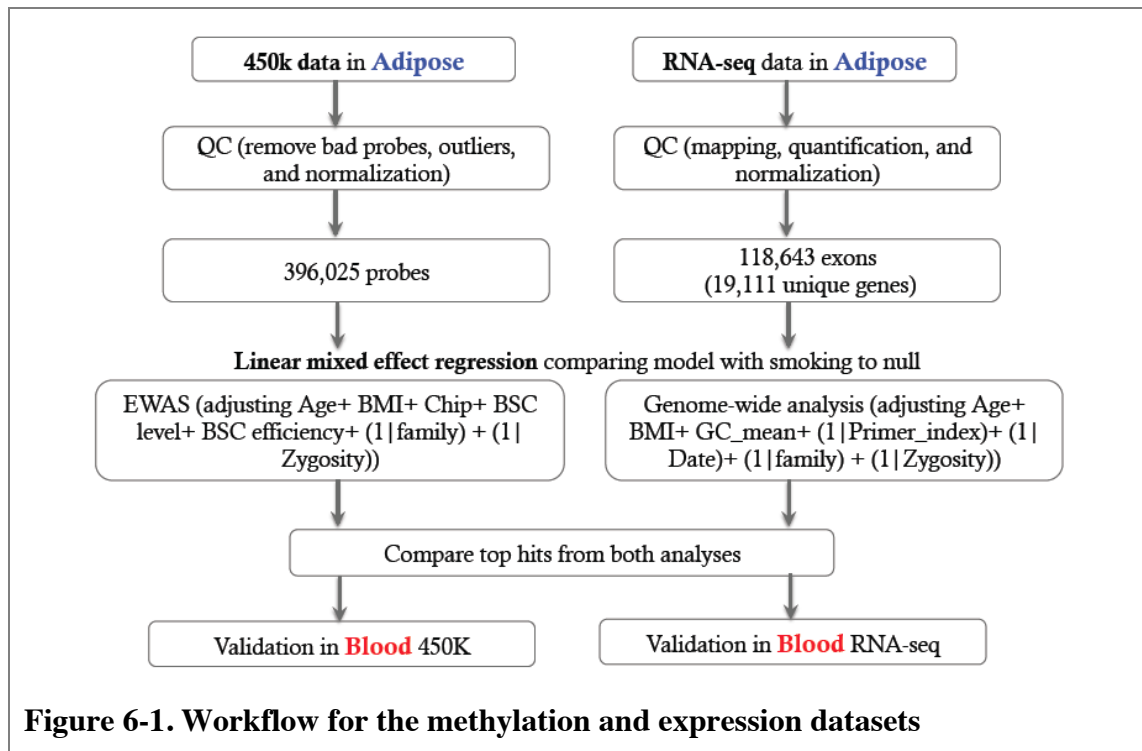
Figure 1 shows the overall schematic of the data analysis for this chapter. Raw DNA methylation betas were first normalized using BMIQ (Teschendorff *et al.*, 2013), and beta values on each probe were normalized to follow the normal distribution then fitted

to two linear mixed effect models (more details about quality control and normalization methods are described in Chapter 3 of this thesis). The linear mixed model adjusted for random effects (zygosity and family structure) and fixed effects (age, BMI, plate, position on the plate, bisulfite conversion, and bisulfite efficiency; the four blood cell counts were further adjusted as covariates in the blood dataset). The normalized betas were fitted with a full model as the outcome, and the predictors consisted of smoking status and all covariates. This full model was compared to the null model (without smoking status). A probe was defined as a smoking-DMP if it passed the false discovery rate of 5%. For smoking status, subjects were categorized into three groups (non-smoker, ex-smoker, current-smoker), and the phenotype was treated as a factor in the analysis. Therefore, a significant smoking-DMP in my study indicates that the methylation levels were different in at least two groups. To test which two groups were statistically significant, a post-hoc test using Tukey's method was performed.

### ***6.2.3.3 Smoking differentially expressed gene analyses***

For the RNA-seq datasets, the raw expression values were first rank normalized prior to analysis. A full model with expression levels as an outcome, and predictors that included smoking status and all the other covariates were compared to the null model without smoking. The genome-wide significance criterion was set at 5% FDR.

The workflow of quality control and analysis is shown on Figure 6-1



#### 6.2.3.4 Correlation between methylation and gene expression levels in adipose tissue

The top findings of the genome-wide scans in adipose tissue for smoking-methylation and smoking-expression were directly compared, as these results were both obtained from the same 349 subjects. The raw values of methylation and expression values were first normalized to follow normal distribution and residuals taken after adjusting for the covariates. Spearman’s correlation test was used to compare the correlation between the residuals of methylation and gene expression levels.

#### 6.2.3.5 Conditional analysis between methylation and gene expression levels in adipose tissue

In cases where the same gene showed both significant DNA methylation and exon expression changes associated with smoking, follow-up analyses were used to try to detect the regulatory pathway underlying the association findings. Three models were considered: (A) Smoking affects methylation, which modulates gene expression; (B) Smoking affects gene expression, which modulates methylation; and (C) Smoking affects methylation and gene expression independently.



To test which model best fits the potential regulatory pathway at the overlapping genes, a Bayesian information criterion (BIC) of the linear regression model was calculated for each of the corresponding models:

Model (A):  $\text{BIC}(\text{Expression} \sim \text{Methylation}) + \text{BIC}(\text{Methylation} \sim \text{smoking}) + \text{BIC}(\text{smoking} \sim 1)$

Model (B):  $\text{BIC}(\text{Methylation} \sim \text{Expression}) + \text{BIC}(\text{Expression} \sim \text{smoking}) + \text{BIC}(\text{smoking} \sim 1)$

Model (C):  $\text{BIC}(\text{Expression} \sim \text{Methylation}) + \text{BIC}(\text{Methylation} \sim \text{Expression}) + \text{BIC}(\text{smoking} \sim 1)$

On each gene, the sum of BIC values was first calculated in each model and then,  $\Delta\text{BIC}$ , the BIC difference of the lowest two BIC values for each comparison was also calculated. In the end, the model with a lower BIC value was the preferred model, and  $\Delta\text{BIC}$  was considered to be a measure of support for the preferred model. The strength of the evidence could be explained, following Kass and Raftery: if  $\Delta\text{BIC}$  was between 0 and 2, it was generally accepted that the support for the preferred model was not very different than the second preferred model. A model was considered to show much more evidence for support compared to the other tested models if it had  $\Delta\text{BIC}$  greater than 6 (Kass & Raftery, 1995).

#### ***6.2.3.6 Methylation quantitative trait locus (meQTL) and expression quantitative trait locus (eQTL)***

There is evidence that methylation and expression are heritable and their levels in certain regions are influenced by the genetic structure (J. T. Bell *et al.*, 2011). Thus, to exclude the possibility that smoking differentially methylated or expressed regions were driven by genetic contribution, genome-wide association scans (GWAS) were performed for the top methylated and expressed regions associated with smoking in both adipose tissue and blood samples. The GWAS results for methylation are referred to as quantitative trait loci (meQTL), and for expression are referred to as expression quantitative trait loci (eQTL). The genome-wide significance level was set at the Bonferroni adjusted P value of  $2.65 \times 10^{-8}$  for both meQTL and eQTL (1,880,781 SNPs).

## 6.3 Results

### 6.3.1 Smoking differentially methylated positions in adipose tissue (smoking-DMPs)

A smoking EWAS was performed in adipose tissue samples from 349 subjects to identify smoking-DMPs, using 396,025 probes. In total, there were 39 CpG sites at 23 known genes and 2 inter-genic regions that passed the 5% false discovery rate ( $P < 4.7 \times 10^{-6}$ ). Among the 39 CpG sites, DNA methylation levels of current smokers were lower than those in non-smokers. Figure 6-2 shows 32 hypo-methylated smoking-DMPs (blue) and 7 hyper-methylated smoking-DMPs (red) were found in adipose tissue. The names listed next to smoking-DMPs were CpG sites identified as smoking-DMPs from previous studies.

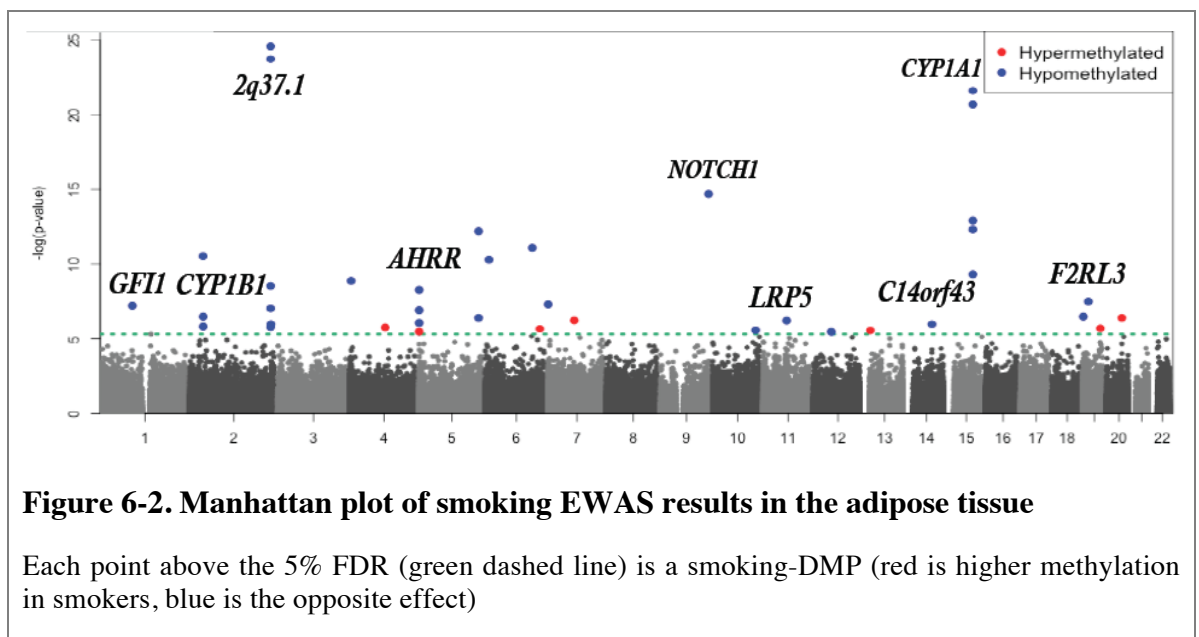


Table 6-2 lists the top 39 smoking-DMPs. The majority of post-hoc comparisons between the current smokers and non-smokers were significant, and half of the comparisons between current smokers and ex-smokers were significant as well. In general, the methylation levels among non-smokers and ex-smokers were similar, and the methylation levels of ex-smokers tended to fall in between the non-smokers and smokers.

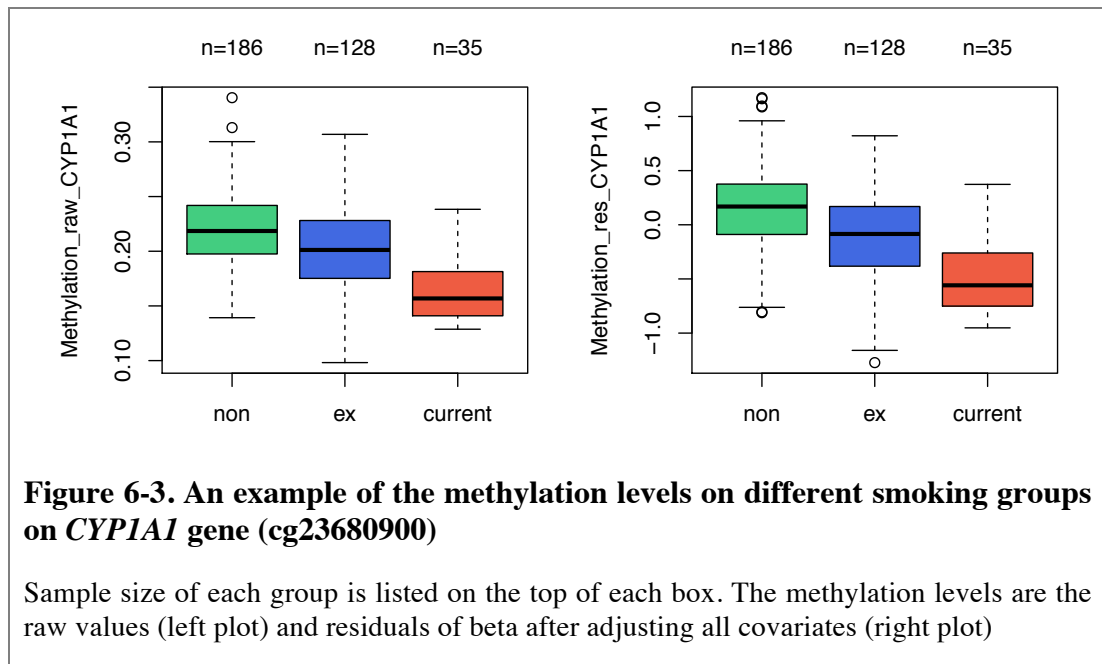
At these 23 known genes and 1 inter-genic region, 9 genes and 1 region have been previously reported as smoking-DMPs, including *GFII*, 2q37.1 region, *AHRR*, *NOTCH1*, *LRP5*, *C14orf43*, *CYP1A1*, *LINGO3*, and *F2RL3* (see Discussion section, Table 7), and 13 genes were novel findings (*CYP1B1*, *CYTL1*, *NEURL1B*, *GPER*, *SAG*, *LATS2*, *PMS2L11*, *NEDD9*, *PDE7B*, *SLC39A8*, *CDC42EP3*, *HTRA1*, and *ACVRL1*). The 9 previously reported smoking-DMPs were mostly identified in blood samples previously, and the direction of association between smoking status and methylation changes in the current study was consistent with these previous studies.

**Table 6-2. Top 39 smoking-DMPs in adipose tissue**

IlmnID	CHR	Gene Name	Location	E vs. N	S vs. E	S vs. N	P value
cg05951221	2	chr2:233284402	-	-0.528	-1.048	-1.576	2.51E-25
cg21566642	2	chr2:233284661	-	-0.397	-1.210	-1.606	1.81E-24
cg23680900	15	<i>CYP1A1</i>	TSS200	-0.564	-0.817	-1.381	2.37E-22
cg26516004	15	<i>CYP1A1</i>	TSS1500	-0.761	-0.545	-1.306	1.95E-21
cg14120703	9	<i>NOTCH1</i>	Body	-0.160	-1.106	-1.266	1.96E-15
cg23160522	15	<i>CYP1A1</i>	5'UTR	-0.289	-0.945	-1.234	1.19E-13
cg10009577	15	<i>CYP1A1</i>	TSS1500	-0.345	-0.458	-0.803	4.73E-13
cg22418620	5	<i>NEURL1B</i>	Body	-0.160	-1.063	-1.224	6.10E-13
cg01985595	6	<i>PDE7B</i>	Body	-0.205	-0.956	-1.161	7.81E-12
cg07992500	2	<i>CDC42EP3</i>	5'UTR	-0.259	-0.892	-1.152	2.90E-11
cg12531611	6	<i>NEDD9</i>	Body	-0.245	-0.754	-0.999	4.77E-11
cg00353139	15	<i>CYP1A1</i>	TSS200	-0.389	-0.571	-0.960	4.69E-10
cg00512031	4	<i>CYTL1</i>	TSS1500	-0.120	-0.786	-0.906	1.26E-09
cg06644428	2	chr2:233284112	-	-0.489	-0.341	-0.830	2.73E-09
cg19405895	5	<i>AHRR</i>	Body	0.016	-0.876	-0.860	5.14E-09
cg03636183	19	<i>F2RL3</i>	Body	-0.152	-0.677	-0.829	3.11E-08
cg11461808	7	<i>C7orf50;GPER</i>	1stExon;5'UTR;Body;TSS1500	-0.470	-0.155	-0.625	4.96E-08
cg14179389	1	<i>GFII</i>	Body	-0.236	-0.634	-0.869	5.93E-08
cg01940273	2	chr2:233284934	-	-0.065	-0.707	-0.773	9.15E-08
cg25648203	5	<i>AHRR</i>	Body	-0.174	-0.756	-0.930	1.21E-07
cg02162897	2	<i>CYP1B1</i>	Body	-0.186	-0.688	-0.873	3.15E-07
cg00378510	19	<i>LINGO3</i>	Body	0.114	-0.902	-0.788	3.15E-07
cg03646542	5	<i>NEURL1B</i>	Body	-0.183	-0.686	-0.868	3.77E-07
cg11841529	20	<i>CD40</i>	TSS200	-0.380	0.879	0.498	3.78E-07
cg01727317	7	<i>PMS2L11</i>	Body	0.303	0.239	0.542	5.75E-07
cg21611682	11	<i>LRP5</i>	Body	-0.006	-0.798	-0.804	5.94E-07
cg05575921	5	<i>AHRR</i>	Body	-0.039	-0.664	-0.703	8.16E-07
cg04341454	2	<i>SAG</i>	3'UTR	0.052	-0.618	-0.567	9.92E-07
cg22851561	14	<i>C14orf43</i>	5'UTR	0.003	-0.806	-0.802	1.01E-06
cg20408276	2	<i>CYP1B1</i>	Body	-0.202	-0.638	-0.840	1.47E-06
cg03329539	2	chr2:233283329	-	-0.179	-0.710	-0.889	1.59E-06
cg13735704	4	<i>SLC39A8</i>	Body	0.162	0.539	0.701	1.68E-06
cg05242523	19	<i>KLK14</i>	TSS1500	-0.287	0.841	0.554	1.92E-06
cg25767832	6	chr6:158210955	-	0.383	-0.002	0.381	2.08E-06
cg08447739	10	<i>HTRA1</i>	TSS1500	0.096	-0.849	-0.753	2.61E-06
cg25576788	13	<i>LATS2</i>	5'UTR	-0.291	0.665	0.374	2.68E-06
cg24980413	5	<i>AHRR</i>	Body	0.126	0.732	0.858	3.18E-06
cg20131897	12	<i>ACVRL1</i>	TSS1500;5'UTR	-0.180	-0.597	-0.776	3.28E-06
cg04135110	5	<i>AHRR</i>	Body	0.039	0.788	0.827	4.07E-06

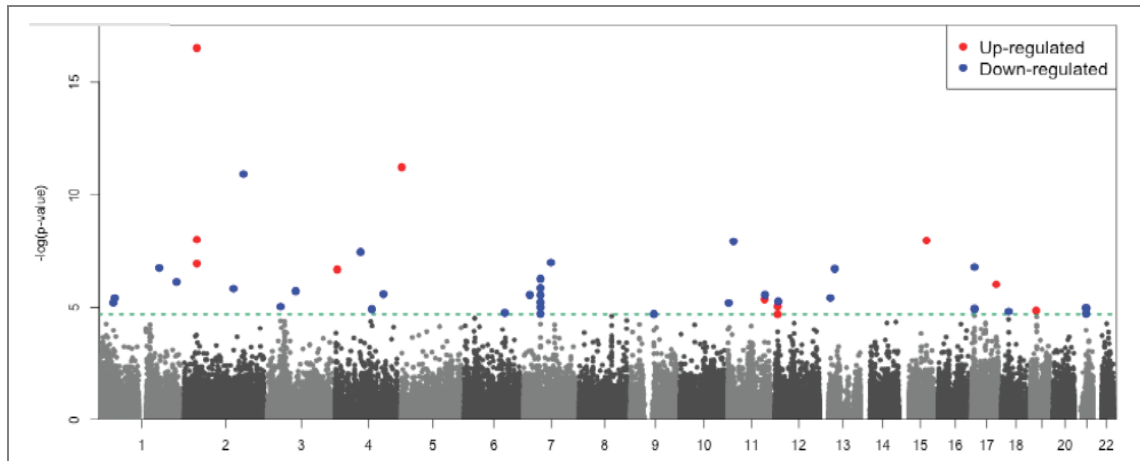
Abbrev: IlmnID, Illumina probe ID; CHR, chromosome; Location: probe location to gene; E vs. N: beta by comparing ex-smoker to non-smoker group; S vs. E: beta by comparing current smoker to ex-smoker group; S vs. N, beta by comparing current smoker to non-smoker group; P value, the global P value for smoking. A negative beta value indicates the methylation in the former group is less than the latter group.

Figure 6-3 shows an example of the methylation levels between the different smoking groups at the *CYP1A1* gene, and the methylation levels were significantly different in all three groups on this CpG site (Table 6-2): lowest methylation levels are found in current smokers, and non-smokers have the highest methylation levels. The ex-smokers have intermediate levels of non-smokers and current smokers.



### 6.3.2 Smoking differentially expressed regions in adipose tissue

The genome-wide scans comparing smoking and gene expression exon estimates were performed in adipose tissue samples from 349 subjects at 118,643 exons. Altogether, 48 exons at 35 unique genes were differentially expressed with smoking at 5% FDR ( $P < 2.05 \times 10^{-5}$ ). Figure 6-4 shows the Manhattan plot of these results. Most of the associated exons were down regulated in smokers (blue) and fewer exons were up regulated in smokers (red).



**Figure 6-4. Manhattan plot of the smoking associated exon expression in the adipose tissue**

Each point above the 5% FDR (green dash) represents a significantly expressed exon, where expression can be higher (red) or lower (blued) in smokers.

The full list of the 48 differentially expressed exons is listed in Table 6-3. Among these 35 genes, there were 3 previously known smoking-expressed genes, *CYP1B1*, *AHRR*, and *CDKN1C*, and 32 novel genes (*ZNF385B*, *EDC3*, *KCNJ11*, *PARM1*, *SEM13E*, *NMNAT2*, *MYH10*, *SMAD9*, *CYTL1*, *CYGB*, *COBL*, *LYST*, *NCAM1*, *F2RL3*, *KIF5C*, *CADM2*, *HHIP*, *PDZD4*, *HDAC9*, *TNFRSF19*, *PZP*, *DLAT*, *ERMAP*, *CD163L1*, *ENTPD3*, *JAM2*, *PGM5*, *TRDN*, *RBMXP4*, *LAMA3*, *CD163L1*, and *TTC9*). Some of these genes were also found as smoking-DMPs in previous studies, e.g. *AHRR*, *CYP1B1*, *CYTL1*, and *F2RL3*. At the 48 differentially expressed exons, expression levels in ex-smokers were more similar to those in non-smokers, and there was no significant difference between these two groups (minimum  $P = 2.7 \times 10^{-4}$ ).

**Table 6-3. Top 48 smoking-expressed exons in adipose tissue**

Exon	CHR	Gene name	E vs. N	S vs. E	S vs. N	P value
ENSG00000138061.7_38294652_38298453	2	<i>CYP1B1</i> <sup>1,2</sup>	-0.003	1.442	1.439	4.25E-18
ENSG00000063438.12_433968_438406	5	<i>AHRR</i> <sup>1,2</sup>	0.198	1.074	1.272	7.92E-12
ENSG00000144331.14_180306709_180308195	2	<i>ZNF385B</i> <sup>1</sup>	-0.279	-0.946	-1.225	1.96E-11
ENSG00000138061.7_38302919_38303323	2	<i>CYP1B1</i> <sup>1,2</sup>	0.048	0.975	1.023	1.01E-08
ENSG00000179151.6_74922899_74925287	15	<i>EDC3</i> <sup>2</sup>	-0.006	0.928	0.922	1.09E-08
ENSG00000187486.5_17407406_17410206	11	<i>KCNJ11</i>	-0.228	-0.845	-1.073	1.18E-08
ENSG00000169116.7_75971373_75975325	4	<i>PARM1</i>	0.002	-0.932	-0.930	3.56E-08
ENSG00000170381.7_82993222_82997354	7	<i>SEMA3E</i>	-0.181	-0.758	-0.939	7.11E-08
ENSG00000138061.7_38301489_38302532	2	<i>CYP1B1</i> <sup>1,2</sup>	-0.098	0.988	0.890	9.51E-08
ENSG00000157064.6_183217372_183221878	1	<i>NMNAT2</i>	-0.088	-0.817	-0.906	1.53E-07
ENSG00000133026.7_8377523_8379266	17	<i>MYH10</i>	-0.245	-0.729	-0.974	1.69E-07
ENSG00000120693.9_37418968_37422956	13	<i>SMAD9</i>	-0.088	-0.849	-0.937	1.82E-07
ENSG00000170891.6_5016313_5016961	4	<i>CYTL1</i> <sup>2</sup>	0.263	0.698	0.960	2.51E-07
ENSG00000161544.4_74523440_74524693	17	<i>CYGB</i>	0.189	0.694	0.883	4.22E-07
ENSG00000106078.12_51152863_51153001	7	<i>COBL</i>	-0.398	-0.427	-0.825	5.24E-07
ENSG00000143669.8_235824341_235826378	1	<i>LYST</i>	-0.028	-0.810	-0.838	7.36E-07
ENSG00000106078.12_51095409_51097288	7	<i>COBL</i>	-0.285	-0.654	-0.939	9.54E-07
ENSG00000149294.11_113145989_113149158	11	<i>NCAMI</i>	-0.246	-0.639	-0.886	1.51E-06
ENSG00000127533.2_17000384_17002830	19	<i>F2RL3</i> <sup>2</sup>	-0.014	0.891	0.878	1.72E-06
ENSG00000168280.11_149879592_149883273	2	<i>KIF5C</i>	-0.183	-0.759	-0.942	1.81E-06
ENSG00000175161.8_86115815_86117944	3	<i>CADM2</i>	-0.305	-0.465	-0.770	1.97E-06
ENSG00000164161.5_145658916_145666423	4	<i>HHIP</i>	0.240	-0.935	-0.695	2.47E-06
ENSG00000106078.12_51111080_51111389	7	<i>COBL</i>	-0.287	-0.591	-0.878	2.75E-06
ENSG00000067840.6_153067621_153070355	X	<i>PDZD4</i>	-0.015	-0.791	-0.806	2.79E-06
ENSG00000048052.14_18705836_18708466	7	<i>HDAC9</i>	-0.241	-0.629	-0.870	2.88E-06
ENSG00000127863.11_24247511_24250232	13	<i>TNFRSF19</i>	0.080	-0.900	-0.820	2.93E-06
ENSG00000106078.12_51092806_51093069	7	<i>COBL</i>	-0.196	-0.681	-0.877	3.27E-06
ENSG00000126838.5_9305721_9305939	12	<i>PZP</i>	0.035	-0.871	-0.836	4.41E-06
ENSG00000150768.10_111933130_111935114	11	<i>DLAT</i>	0.272	0.624	0.896	4.51E-06
ENSG00000162367.6_47681963_47685846	1	<i>TALI</i>	-0.056	-0.763	-0.819	4.77E-06
ENSG00000164010.9_43308188_43310660	1	<i>ERM1P</i>	-0.157	-0.754	-0.911	4.98E-06
ENSG00000106078.12_51261076_51261286	7	<i>COBL</i>	-0.346	-0.448	-0.794	5.05E-06
ENSG00000129757.8_2905229_2907111	11	<i>CDKN1C</i> <sup>1</sup>	-0.337	-0.293	-0.629	6.04E-06
ENSG00000177675.4_7525916_7526236	12	<i>CD163L1</i>	-0.193	0.937	0.744	6.41E-06
ENSG00000106078.12_51083909_51085265	7	<i>COBL</i>	-0.223	-0.655	-0.878	7.00E-06
ENSG00000168032.4_40464341_40464613	3	<i>ENTPD3</i>	-0.189	-0.639	-0.828	7.35E-06
ENSG00000133026.7_8383428_8383636	17	<i>MYH10</i>	-0.205	-0.648	-0.853	7.48E-06
ENSG00000154721.9_27066068_27066220	21	<i>JAM2</i>	0.134	-0.797	-0.662	9.42E-06
ENSG00000154330.6_71098781_71098964	9	<i>PGM5</i>	-0.184	-0.704	-0.888	1.01E-05
ENSG00000186439.7_123537483_123539885	6	<i>TRDN</i>	-0.221	-0.651	-0.872	1.03E-05
ENSG00000154721.9_27011584_27012200	21	<i>JAM2</i>	-0.102	-0.738	-0.839	1.09E-05
ENSG00000133026.7_8387456_8387584	17	<i>MYH10</i>	-0.155	-0.687	-0.843	1.10E-05
ENSG00000249465.1_110267482_110268615	4	<i>RBMXP4</i>	0.015	-0.832	-0.817	1.22E-05
ENSG00000154721.9_27086952_27089874	21	<i>JAM2</i>	0.076	-0.805	-0.729	1.35E-05
ENSG00000154721.9_27070989_27071191	21	<i>JAM2</i>	0.014	-0.776	-0.763	1.53E-05
ENSG00000053747.9_21519187_21519350	18	<i>LAMA3</i>	0.147	-0.844	-0.697	1.54E-05
ENSG00000177675.4_7527877_7528191	12	<i>CD163L1</i>	0.037	0.792	0.829	1.67E-05
ENSG00000133985.2_71137793_71142077	14	<i>TTC9</i>	0.202	-0.865	-0.663	1.92E-05

Exon, exon location on the gene; CHR, chromosome; Gene name, UCSC gene name; Beta ex-smoker vs. non-smoker, beta by comparing ex-smoker to non-smoker group; Beta smoker vs. ex-smoker, beta by comparing current smoker to ex-smoker group; Beta smoker vs. non-smoker, beta by comparing current smoker to non-smoker group; Overall P value, the global significance P value for smoking

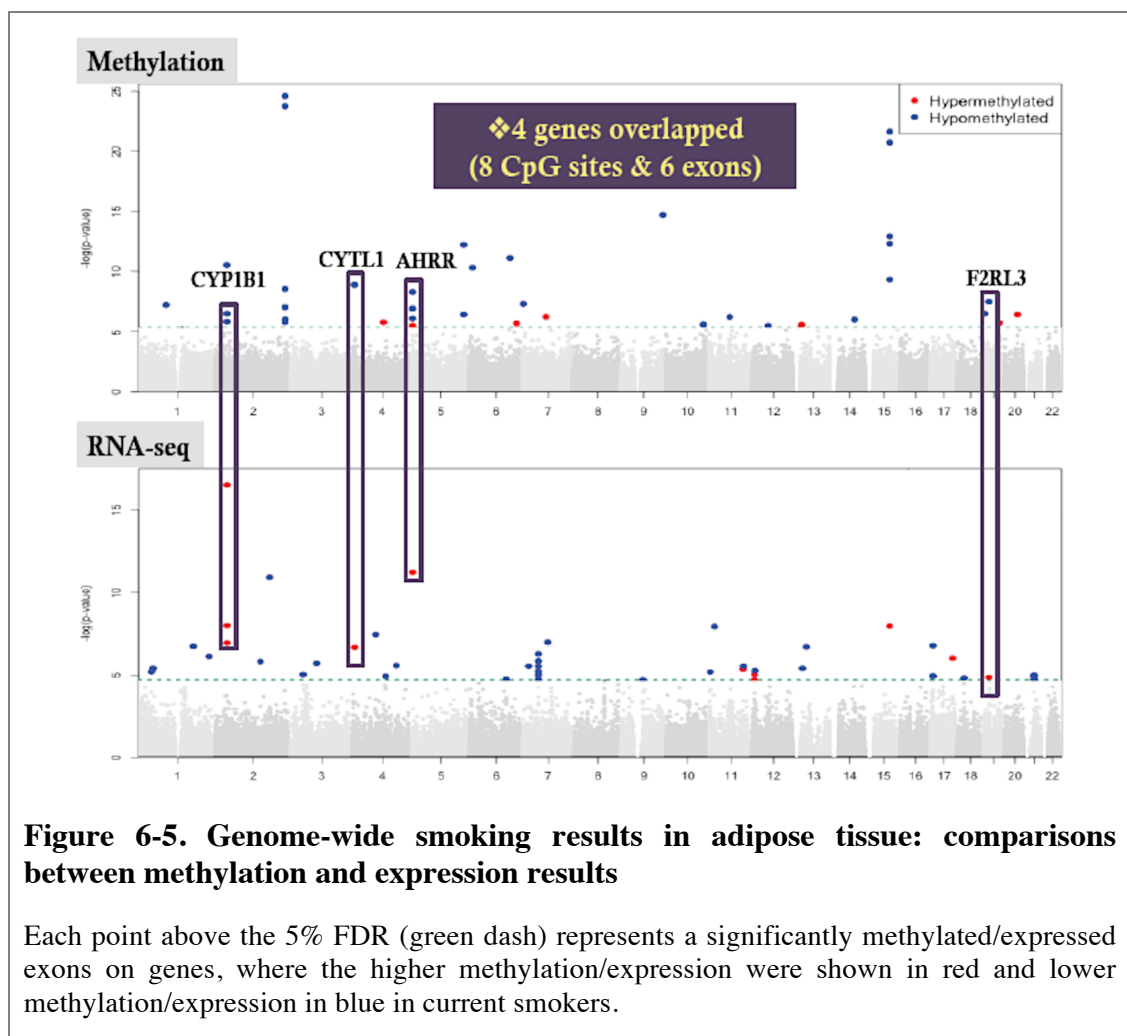
<sup>1</sup>These genes were previously identified to be differentially expressed with smoking: *CYP1B1* (Chang *et al.*, 2003; van Leeuwen *et al.*, 2007), *AHRR* (Monick *et al.*, 2012), and *CDKN1C* (Harvey *et al.*, 2007).

<sup>2</sup>These genes were identified as smoking-DMPs previously

### 6.3.3 Comparison between the smoking EWAS and genome-wide expression results

#### 6.3.3.1 Four genes overlap between methylation and expression adipose results

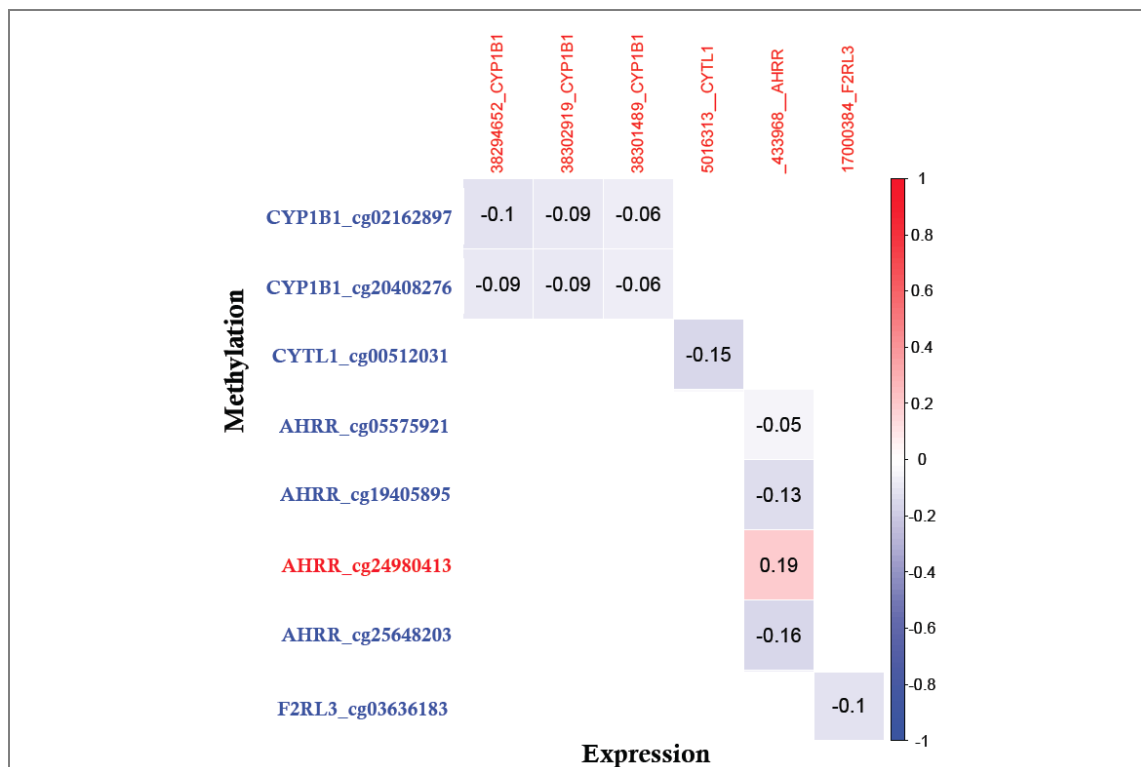
A comparison of the 39 smoking-DMPs and 48 differentially expressed exons showed that 4 genes (*CYP1B1*, *CYTL1*, *AHRR*, and *F2RL3* genes) overlapped between the smoking genome-wide methylation and exon expression analyses (Figure 6-5). There were 8 CpG sites on the 4 genes, including 2, 4, 1 and 1 CpG sites in the gene bodies of *CYP1B1*, *AHRR*, and TSS1500 regions of *CYTL1* and *F2RL3*, respectively. For the 6 expressed exons, 3 of them were in *CYP1B1*, and 1 for each of the remaining 3 genes. Apart from 1 CpG site on *AHRR*, current smokers showed lower methylation levels compared to non- and ex-smokers, and all the exons were up-regulated in current smokers.



The next analysis focused on testing whether there was a correlation between the methylation and gene expression levels at these 4 genes, with the hypothesis that DNA methylation (particularly in the promoter region) could down regulate gene expression.

### 6.3.3.2 Correlations between methylation and expression of the 4 overlapping genes

Figure 6-6 shows the correlation matrix between the methylation and expression of the 4 genes. Here, the predominant trend appears to be a weakly negative correlation between methylation and expression.

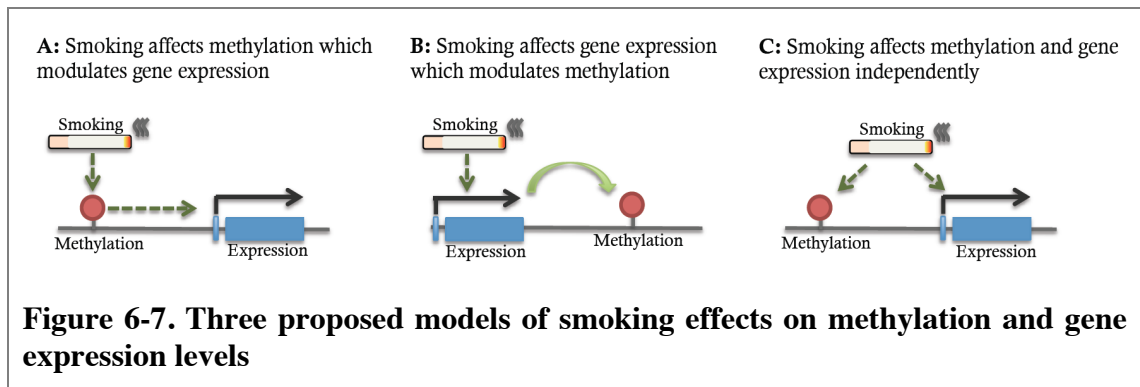


**Figure 6-6. Correlation matrix between methylation and expression of the 4 genes in adipose tissue**

Each column is a single exon and each row is a single CpG site, the colour shows whether the exon/CpG site is up regulated/hyper-methylated (red) or down regulated/hypo-methylated (blue). The number and corresponding colour scale show the correlation coefficients from -1 to 1 (blue to red).

To explore the underlying mechanisms of smoking-associations with both methylation and expression at these 4 genes, a conditional analysis was performed. Figure 6-7 shows the 3 proposed regulation models.





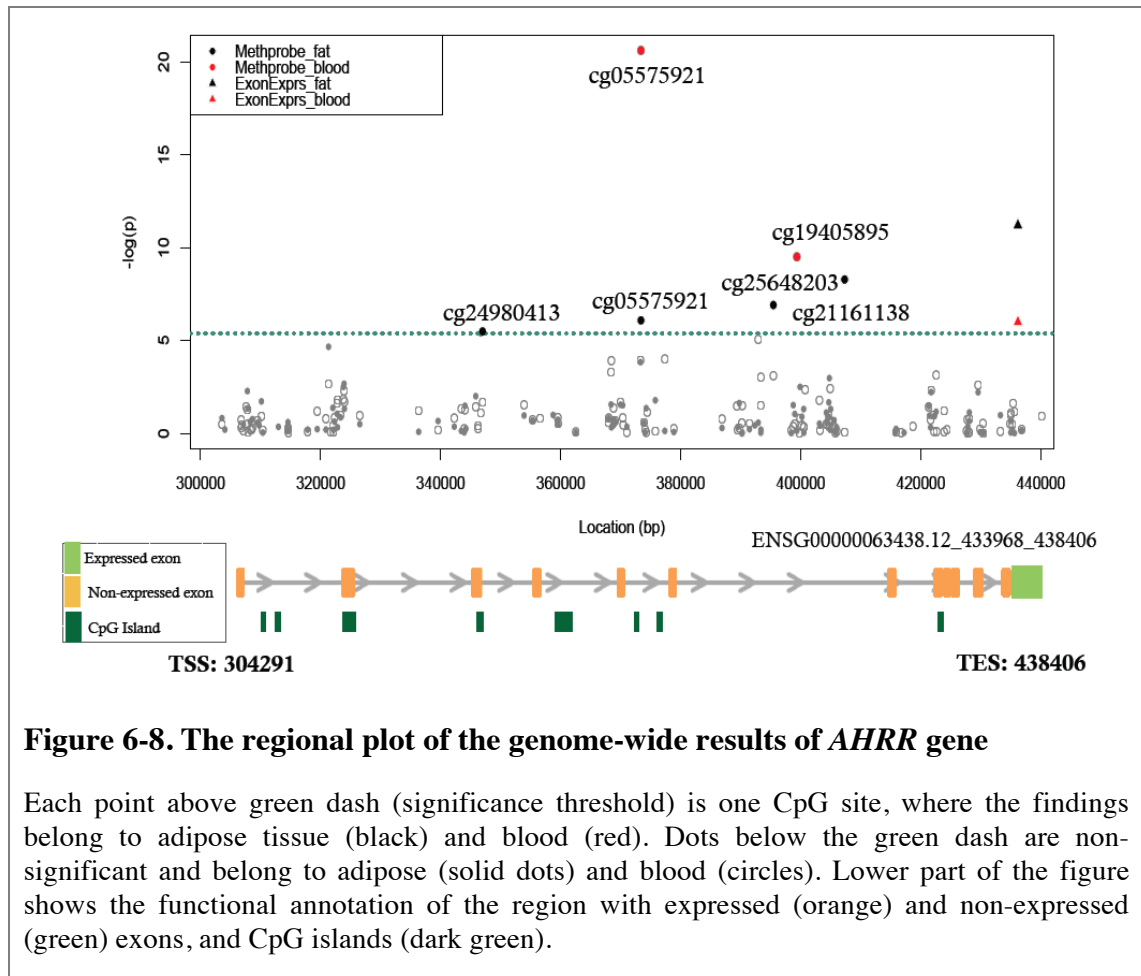
Among the 12 comparisons (Table 6-4), model (B) was the preferred model in 7 comparisons, model (A) was the preferred model in the remaining 5 comparisons, and none of the results supported a model of independent effects of smoking on methylation and expression (model (C)). If we only consider models where the fit was ‘significantly’ consistent with the preferred model (that is,  $\Delta\text{BIC} > 6$ ), there are 8 comparisons in three of the loci (*CYP1B1*, *AHRR*, and *F2RL3*) that meet these criteria. At locus *F2RL3* the result is consistent with the scenario where smoking affects DNA methylation, which regulates gene expression. On the other, at locus *AHRR* the results are consistent with the scenario where smoking affects gene expression, which has a modulatory effect on DNA methylation. However, at locus *CYP1B1* there were multiple DNA methylation probes and exons that were associated with smoking, and different combinations of these show support for either models (A) or (B), suggesting the possibility of more complex regulatory processes at this region and potential alternative transcript regulation.

**Table 6-4. Results of conditional analysis**

IlmnID	Exon start	Exon end	CHR	Gene	$\Delta\text{BIC}$	BIC (A)	BIC (B)	BIC (C)
cg02162897	38294652	38298453	2	CYP1B1	5.96	1485.7	1479.8	1502.8
cg02162897	38302919	38303323	2	CYP1B1	2.85	1964.8	1967.7	1982.7
cg02162897	38301489	38302532	2	CYP1B1	<b>11.61</b>	<b>1900.9</b>	1912.5	1920.2
cg20408276	38294652	38298453	2	CYP1B1	<b>7.67</b>	1446.7	<b>1439.1</b>	1463.0
cg20408276	38302919	38303323	2	CYP1B1	1.14	1925.2	1926.3	1941.6
cg20408276	38301489	38302532	2	CYP1B1	<b>9.90</b>	<b>1860.9</b>	1870.8	1878.5
cg00512031	5016313	5016961	4	CYTL1	0.06	1782.6	1782.6	1799.0
cg05575921	433968	438406	5	AHRR	<b>21.54</b>	1548.5	<b>1527.0</b>	1559.1
cg19405895	433968	438406	5	AHRR	<b>18.95</b>	1950.3	<b>1931.4</b>	1958.9
cg24980413	433968	438406	5	AHRR	<b>19.99</b>	1662.5	<b>1642.5</b>	1662.4
cg25648203	433968	438406	5	AHRR	<b>15.00</b>	1659.4	<b>1644.4</b>	1668.8
cg03636183	17000384	17002830	19	F2RL3	<b>10.57</b>	<b>1499.8</b>	1510.3	1515.9

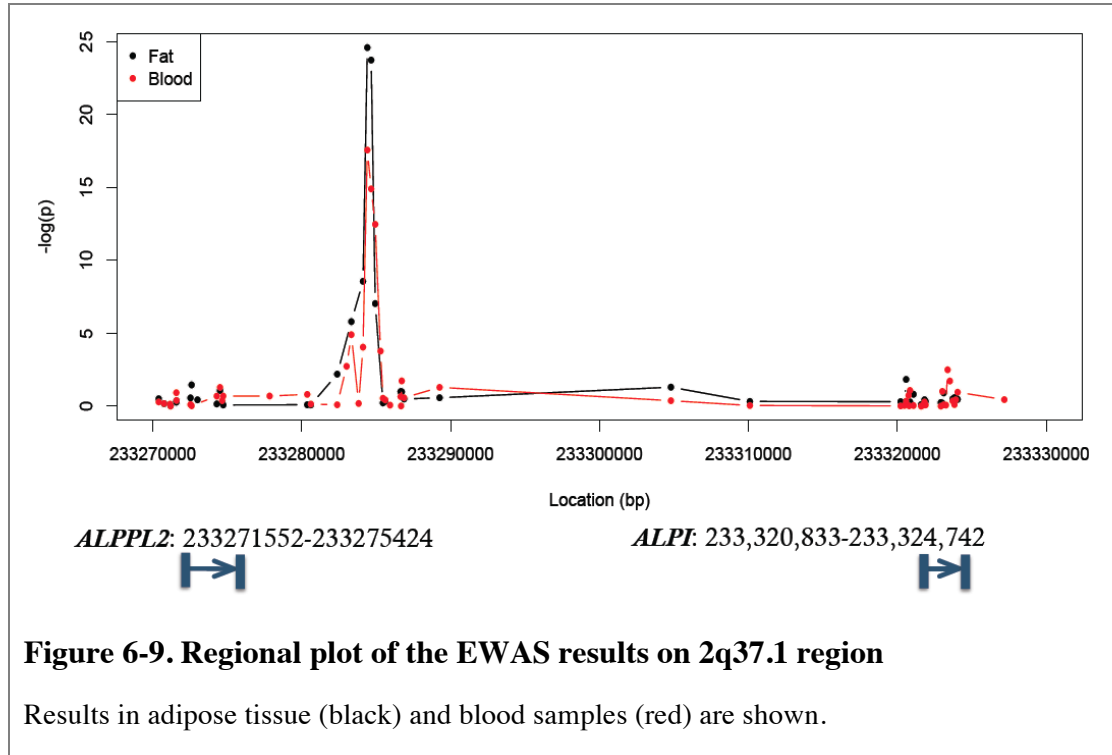
IlmnID, Illumina probe ID; BIC (A), smoking affects methylation, which modulates gene expression; BIC (B), smoking affects gene expression, which modulates methylation; and BIC (C), smoking affects methylation and gene expression independently. Numbers in bold indicated the lower level between the 3 models, which is the best model.

In the gene expression adipose-blood comparison, the only region that showed tissue-shared results was an exon of the *AHRR* gene. A visualization of this region with the DNA methylation results is shown in Figure 6-8. If we consider all available CpG sites of the *AHRR* gene (Figure 10, TSS: chr5:304,291 to TES: chr5:438,406), we observe that the expressed exon is the last exon of the *AHRR* gene (green block on the lower region of the figure). The significant smoking-DMP CpG-sites in *AHRR* are not located on CpG islands, nor in the exons, they all fall within introns.



At another of the 4 overlapping regions, the inter-genic region at 2q37.1, several smoking-DMPs were found in both tissues. These significant CpG sites were not located in any genes, and were 10 kb away from the *ALPPL2* gene and 35 kb away from the *ALPI* gene. Figure 11 is a visualization of these results and shows that the smoking-methylation results between blood and adipose were very similar across this genomic region. There was a strong peak of association at chr2:233,284,402 and chr2:233,284,661. Look ups into the chromatin signature of these regions using the

ENCODE data in the UCSC Genome Browser (Meyer *et al.*, 2013), shows the presence of H3K27m3, but not H3K27ac histone mark in GM12878 cells, suggesting there might be absence for transcription factor binding on the region.



### 6.3.4 Genetic contributions to methylation and exon expression levels

To investigate if there were genetic contributions to methylation and expression levels at the smoking-associated loci, GWAS was performed on the residuals of the methylation and expression levels at the top adipose and blood results. Due to missing genotype data in a few subjects, the GWAS was performed on a reduced sample of 250 individuals. Table 6-5 shows that at a significance level of  $P < 10^{-7}$  there were no eQTLs, but several SNPs were significantly associated with 3 methylation probes. These three probes were smoking-DMPs in adipose tissue (smoking-DMP minimum  $P = 3.5 \times 10^{-7}$ ). At the *LINGO3* and *HTRA1* gene, all the SNPs that associated with the methylation sites were within 25 kb suggesting that these two sites were cis-meQTLs.

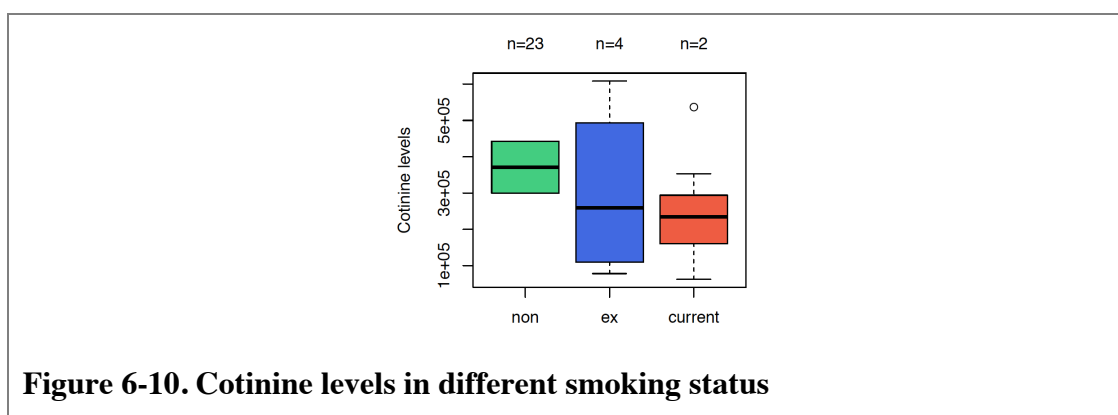
**Table 6-5. List of meQTLs found in the top EWAS results**

IlmnID	SNP	Meth	Meth gene	SNP	RS gene name	RS left gene	RS right gene	P value
cg00378510	rs757322	19	LINGO3	19	NA	SPPL2B	TMPRSS9	7.32E-12
cg00378510	rs11671	19	LINGO3	19	SPPL2B	LSM7	TMPRSS9	1.19E-10
cg00378510	rs3795039	19	LINGO3	19	SPPL2B	LSM7	TMPRSS9	1.17E-08
cg00378510	rs2074546	19	LINGO3	19	SPPL2B	LSM7	TMPRSS9	1.56E-08
cg00378510	rs3746289	19	LINGO3	19	SPPL2B	LSM7	TMPRSS9	3.27E-08
cg00378510	rs730417	19	LINGO3	19	NA	LINGO3	LSM7	5.02E-08
cg00378510	rs7251424	19	LINGO3	19	LSM7	LINGO3	LSM7	8.26E-08
cg08447739	rs3750848	10	HTRA1	10	ARMS2	PLEKHA1	HTRA1	4.15E-10
cg08447739	rs10490924	10	HTRA1	10	ARMS2	PLEKHA1	HTRA1	7.89E-10
cg08447739	rs3750847	10	HTRA1	10	ARMS2	PLEKHA1	HTRA1	2.03E-09
cg08447739	rs3793917	10	HTRA1	10	HTRA1	ARMS2	HTRA1	2.23E-09
cg08447739	rs932275	10	HTRA1	10	HTRA1	ARMS2	DMBT1	8.69E-09
cg11841529	rs3746226	20	CD40	19	LOC100131296	ZNF543	ZNF304	1.89E-09
cg11841529	rs13345625	20	CD40	19	ZNF543	ZNF460	LOC100131296	3.08E-09

IlmnID, Illumina probe ID, Meth CHR, the chromosome where methylation probe is located; SNP CHR, the chromosome where SNP is located

### 6.3.5 Association among cotinine levels, smoking status, and methylation levels

I examined the association of cotinine values in 29 subjects (2 non-smokers, 4 ex-smokers, and 23 current-smokers) with self-reported smoking status and methylation levels at the 39 smoking-DMPs. In theory, there should be a negative correlation between the cotinine levels with methylation if the methylation was indeed lower in smokers. However, among the 39 smoking-DMPs, only 3 sites had nominally significant effects, and none of the correlation effects were consistent with the self-reported smoking effects. This was because there were very few (only 2) non-smokers in the data, and their cotinine levels were in fact higher than those of most of the current- or ex-smokers (Figure 6-10), which indicates that the cotinine levels of the 29 subjects were likely not representative of their smoking status.



**Figure 6-10. Cotinine levels in different smoking status**

## 6.4 Discussion

The genome-wide results from this chapter show that smoking can impact DNA methylation and gene expression levels in adipose tissue. To my knowledge, this is the first study that performed genome-wide analyses of smoking in adipose tissue DNA methylation and gene expression profiles. The key result was that a subset of smoking-DMPs identified in this chapter showed tissue-shared effects with blood, while the remainder are novel adipose-specific smoking DMPs.

This study found 39 smoking-DMPs and 35 smoking-differentially expressed genes (48 exons) in adipose tissue. Together, they overlapped at 4 genes (*CYP1B1*, *CYTL1*, *AHRR*, *F2RL3*). Some of these genes are associated with metabolic diseases, such as type 2 diabetes and hypertension from disease-related enrichment analysis (details in Chapter 5). At these 4 overlapping genes, methylation levels were mostly negatively correlated with gene expression levels (Figure 6-5, 11 negative and 1 positive correlation). Of the 12 methylation probes in the 4 overlapping genes, only probe cg00512031 in the *CYTL1* gene was at the gene promoter, and a negative association here is consistent with the expectation that promoter-based CpG-sites negatively associate with gene expression (Eckhardt *et al.*, 2006; Ball *et al.*, 2009; Lister *et al.*, 2009). However, the majority of the remaining probes were in the gene body, thus a negative association with expression in these genes is against the expectation of a positive correlation between methylation and gene expression if the CpG site was located in the gene body. Some studies have reported both positive and negative correlations between methylation and expression on the gene body (Zilberman *et al.*, 2007; Zemach *et al.*, 2010; Jjingo *et al.*, 2012; Gutierrez-Arcelus *et al.*, 2013). An explanation for this finding could also be that DNA methylation sites in the gene body are in fact located in alternative promoters that regulate the expression of particular isoforms, as in the case of alternative splicing.

Follow-up the adipose results, in blood, I identified some tissue-shared smoking-DMPs. In the blood dataset alone, there were 12 smoking-DMPs in 6 genes (*AHRR*, *GPR15*, *F2RL3*, *GFII*, *CCL28*, *PRSS23*) and 2 inter-genic regions (chromosomes 2q37.1 and 6p21.33), which were all previously identified in the literature. Tissue-shared smoking effects in adipose and blood methylation could be observed on 4 genes (*AHRR*, *F2RL3*, *GFII*, 2q37.1) in our data alone. The one result that overlapped across all four datasets

(methylation and expression in adipose and blood) was the *AHRR* gene. This region appears to have the strongest and most stable smoking-related changes in our data, and the most likely model explaining the associations at this region is a scenario where smoking impact expression levels, which in turn modulate DNA methylation patterns.

### 6.4.1 Tissue-shared and adipose-specific smoking-DMPs

To date, many studies have performed the smoking EWAS in adult samples and in newborns, in order to detect the maternal smoking effects on infant methylome. The major findings have been smoking differential methylation sites in multiple tissues, and replication in different populations. Some studies also found the methylation levels are ‘reversible’ on some of these differential sites, that is, methylation levels of ex-smokers could gradually change more towards to non-smokers after quitting smoking. Table 6-6 summarise the most up to date smoking EWAS using Illumina platforms (Illumina 27k and Illumina 450k).

**Table 6-6. Overview of recent smoking EWASs**

Smoking	Tissue	Platform	Key findings
Adult <sup>1</sup>	Whole blood, peripheral blood mononuclear cells, small airway epithelial tissue, lymphoblast, alveolar macrophage	Illumina 27k; Illumina 450k; pyrosequencing (validation)	Numerous smoking differential methylation sites have been identified in multiple tissues and in different populations (e.g. <i>AHRR</i> , <i>CYP1B1</i> , <i>2q37.1</i> , <i>GFII</i> , <i>GPR15</i> , <i>F2RL3</i> ). Methylation levels of some of these sites are ‘reversible’ when quitting smoking (e.g. <i>F2RL3</i> ).
Newborn (prenatal) <sup>2</sup>	Cord blood, placenta	Illumina 27k; Illumina 450k; pyrosequencing (validation)	Maternal tobacco use is associated with newborn methylation changes on plentiful methylation sites. Some of these CpG sites showed persistently patterns into adolescent.

<sup>1</sup>(Breitling *et al.*, 2011; Monick *et al.*, 2012; Wan *et al.*, 2012; Buro-Auriemma *et al.*, 2013; Philibert *et al.*, 2013; Shenker *et al.*, 2013; Sun *et al.*, 2013; Zeilinger *et al.*, 2013; H. Zhang *et al.*, 2013; Besingi & Johansson, 2014; Dogan *et al.*, 2014; H. R. Elliott *et al.*, 2014; Harlid *et al.*, 2014; Guida *et al.*, 2015)

<sup>2</sup>(M. Suter *et al.*, 2011; Joubert *et al.*, 2012; Breton *et al.*, 2014; Markunas *et al.*, 2014; Ivorra *et al.*, 2015; K. W. Lee *et al.*, 2015; Richmond *et al.*, 2015)

Table 6-7 summarizes 9 of our adipose smoking-DMPs that are also well-known smoking-DMPs from previous 27k or 450k EWAS studies (see full reference list on Table 6-6). These studies were mainly done in blood or lung samples and in different ethnic groups, suggesting that there is a consistent smoking effect on DNA methylation across multiple tissues in the human body.

**Table 6-7. List of well-known smoking-DMPs identified from previous studies**

CHR	Gene name	Studies
1	<i>GFII</i>	Joubert et al. 2012; Zeilinger et al. 2013; Besinigi and Johansson 2014; Dogan et al. 2014; Elliott et al. 2014
2	2q37.1 region	Shenker et al. 2013; Sun et al. 2013; Zeilinger et al. 2013, Besinigi and Johansson 2014; Dogan et al. 2014; Elliott et al. 2014; Markunas et al. 2014
5	<i>AHRR</i>	Monick et al. 2012; Philibert et al. 2013; Shenker et al. 2013; Sun et al. 2013; Zeilinger et al. 2013, Besinigi and Johansson 2014; Dogan et al. 2014; Elliott et al. 2014; Markunas et al. 2014
9	<i>NOTCH1</i>	Dogan et al. 2014
11	<i>LRP5</i>	Zeilinger et al. 2013; Besinigi and Johansson 2014; Dogan et al. 2014
14	<i>C14orf43</i>	Zeilinger et al. 2013; Dogan et al. 2014; Elliott et al. 2014
15	<i>CYP1A1</i>	Joubert et al. 2012; Buro-Auriemma et al. 2013; Zeilinger et al. 2013
19	<i>LINGO3</i>	Zeilinger et al. 2013
19	<i>F2RL3</i>	Breitling et al. 2011; Joubert et al. 2012; Wan et al. 2012; Shenker et al. 2013; Sun et al. 2013; Zeilinger et al. 2013; Besinigi and Johansson 2014; Dogan et al. 2014; Elliott et al. 2014; Markunas et al. 2014; Zhang et al. 2014

In addition to tissue-shared smoking-DMPs, our results also highlight adipose-specific smoking-DMPs, which are novel smoking-DMPs that have not been previously identified by other smoking EWAS and do not validate in the twin blood sample. There 13 novel smoking-DMPs identified in adipose tissue in the following genes: *CYP1B1*, *CYTL1*, *NEURL1B*, *GPER*, *SAG*, *LATS2*, *PMS2L11*, *NEDD9*, *PDE7B*, *SLC398A*, *CDC42EP3*, *HTRA1*, and *ACVRL1*. I discuss some of these results in section 6.4.2 below. Further studies of smoking-methylation effects in adipose samples will be required to replicate these results.

Three published smoking EWAS in blood samples were further compared to better understand the extent of tissue-shared effects of the top 39 CpG sites found in my adipose smoking EWAS. These 3 studies included a large sample EWAS in Germans (Zeilinger *et al.*, 2013), and two smaller sample size EWAS in Asians (H. R. Elliott *et*

*al.*, 2014) and in African Americans (Dogan *et al.*, 2014). The results are summarized in Table 6-8. The overall pattern showed that *GFII*, 2q37.1 region, *AHRR*, *LRP5*, *C14orf43*, *F2RL3*, *HTRA1*, *CYP1A1*, and *LINGO3* were consistently reported in more than one study. The methylation levels on several CpG sites, for example, on 2q37.1 region (cg05951221, cg21556642, cg09140273), on *AHRR* (cg05575921), on *LRP5* (cg21611682), on *C14orf43* (cg22851561), and on *F2RL3* (cg03636183) were constantly lower in current smokers.

**Table 6-8. Tissue-shared smoking differential sites in adipose EWAS**

Chr	Gene	CGid	TwinsUK/ Adipose/349	TwinsUK/ Blood/306	Zeilinger/ blood/2272	Elliott/blood 192/Asians*	Dogan/blood 111.African A.
1	GFII	cg14179389	●	●	(5 ● )	(1 ● )	(2 ● )
2	CDC42EP3	cg07992500	●	●			
2	CYP1B1	cg02162897 cg20408276	● ●	● ●			
2	2q37.1 region	cg03329539 cg06644428 cg05951221 cg21566642 cg01940273	● ● ● ● ●	● ● ● ● ●	● ● ● ● ●	● ● ● ● ●	● ● ● ● ●
2	SAG	cg04341454	●	●			
4	CYTL1	cg00512031	●	●			
4	SLC39A8	cg13735704	●	●			
5	AHRR	cg24980413 cg05575921 cg25648203 cg19405895	● ● ● ●	● ● ● ●	● (13 ● 2 ● )	● (7 ● 1 ● )	● (11 ● )
5	NEURL1B	cg22418620 cg03646542	● ●	● ●			
6	NEDD9	cg12531611	●	●			
6	PDE7B	cg01985595	●	●			
6		cg25767832	●	●			
7	GPER;C7orf50	cg11461808	●	●			
7	PMS2L11	cg01727317	●	●			
9	NOTCH1	cg14120703	●	●			(11 ● )
10	HTRA1	cg08447739	●	●			
11	LRP5	cg21611682	●	●	● (2 ● )		(1 ● )
12	ACVRL1	cg20131897	●	●			
13	LATS2	cg25576788	●	●			
14	C14orf43	cg22851561	●	●	● (5 ● )	(1 ● )	●
15	CYP1A1	cg23160522 cg00353139 cg23680900 cg10009577 cg26516004	● ● ● ● ●	● ● ● ● ●	(1 ● )		
19	LINGO3	cg00378510	●	●	(1 ● )		
19	F2RL3	cg03636183	●	●	●	●	●
19	KLK14	cg05242523	●	●			
20	CD40	cg11841529	●	●			

The column labels, separated by back slash in order, are study name, tissue, and sample size. The top 39 probes discovered from adipose tissue are indicated in Column 3, 4, and those correspondingly validated in blood (column 5) and three EWAS studies (column 6, 7, 8). Directions for smoking-DMPs are indicated as hypo-methylation (blue) or hyper-methylation (red). Deeper colours (of blue and red) indicate significance for the CpG site and lighter colours indicates not significant. Parenthesis indicates extra probes not overlapping with adipose tissue. For example, the probe in the first row shows significantly hypo-methylation in adipose tissue, but not significant hypo-methylation in blood, and other studies reported 5, 1, and 2 extra significant probes for the *GFII* gene. In total, 5 probes across 3 genes (2q37.1, *AHRR*, *F2RL3*) overlapped between adipose and blood (*GFII* was significant however not sharing the same probe).



## 6.4.2 Highly replicated Smoking-DMPs

In this section I discuss several of the most significant smoking-DMPs in my study that may also be replicated in the literature (*AHRR*, *F2RL3*, and chromosome 2q37.1).

### 6.4.2.1 *AHRR* (aryl hydrocarbon receptor (*AhR*) repressor) gene

The most-replicated smoking-DMP is in the *AHRR* gene. *AHRR* is a protein-coding gene for the AhR signalling cascade that mediates dioxin toxicity and cell growth and differentiation. Smoking might activate AhR similar to dioxin intake (Kasai *et al.*, 2006). *AHRR* is also a tumour suppressor gene, and persistent down-regulation of *AHRR* mRNA is found in multiple human malignant tissues (Zudaire *et al.*, 2008).

On the *AHRR* gene, multiple CpG sites were smoking-DMPs, including the most significant replicated marker cg05575921. This marker was the most significant hit that I identified in the blood EWAS ( $P = 2.42 \times 10^{-21}$ ), and remained highly significant with only 26 current smokers included in our study. In the adipose EWAS, 4 smoking-DMPs at the *AHRR* have a less significant P value. The methylation levels of these 4 markers were weakly negatively correlated with expression despite being located on the gene body. One potential explanation is that smoking is associated with a particular isoform of *AHRR*, because it was the last exon of *AHRR* that was up regulated in non-smokers in both adipose and blood. It is possible that the smoking-DMPs in *AHRR* may regulate the expression of this isoform. Another study also showed a negative correlation between increasing methylation and expression at cg05575921 ( $P < 0.03$ ) (Monick *et al.*, 2012). This further suggests that there might be some functional regulation between methylation and expression on this marker.

Furthermore, methylation on cg05575921 was strongly associated with smoking cessation time (Zeilinger *et al.*, 2013). This study found 174/187 smoking-DMPs that were significantly differentially methylated between ex-smokers and non-smokers. The methylation levels of 36 CpG sites decreased with cessation years of ex-smokers. Among the 36 CpG sites, cg05575921 was the most significant 'cessation site' ( $P = 7.73 \times 10^{-44}$ ) and cessation time could explain 21.48% of the methylation variance. On these cessation sites, the methylation levels in ex-smokers would gradually move closer

towards the methylation levels in non-smokers after quitting smoking, and in 60 years time the methylation could return to the same levels as that in non-smokers. Although the time since cessation not recorded in my dataset, 27/39 CpG sites showed the same trend, where the methylation levels of ex-smokers are in between the levels of non-smokers and current smokers (Table 6-2). It suggests that methylation levels at these 27 CpG sites might change with cessation time.

#### **6.4.2.2 *F2RL3* (coagulation factor II receptor-like 3) gene (also known as PAR-4)**

The CpG site cg03636183 on *F2RL3* gene was consistently identified as a smoking-DMP in the majority of smoking EWASs. The CpG site is located on the second exon of the gene on chromosome 19, and is the only associated CpG site in *F2RL3*. A study reports that *F2RL3* methylation is highly associated with mortality from coronary disease (Breitling *et al.*, 2011). The methylation levels on cg03636183 are lower in smokers, and the methylation levels increased after smoking cessation (Breitling *et al.*, 2011; Zeilinger *et al.*, 2013; Y. Zhang *et al.*, 2014). In a mouse study (N = 5), the gene expression levels of *F2RL3* increased after smoking exposure (Shenker *et al.*, 2013), although the results did not reach nominal significance, they suggested that smoking could lead to the methylation changes on *F2RL3* gene. In our data, smoking associates with both DNA methylation and gene expression changes, and the most likely model explaining this association is that smoking impacts expression via methylation. The pattern of association is also consistent with previous studies in both humans and in the mouse.

#### **6.4.2.3 *2q37.1* region (close to *ALPP/ALPPL2* genes)**

CpG-sites in the inter-genic region on chromosome 2 were the top loci in the adipose smoking EWAS and second top hit in the blood smoking EWAS. Several smoking-associated CpG sites were identified in this region in many of the previous smoking EWAS. Three alkaline phosphatase genes surrounded this region: alkaline phosphatase placental-like 2 (*ALPPL2*), alkaline phosphatase intestinal (*ALPI*), and alkaline phosphatase placental (*ALPP*). However, all significant CpG sites identified in our EWAS are located in the 3'UTR of both *ALPPL2* and *ALPI*, and the functional

significance of these methylation changes is still unknown. In my expression analysis, none of the exons of these genes were differentially expressed with smoking status, and there was no correlation between methylation and expression levels. One possible explanation for the methylation variation in this region might be a genetic contribution. Among the 5 significant CpG sites at the 2q37.1 region, 2 sites were identified as trans-meQTLs at a marginal significance (the most significant P value for cg03329539 is  $P = 2.90 \times 10^{-6}$  and  $P = 3.08 \times 10^{-6}$  for cg05951221). This suggests that there could be an interaction between genotype and smoking on the methylation levels in this region, but further studies are required to confirm this.

#### **6.4.2.4 Lung cancer related genes**

Several studies examined whether smoking-DMPs were also associated with lung cancer, and I also checked whether the top 39 smoking-DMPs found in adipose tissues and 12 smoking-DMPs found in blood samples were also candidate DMRs for lung cancer. In a Korean population, methylation at one of my top smoking-DMPs (*CYP1B1*) was found to correlate with smoking status by comparing 80 non-small cell lung carcinoma (NSCLC) tissue samples to 16 normal lung tissues (Kang *et al.*, 2012). Furthermore, two of the top findings *LRP5* and *ACVRL1* in adipose tissue overlapped with their findings, and they found that gene expression levels at these genes were significantly associated with methylation levels. In another study, the methylation levels of 59 matched lung adenocarcinoma and non-tumour lung pairs were compared, and 164 hyper-methylated smoking genes and 57 hypo-methylated smoking genes were identified (Selamat *et al.*, 2012), the majority of which were negatively associated with gene expression. Two of my smoking-DMPs, *CYTL1* and *ACVRL1* overlapped with their findings. In addition, some of the smoking-DMPs and smoking-expressed genes were differentially expressed (*AHRR*, *ZNF385B*, *EDC3*, *CYGB*, *COBL*, and *F2RL3*) and differentially methylated (*CYTL1*, *LRP5*, and *ACVRL1*) in lung cancer. In conclusion, there is some evidence that the smoking-DMPs identified in this study are also smoking-DMPs in lung cancer tissue.

#### **6.4.2.5 Maternal smoking effect on newborns**

Several EWAS studies have examined the prenatal maternal smoking effect on the newborn methylome (see full reference list on Table 6-6). One compared 1,062 newborn cord blood methylomes to plasma cotinine levels in the mother, and identified 26 CpG sites (that mapped to 10 genes) that were significantly associated (Joubert *et al.*, 2012), and 3 genes (*AHRR*, *CYP1A1*, and *GFII*) were replicated in an independent cohort. In another study, 185 CpG sites were identified with altered methylation by maternal smoking in 889 infants, and the same 3 genes were validated (Markunas *et al.*, 2014). Two of these genes, *AHRR* and *GFII*, were both identified as smoking-DMPs in adipose tissue and blood samples in my study, and *CYP1A1* was identified in adipose tissue. Therefore, effects at these genes are not only robust across tissues in adults of different ethnicities, but also appear for smoking exposure at a different developmental stage. However, most of the smoking-DMPs from the adult samples were not found in the newborn studies, suggesting a distinct effect of direct smoking exposure in adult versus indirect exposure *in utero*.

#### **6.4.2.6 Smoking-DMPs and disease**

Apart from *CYTL1*, *CYP1B1* and *ACVRL1* genes that have been found to differentially methylate in the lung cancer tissues (Kang *et al.*, 2012; Lokk *et al.*, 2012; Selamat *et al.*, 2012), other novel smoking-DMPs were differentially methylated with certain diseases. Some genes were differentially methylated with cancers, mostly breast cancer, such as *CYP1B1* (acute lymphocytic leukaemia and breast cancer), *GPER* (breast cancer), and *HTRAI* (breast cancer). *LATS2* was previously identified as a tumour suppressor gene. In breast cancer, smoking is considered as a risk factor because it is a carcinogen. From a large longitudinal cohort of 111,140 participants, the hazard ratio of breast cancer was 1.06% in ever smokers (current- and ex-smokers) relative to never smokers (Xue *et al.*, 2011). This suggests that methylation on these genes, smoking, and breast cancer may be causally related.

### **6.4.3 Cotinine levels, smoking status, and methylation**

Serum cotinine has been a gold standard for measuring smoking status. In the TwinsUK cohort, cotinine was only available for limited subjects. A critical issue was that cotinine was not necessarily obtained on the same clinical visit date as the date of DNA sample for DNA methylation analysis. Another problem was that the half-life of cotinine is very short (< 24 hours) and in many cases did not match the date of blood collection. In future, I would like to extend this work by performing a cotinine EWAS in a larger sample of twins.

### **6.4.4 Genetic contributions to smoking-DMPs and differentially expressed exons**

In adipose tissue, I found that 19/39 smoking-DMPs that were meQTLs and no exons (out of 48 exons) were eQTLs at a stringent genome-wide significance level of  $P < 10^{-7}$ . Comparing this result with other studies, Grundberg et al. identified more meQTLs at 1% FDR from 648 samples (my sample of 349 subject was a subset of their dataset) (Grundberg *et al.*, 2013), and for the expression data, an on going EuroBATS eQTL analysis conducted by Buil *et al.* (Buil *et al.*, in prep) identified 19/35 expressed exons as eQTLs.

Among the 19 meQTLs, there were 4 CpG sites on the *AHRR* gene and 2 CpG sites on 2q37.1 region identified as *cis*-meQTLs, suggesting that genetic contribution has some impact on the methylation changes in addition to the smoking effect. If methylation was affected by genetic variants, the effect should presumably be present across tissues. Since smoking-DMPs at these two regions were actually identified in both adipose tissue and blood samples, the methylation changes on these two regions could be influenced by the underlying genetic sequence. Future analyses will focus on identifying gene-by-smoking interactions of DNA methylation levels.

### **6.4.5 Smoking-DMPs or expressed genes overlap with GWAS results**

Several differentially methylated and expressed genes from my results were also associated with smoking-related diseases. There were candidate genes for lung disease, for example, *CYP1A1* and *CYP1B1* (lung cancer), and *HHIP* (pulmonary function). Several genes were identified as candidate genes for metabolic or cardiovascular diseases, such as *ZNF385B* (sudden cardiac arrest), *KCNJ11* (type 2 diabetes), *NMNAT2*, *CADM2*, and *PZP* (obesity-related traits), *HDAC9* (coronary artery disease and stroke), *LRP5* (bone mineral density), *SLC39A8* (cholesterol, BMI, and blood pressure).

### **6.4.6 Conclusion**

In conclusion, I have identified adipose-specific and tissue-shared DNA methylation changes related to smoking and corresponding gene expression associations with smoking. Smoking exerts a strong effect on DNA methylation and gene expression levels across at least two tissue types (blood and adipose tissue). These results indicate that DNA methylation levels at smoking-DMPs may be good biomarkers of smoking status and strongly suggest that smoking should be incorporated as a covariate in EWAS studies.

# Conclusions, Discussions and Future Perspectives

---

Epigenetics is an invaluable area of biology that can bridge the gap between the study of genotype and phenotype. Results from 1300 GWAS studies published since 2005 show that the genetic contributions cannot fully explain many diseases and phenotypic traits. Therefore, studies have searched for unknown modifications, such as epigenetic modifications, to explain the ‘missing heritability’. The missing heritability typically refers to the phenotypic variance that cannot be fully explained by the genetic effect. To possibly contribute to the heritability, epigenetic changes need to be stably maintained throughout lifetime, and faithfully passed down to different generations (i.e. meQTLs). Feinberg and Irizarry suggest that the variability of phenotype can be mediated by epigenetic modifications (Feinberg & Irizarry, 2010). For example, the epigenetic modifications act as a mediator between the gene and environment, and then contribute to disease.

The study of epigenetics in the context of human complex traits and environmental exposures has expanded rapidly in the past years. The number of publications using keywords related to epigenetics according to Thomson Reuters Web of Knowledge has quadrupled from 2000 to 8000, during 2001-2011. When my PhD began in 2010, there was a transition from epigenetic studies of target genes to applications of the genome-wide Illumina 27k array. Now, in the span of four years, the most commonly used platforms provide a nearly twenty-fold increase in the number of probes on the Illumina 450k, and EWAS studies are widely performed with more phenotypes, diseases and environmental exposures.

My thesis covers broad aspects of EWAS studies, with applications to age and age-related phenotypes. It involves the analysis of statistical power, exploring DNA

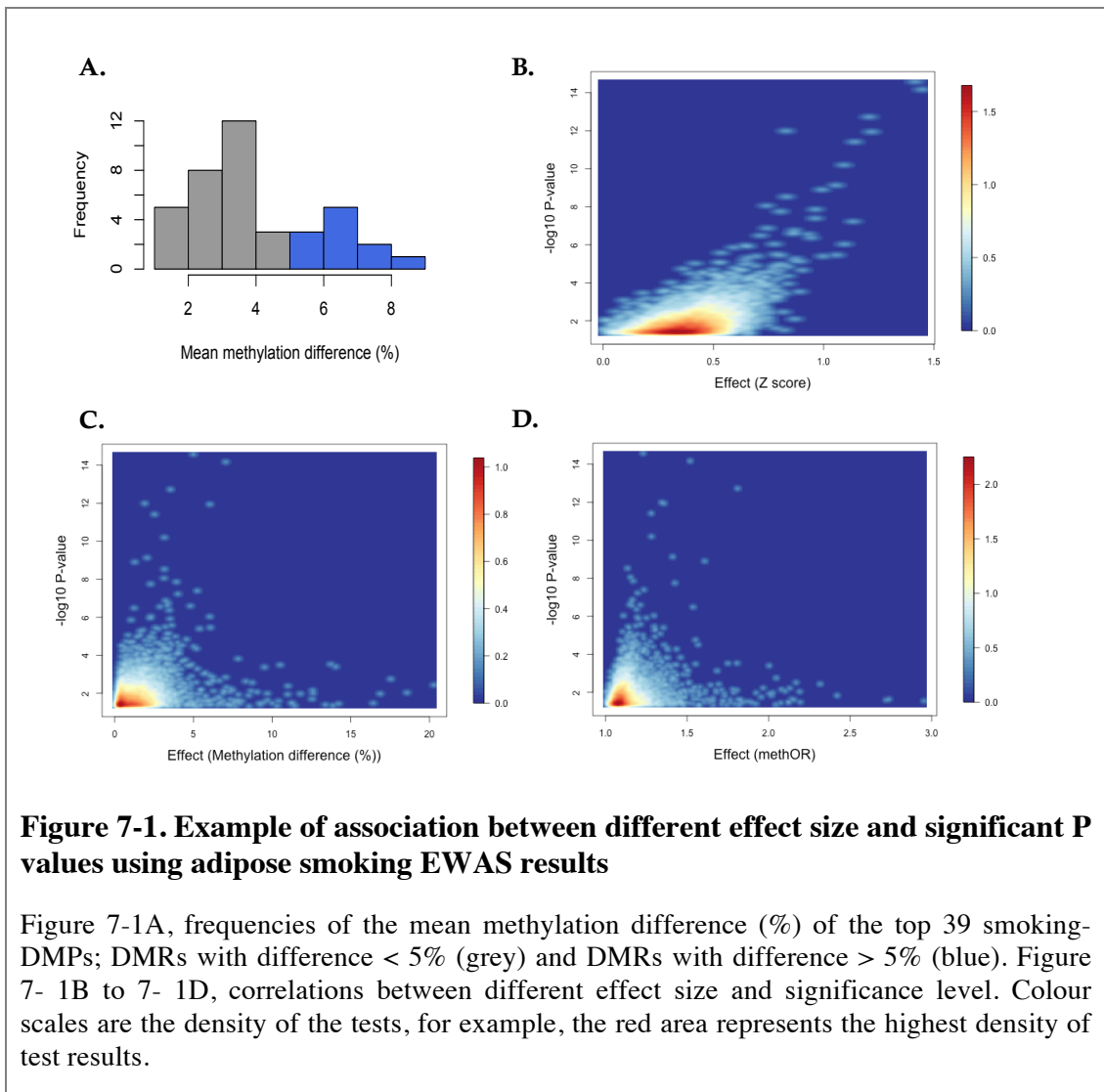
methylation quality control and methodology, finding the covariates that influence an EWAS study, and applying EWAS methods to the study DNA methylation changes related to ageing, birth weight, and smoking. Below is a brief discussion of what I consider to be the key findings from each of my research chapters.

The estimation of statistical power in EWAS is very important, but has not been comprehensively explored. Most studies use power and sample size estimation of a traditional genetic epidemiology study. However, unlike the power estimation in epidemiological studies where power can be often increased by simply acquiring larger sample sizes, there are qualitative differences in an epigenetic study, since the epigenome changes throughout the life of an individual (Relton & Davey Smith, 2010; Heijmans & Mill, 2012). My power estimates in EWAS were based on a disease discordant MZ study and a case-control study of unrelated subjects. Under a number of assumptions, I simulated several scenarios of DNA methylation effects on complex traits, and explored the sample size needed to reach power to detect differential methylation, as well as the factors that influence power. I found that the power of an EWAS study is influenced by the sample size, the methylation effect size, the variance in DNA methylation within and across groups, and the correlation in DNA methylation levels between groups. The power estimates that I provide are informative for future and on-going EWAS of both discordant twins and case-control studies, and my results for the case-control design are consistent with previous estimates.

One of the findings from my results on power estimation was that the methylation difference between groups on its own is not a very good measure of effect size, because the same DNA methylation difference could result in different EWAS power estimates. The smoking EWAS results are a good example to explain why using the mean methylation difference is not a good representative of effect size, as I discussed in Chapter 2. Here, I compared 35 current smokers and 186 non-smokers, and found 39 smoking differential methylation positions/regions at genome-wide significant  $P < 4.72 \times 10^{-6}$ . From the absolute mean methylation difference at these 39 smoking-DMPs, only 11 (28.2%) had  $> 5\%$  difference between smokers and non-smokers (Figure 7-1A). If I only focus on nominal significance ( $P = 0.05$ ), there were 12,874 smoking-DMPs/DMRs. I compared the correlation between the nominal  $P$  values ( $-\log_{10}$  transformed) and effect size, using the following measures of effect size: (1) the



absolute methylation difference between current and non-smokers (percentage); (2) methylation odds ratio (methOR, where all values were transformed to be greater than 1 as ‘risk OR’); and (3) the Z score (absolute methylation difference/pooled SD of the two groups). Figure 1B-1D shows the correlations between the P values and the relative effects. The Z score had a linear relationship with significance as expected, but not the methylation difference or methOR. In fact, there were no significant tests found in the sites with large effect sizes (top 50<sup>th</sup> percentile). In addition, the majority of significant test results (99%) had < 5% of methylation difference. This shows the importance of exploring several measures to determine the methylation ‘effect size’ in an EWAS. Smoking is a good example to explore these concepts with, because the smoking-DMPs/DMRs used here have been replicated by multiple studies.



EWAS criteria for genome-wide significance also need to be discussed. Across the genome, there are regions of variable DNA methylation, as well as less variable regions. Using a Bonferroni correction is a stringent method to account for multiple testing and it is questionable whether all the probes should be included in the denominator for Bonferroni correction. Another consideration is that methylation patterns from nearby probes are often co-methylated, while the Bonferroni denominator assumes independence. Other methods should also be used to take multiple testing into account, such as the false discovery rate or ideally permutation-based methods, which can take the correlated structure of the data into account.

The normalization and quality control procedure in an EWAS are constantly being updated and refined. In this thesis, I used many iterations and tests of these procedures as the Illumina 27k was updated with the Illumina 450k array, and multiple methylation analysis packages were proposed in the literature. The major pitfalls in analysing the Illumina 450k data are batch effect adjustment and data normalisation. One study suggests that batch effect in Illumina 450k data can increase the false discovery of differential methylation, so it is important to randomly allocate samples during experiment (Harper *et al.*, 2013). Because the Illumina 450k platform requires bisulfite-treated DNA, the incomplete conversion of unmethylated cytosines will be detected as methylated cytosines and therefore lead to incorrect methylation levels. Two common ways to adjust for batch effects (i.e. plate number, position on the plate) are either to apply a surrogate variable analysis (SVA; (Leek & Storey, 2007)) or to incorporate the batch effects as covariates. Both of these methods have their own limitation: the former requires at least two samples on a single plate, while later may cause the problem of overfitting when using multivariable model in a dataset of low sample size. Although I have not included explicit tests comparing different normalization approaches to the same datasets, these were performed for some of the data chapters of the thesis, often numerous times. The final pipeline that I used included quality control steps to ensure that only data from reliable probes were considered, steps to ensure that there were no individual outliers in the sample, and several steps to identify and correct for covariates that may influence DNA methylation patterns. Currently, there is no consensus step-by-step pipeline for analysing Illumina 450k data, although many packages deal with normalization. I have compiled what I thought were the most relevant and important

quality control steps for the datasets presented in this thesis. Generally, quality control and analyses should be customized for each dataset.

It is worthwhile to note that there are limitations to performing EWAS on the Illumina 450k array. Compared to sequencing-based platform, Illumina 450k shows relatively low coverage of 5% of  $10^7$  CpG sites across genome, therefore it might not be an ideal validation for sequencing-based platform. In addition, this platform was designed for human samples only, and cannot be used for allele-specific methylation detection. One should be careful to define a ‘significant’ differential methylation purely based on the statistical P values. In some EWAS studies, I have noticed certain differential methylated sites could shift back and forward of being ‘significant’ under different analysis model. This highlights the importance of interpreting results in context of their biological meanings rather than only by statistical meanings (Bock, 2012).

My first phenotype-related chapter aimed to identify methylome changes related to the ageing process. I identified a large number of a-DMPs across tissues and samples, and found that a-DMPs shared across tissues tended to be hyper-methylated and enriched on CpG islands. These a-DMPs require further investigation into their biological role, as previous studies have suggested that a-DMPs are concentrated in bivalent chromatin domain promoters and polycomb group protein target genes, and hyper-methylation of CpG islands is implicated with gene silencing. In this study, thousands of CpG sites showed age-related changes in methylation. The majority of these effects were hyper-methylated with age, a large proportion replicated in an independent sample, and some changes were observed in multiple tissues. These findings indicated that a-DMPs are less likely stochastic events, but instead associate with biological mechanisms involved in ageing and potential longevity.

The consistent and replicated a-DMP results in the literature led to the development of prediction models using a-DMPs to predict chronological age, and translate the predictions into biological function by proposing the concept of ‘methylation age acceleration’ relative to chronological age. Among these prediction models, I selected Horvath's method because it was built on a large database across different tissue samples. Using his method on my 450k datasets, I found that methylation age generally correlated well with chronological age, however I detected more variation in the blood over skin and adipose tissues. Another central idea is that age acceleration relates to

age-related phenotypes. If so, methylation-based measures of age acceleration could become an important biomarker for age-related diseases. A further question is that if this 'ageing clock' can be slowed down with certain intervention, this may lead to treatment or amelioration of age-related diseases. In my analysis I found some correlations of methylation-based measures of age acceleration with age-related phenotypes, however, more work needs to be done to elucidate the exact meaning of these results.

Birth weight is the second phenotype that I examined. My working hypothesis was that the maternal environment triggers changes in birth weight with corresponding changes to the newborn's methylome that persist and can be observed in adult life. This is the first study to compare the twin birth weight difference as a continuous trait with the difference in the adult methylome in birth weight discordant MZ twins. I did not find any genome-wide significant BW-DMP because the study was largely underpowered. One explanation is that the methylation patterns between healthy co-twins are quite similar in the relatively modest sized datasets that I explored. My results were consistent with the small number of BW-DMPs identified in newborns (Gordon *et al.*, 2012), and with lack of genome-wide significant results from another sample of adult MZ twins (Souren *et al.*, 2013). If low birth weight is epigenetically driven, the signal could further be reduced or wiped out during the ageing process. From the pathway analysis, when I loosen the significance criteria of BW-DMPs, they appear to relate to genes that associate with metabolism and cardiovascular diseases.

Smoking is the last phenotype that I examined, and is considered to be the strongest environmental effect to the methylome identified to date. The smoking-DMPs have been well defined, replicated, and validated across tissues. I found previously identified smoking-DMPs and smoking differentially expressed genes as well as novel DMRs. This is the first study to identify smoking effects in the methylome and transcriptome in adipose tissue, which could be a highly informative for disease since both fat and smoking are risk factors for the cardiovascular disease. A striking finding is that many of the smoking-DMPs that occur in adipose tissue, have been found in blood, which indicates tissue-shared effects. Four of these genes were consistently methylated and differentially expressed with smoking status.

Few studies have documented the genome-wide smoking effect not only in DNA methylation, but also on the gene expression at the exon level. I found several differentially expressed genes that also harboured smoking-DMPs and were differentially expressed in lung cancer. The majority of smoking-DMPs were hypomethylated and the majority of the differentially expressed genes were up regulated, suggesting that there could be a biological relationship between methylation and expression, potentially triggered by smoking. The direct correlations between methylation and gene expression were lower than expected, given that the methylation probes were located in the gene body, where methylation typically shows a positive correlation with gene expression. However, in an extended analysis I found that there could be a trans- rather than cis- effect between the two levels. For example, the smoking related *AHRR* gene was thought to govern two smoking-related gene *CYP1B1* and *CYP1A1*, so the methylation of *AHRR* might impact on expression of *CYP1B1* and *CYP1A1*.

Current EWAS studies have progressed onto deeper coverage sequencing data, such as WGBS and MeDIP-seq data. These new sequencing technologies could reveal a more complete view of the methylome, and discover more differential methylation sites across the whole genome. Similar to the current array data, challenges faced by NGS data are the data QC and analysis. There will be even lower thresholds for multiple testing with more coverage, and most of these platforms are expensive so fewer samples can be profiled. These issues will impact power of future EWAS.

My recommendation for future methylation studies is to integrate epigenetics with other '-omics' data, for example, exploring methylation and gene expression to understand gene regulation mechanisms, and studying the interplay between methylation and histone modifications and their relation to chromatin structure. Currently, many studies are focused on epigenetic epidemiology lack biological evidence of progression from methylation change to phenotype and disease. Longitudinal studies are needed to prove causal or a consequential association between the methylation and phenotype.

In summary, the phenotypes that I examined in the thesis have yielded novel differentially methylated site, as well as replicating previous findings. One of the key messages of my results is that covariates, such as age and smoking, should always be included in EWAS studies because these a-DMPs and smoking-DMPs demonstrated a

strong and consistent effect across multiple population samples and cell types. The findings from the power estimation and BW EWAS, both showed the importance of effect size and sample size required for EWAS to reach statistical power to detect differential methylation effects in human complex traits.

# Appendix A: Epigenome-Wide Association Scans in Osteoarthritis

---

Here I present my early work using EWAS method in osteoarthritis, both with discordant MZ twin design and case-control design. Due to the low sample size, this study is under power and therefore included as an appendix. Although the differential methylation identified in the study does not meet genome-wide significance, some of the top genes are found to be associated with osteoarthritis in genetic studies, suggesting it might potentially have biological meanings.

---

## A1. Introduction

Osteoarthritis (OA) has been defined as the clinical and pathological outcome of a range of disorders resulting in the structural and functional failure of synovial joints. It is characterized by the synovial inflammation, destruction of the extracellular matrix of articular cartilage, and bone remodelling. OA is age-related and shows a higher prevalence in older individuals, for example, 1 in every 5 adults, aged between 50 and 59 has OA, and furthermore, almost 1 in every 2 adults aged 80 and above has painful OA in one or both knees. In terms of its anatomical distribution, OA frequently affects the joints of the hand, spine, hips, and knees. Previous studies have identified several risk factors, such as age, female gender, BMI, bone mass, history of injury/trauma, and genes (Valdes & Spector, 2011). For this study, we have concentrated on OA of the hips and knees, as they represent a considerable morbidity, to the extent that the most severe forms have resulted in over 100,000 total joint replacement surgeries in the UK every year.

From large-scaled GWAS studies, several genes have been identified to confer OA susceptibility. These candidate genes are associated with the function of the cartilage,

and obesity. Using 1,341 cases and 3,496 controls in European population, Kerkhof et al. (Kerkhof *et al.*, 2010) have found an OA candidate gene *COG5* (OR = 1.14,  $P = 8 \times 10^{-8}$ ) that locates on chromosome 7q22 with replication in a separate 13,497 cases and 40,000 controls. Residing on the same chromosome region, the *DUS4L* gene (OR = 1.15,  $P = 6 \times 10^{-8}$ ) is another OA candidate gene obtained using a meta-analysis of European and East Asian population (Evangelou *et al.*, 2011). Aside from the genes that locate on chromosome 7, other OA-candidate genes have been reported: *MCF2L* (13q34, OR = 1.17,  $P = 2 \times 10^{-8}$ ) (Day-Williams *et al.*, 2011), *DOTIL* (19p13.3,  $P = 1 \times 10^{-11}$ ) (Castano Betancourt *et al.*, 2012), *GNL3*, *PBRM1*, *SNORD19* (3p21.1, OR = 1.09,  $P = 5 \times 10^{-9}$ ), *MIR4642*, *NUDT19P4* (6p21.1, OR = 1.08,  $P = 6 \times 10^{-7}$ ), *RIMKLBP2*, *ZC3H11B* (1q41, OR = 1.07,  $P = 1 \times 10^{-6}$ ), and obesity related gene *FTO* (16q12.2, OR = 1.07,  $P = 4 \times 10^{-6}$ ) (Zeggini *et al.*, 2012). Specifically, those that associated with OA of the knee, Valdes et al. (Valdes *et al.*, 2008) located two candidate regions, *PTGS2* and *PLA2G4A* on 1q31 (OR = 1.59,  $P = 3 \times 10^{-6}$ ) and *PARD3B* on 2q33 (OR = 1.46,  $P = 6 \times 10^{-6}$ ) in the European population. In another study, the OA of the knee candidate gene *BTNL2* (OR = 1.31,  $P = 5 \times 10^{-9}$ ) was identified in the Japanese population using sequencing techniques (Nakajima *et al.*, 2010).

The heritability in OA of hip is approximately 60% (Spector & MacGregor, 2004), and suggests that missing heritability could have an epigenetic component. Indeed, epigenetic studies have revealed several OA-DMPs and OA-DMPs. The majority of OA-differential methylation are known to be involved with the modelling and maintenance of articular cartilage that is composed of chondrocytes, collagen, and extracellular matrix (ECM), and signal responses to synovial inflammation and pain. Other OA-DMPs would likely to derive from gene expression studies. In gene promoter regions, evidence has shown DNA methylation can down-regulate gene expression to induce phenotypic changes. The promoter of some of metalloproteinase genes, such as *MMP3*, *MMP9*, *MMP13*, and *ADAMTS4* are up-regulated in OA and could influence the transcription-binding factors (Roach *et al.*, 2005). Other genes, such as proinflammatory cytokine *IL-1 $\beta$* , growth differentiation factor *GDF-5*, chondrocyte differentiation gene *SOX9*, and obesity-related *LEP* genes are also differentially methylated with OA (Hashimoto *et al.*, 2009; Barter *et al.*, 2012). However, the association between methylation and expression remains unclear, and the direction of association is at times conflicting. For example, the genes, such as type II collagen



(*COL2A1*) and aggrecan (*ACAN*) are differentially expressed in OA yet the methylation of these genes remains non-differentially hypo-methylated in the healthy or OA subjects (Poschl *et al.*, 2005; Zimmermann *et al.*, 2008).

Beside candidate gene studies, two epigenome-wide association studies (EWAS) have been undertaken for OA, osteoporosis, and healthy subjects using the *Illumina 27k array* (Delgado-Calle *et al.*, 2013; Fernandez-Tajes *et al.*, 2013). One study (Delgado-Calle *et al.*, 2013) found that the methylation levels of the bones of 27 subjects with hip fractures and 26 hip OA subjects have 241 CpG sites that located on the promoter regions of 228 genes to be differently methylated (at Bonferroni-corrected  $P < 0.05$ ). Most of the DMRs identified ( $n = 217$ ) are more methylated in the OA subjects, and these regions enriched for association with bone traits and involved in multiple functional categories, such as homeobox (*HOX*). In the other study, 91 OA differential CpG site are obtained from directly comparing the methylation levels from the cartilage sample of 25 OA subjects to 20 healthy controls. Furthermore, a tight cluster of 1,357 DMRs are found in 7 OA subjects, and 450 of these genes are differentially expressed. These DMRs are established in biological functions, such as regulation of phosphorylation, the protein kinase cascade, morphogenesis and development, inflammatory, and lipid metabolism (Fernandez-Tajes *et al.*, 2013).

To date, many gene expression and methylation studies have been performed on bone and cartilage samples, thus subject sample sizes tend to be small due to the invasiveness of the procedures, which may include complications, thus are difficult to acquire quantity in practice. In this study, we hypothesize that the methylation levels for the OA patients' blood samples that are far less invasive are also differentially methylated. Additionally, our study is designed with age-matched cases and controls as well as focus on the monozygotic twins (MZ) who are genetically identical but discordant for the OA presence.

## A.2 Material and methods

### A.2.1 Datasets

From the TwinsUK cohort, we have identified 9 OA discordant MZ pairs and 1 OA-concordant MZ pair, and total 16 subjects with knee or hip OA. To concurrently examine the epigenetic and genetic association with OA, we used the discordant twin design and case-control design.

In the 9 OA discordant MZ twin design, individual age ranged 57.29 to 80.90 during their visit to our department. In the case-control design, only unrelated subjects were included as cases and controls, so total 16 cases and 30 controls that had matched for age were included. The details of age, height, weight, and Body Mass Index (BMI) are presented in Table A1. When compared to healthy co-twin, the OA subjects have similar heights however increased weight and BMI.

**Table A1. Demographic characteristics of our two study designs**

	OA discordant MZ			Case-control		
	OA-twin	Healthy co-twin	$P^1$	OA (N = 16)	Healthy (N = 30)	$P^2$
Age	69.67 ± 7.90	69.67 ± 7.90	-	68.66 ± 7.48	68.02 ± 7.10	0.87
Height	156.48 ± 5.00	157.10 ± 4.63	0.439	158.37 ± 5.47	160.57 ± 6.01	0.20
Weight	74.21 ± 15.45	68.60 ± 11.35	0.024	71.47 ± 14.73	68.02 ± 12.66	0.53
BMI	30.34 ± 6.45	27.82 ± 4.67	0.059	28.54 ± 6.06	26.36 ± 4.55	0.26

<sup>1</sup>P:  $P$ -value from Wilcoxon paired test; <sup>2</sup>P:  $P$ -value from Wilcoxon rank-sum test

### A.2.2 Phenotypes

In the TwinsUK Adult Twin Registry, all twins were recruited from the general population with self-report of disease status. To verify their disease status, knee X-ray was collected. For cases, the DNA samples were obtained after disease onset, and the DNA samples from the matched controls were obtained within 1-year. For most of the cases ( $n = 14$ ), we matched 2 healthy controls, however 2 of them are paired with 1 control due to the old age. Because BMI has been associated with OA, the height, weight, and BMI information that close to the date for subjects' visit date was included. We have considered the age in analysis as a covariate, and it is defined as the time of DNA extraction instead of gestational age.

### **A.2.3 Illumina Methylation450K data**

The DNA methylation levels were obtained white blood cells from whole blood. We excluded 17,664 probes that mapped to multiple loci in the human genome (hg19) within 2 mismatches (method see (J. T. Bell *et al.*, 2011)), probes with missing value, and probes on sex chromosomes. Out of the initial 485,577 probes, we obtained 454,601 probes for the discordant twin design and 432,827 probes for the case-control design. To confirm those genes that probes mapped to, each probe was mapped to the Homo sapiens genes (GRCh37) and crossed checked using the UCSC genome browser. A probe can be locates on one gene or with a 30 kilobase (kb) in front of the gene start site or after the gene end site depends on the strand direction (forward and reverse, respectively). The methylation distribution of Illumina 450k array has been previously described on Chapter 3, Table 1.

### **A.2.4 Gene expression data**

The gene expression data from LCL (lymphocytes), fat, and blood tissues were measured by *Illumina expression array HumanHT-12 V3*, and the results obtained from the MUTHER study (previously described, see (Grundberg *et al.*, 2012)). The covariates, such as age, batch, and skin concentration levels are known to be confounders and adjusted in the analysis.

### **A.2.5 Statistical analyses**

#### ***A.2.5.1 Quality Control for Illumina 450k data***

As quality control, the methylation distribution was checked for all subjects. To identify outliers, heatmap of correlation of clustering among subjects and boxplot for the methylation distribution were produced. Probes with missing values across subjects were also removed. The mean, median and principle components were used to identify batch effect. For the final analysis, the plate and bisulfite conversion levels were included as the systemic batch effects and biological effect, such as age and BMI were also adjusted in analysis. Due to the smaller sample size of OA discordant twin pairs, the position on the plate was not taken into consideration as a systemic batch effect.

### **A.2.5.2 OA EWAS**

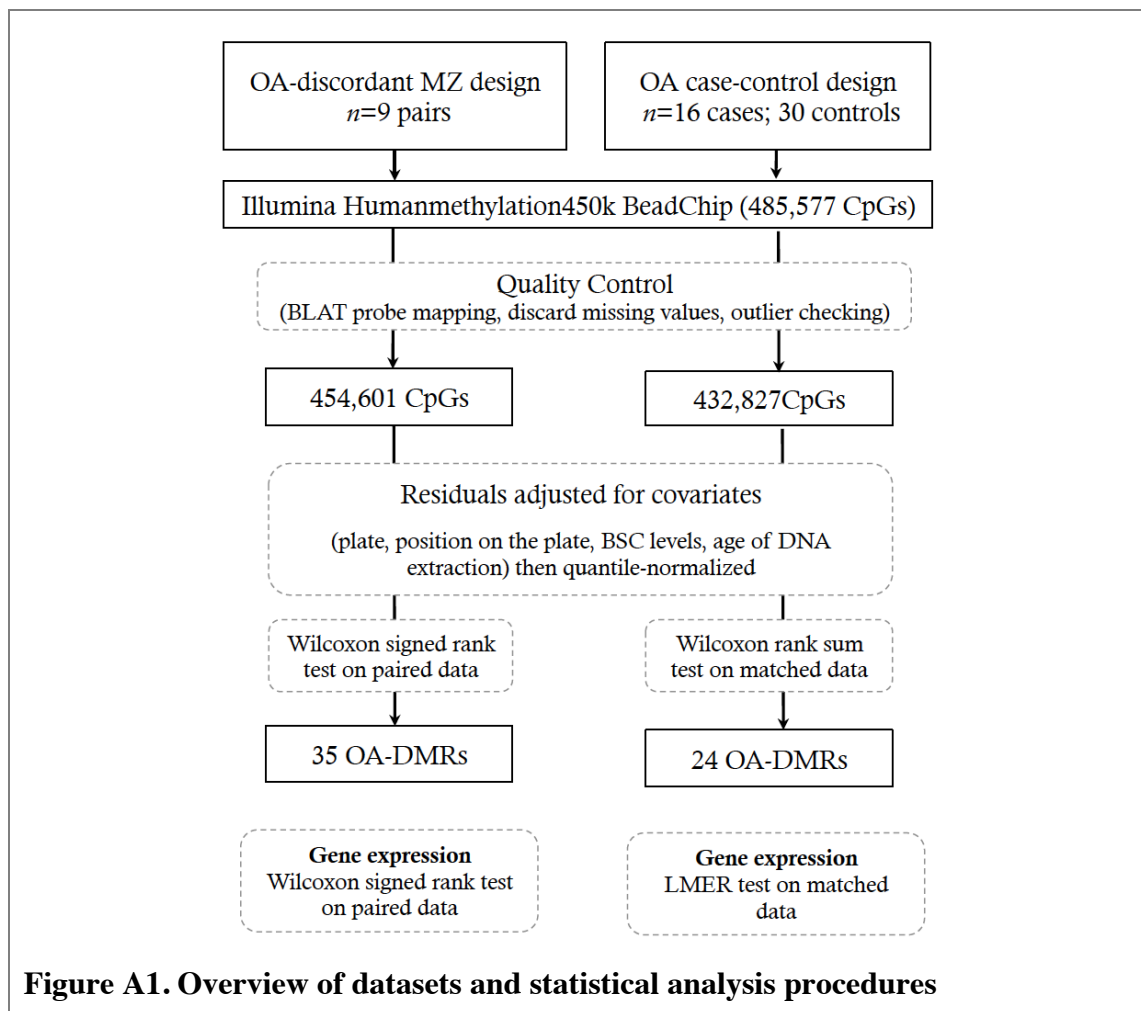
Prior to analysis, raw methylation levels were adjusted for all the covariates and residuals taken from linear model. To better compare between individuals, the methylation residuals were quantile-normalized (Bolstad *et al.*, 2003) across subjects where residuals were ranked to make the overall methylation distribution the same in each subject. For each probe, the Wilcoxon signed rank test was applied between the paired discordant pairs, and the Wilcoxon rank sum test was applied between the OA subjects and healthy controls. Due to the small sample size of both datasets, probes were considered to be a candidate gene at a locus-specific P value of 0.01.

### **A.2.5.3 Gene expression association with OA and methylation**

The gene expression datasets from MuTHER study have been described previously (Grundberg *et al.*, 2012). Gene expression probes of genes proximal to top hits found in the two designs were extracted from the whole expression array. In the expression data, 3 tissues, LCL, Fat, and Skin were included for most of the total 866 subjects. Because few of our cases and discordant twins from the methylation analysis also have the gene expression data, the new case-controls and discordant twin pairs were included in the analysis. In summary, there are 13, 11, and 10 OA discordant pairs in Fat, LCL, and skin tissues. For the case-control design, 13 unrelated cases and 26 healthy age-matched controls were included. To check the association between the DNA methylation and gene expression of the top hits, the expression levels of genes identified from the methylation analysis were extracted ( $n = 70$ ). Gene expression levels on each probe were quantile-quantile normalized before fitting to a linear model adjusted for covariates (age, BMI, batch, skin concentration levels), and quantile-normalized across subjects. Wilcoxon signed rank test was applied to compare within pair differences in the discordant twin design. In the case-control design, the age-matched case and two controls were assigned to the same group, and taken as a random effect of the analysis. A linear mixed effect model (LMER) that comprised of group and disease status was applied to the methylation residuals for each probe. For differential expression with OA, probes with P value less than 0.05 were considered.

#### A.2.5.4 Methylation quantitative trait locus (meQTL)

To examine the genetic contribution to the top probes we have identified from both designs, the association between methylation probes and SNPs within 500kb regions were checked using PLINK. There were 135 unrelated subjects with non-missing genotypes and phenotypes included in the cis-meQTL checking. The methylation levels on these probes were adjusted for age, BMI and other batch effects then quantile-quantile normalized to fit the linear association model. A probe is considered to be cis-meQTL with the genome-wide significance level of  $10^{-8}$ . An overall scheme of our analysis is shown in Fig A1.



## A.3 Results

### A.3.1 OA differential methylation analysis

#### A.3.1.1 OA discordant MZ twin design

To identify the OA-associated methylation regions, we have compared methylation differences within the 9 discordant MZ pairs. At P value < 0.01, 35 probes were differentially methylated with OA and their details shown in Table A2. The genotypes and gene expression of several genes, such as *FOXO3*, *SAMD11*, *COL9A2*, *BET3L/TRAPPC3L*, and *FOXO4* have been previously associated with OA either in humans or mouse. Though not directly associated with OA, genes such as the zinc finger transcription factors (*ZNF841* and *ZNF432*), *RPS6KA2*, *ROBO3*, and *PMF1-BGLAP*, have been implicated in other skeletal diseases, cartilage differentiation, and osteoporosis.

**Table A2. Top 35 OA-DMPs found in the discordant MZ twin design**

IlmnID	CHR	OA-dir <sup>1</sup>	Gene <sup>2</sup>	Related disease/phenotype
cg12801619	6	Hypo	<i>FOXO3</i>	The OA in both human and mice is associated with the changes in <i>FOXO</i> expression and activation, and therefore involved in cartilage aging. A pathway analysis of gene expression profile indicates the cluster of <i>FOXO3</i> , <i>ZBTB16</i> , and <i>SLC6A3</i> genes are related to OA which involved reproductive process in a multicellular organism (B. Zhang <i>et al.</i> , 2013)
cg05527507	1	Hypo	<i>SAMD11</i> , <i>LOC100130417</i>	<i>SAMD11</i> could promote cell proliferation. Compare subjects with AIS to non-AIS osteoblasts, <i>SAMD11</i> down-regulated gene expression. (Fendri <i>et al.</i> , 2013)
cg05518543	4	Hypo	<i>MAEA</i>	Lung cancer, multiple myeloma, prostate cancer
cg27584097	19	Hyper	<i>RYR1</i>	Central core disease, myopathy, hyperthermia
cg18186394	1	Hypo	<i>COL9A2</i>	Candidate gene for hip osteoarthritis (identified by MRI) using 345 twins (Nakki <i>et al.</i> , 2011) up-regulated gene associated with loss of cells and abnormalities of matrix in degenerated discs (Y. G. Zhang <i>et al.</i> , 2010)
cg12020682	6	Hypo	<i>FUCA2</i>	Gastric cancer and other carcinoma
cg02660117	1	Hypo	<i>BCAR3</i> , <i>MIR760</i>	BCAR3- breast cancer; MIR760- early detection of colorectal cancer
cg09391898	11	Hypo	<i>FOXRED1</i> , <i>SRPR</i> , <i>TIRAP</i>	FOXRED1- neuropathy, breast cancer, Parkinson's disease; SRPR- lung disease; TIRAP- immune diseases, such as RA, leukaemia, SLE, etc.
cg22824291	11	Hypo	<i>Loc100996455</i>	
cg01011367	10	Hypo	<i>ACADSB</i>	Hypertension, isovaleric academia, Alzheimer's diseases, TB
cg01284619	19	Hypo	<i>HCN2</i>	Epilepsy, inflammatory and neuropathic pain

1

**Table A2. Top 35 OA-DMPs found in the discordant MZ twin design (continued)**

IlmnID	CHR	OA-dir <sup>1</sup>	Gene <sup>2</sup>	Related disease/phenotype
cg27325460	13	Hypo	<i>TPT1-AS1</i>	
cg12909732	7	Hypo	<i>RAMP3</i>	Heart disease, prostate and pancreas disease
cg17437086	5	Hypo	<i>LPCAT1</i>	Colorectal cancer
cg20550677	22	Hypo	<i>BIK, TLL1</i>	BIK- multiple cancer and carcinoma; <i>TLL1</i> - ciliary dyskinesia, prostate cancer
cg14379719	8	Hypo	<i>GATA4</i>	Heart disease, tumour
cg22849672	6	Hypo	<i>MPC1 (BRP44L)</i>	Malaria
cg24480379	1	Hypo	<i>SELENBP1</i>	Multiple cancers and carcinoma
cg12354014	6	Hypo	<i>FAM260, TRAPPC3L (BET3L)</i>	<i>TRAPPC3L</i> and <i>BET3L</i> are paralogues and both of them are involved in the network forming collagen genes (e.g. <i>COL8A2, COL10A1A</i> ) (Aldea <i>et al.</i> , 2013)
cg03353124	18	Hypo	<i>TMEM200C</i>	
cg25683989	12	Hyper	<i>HVCN1</i>	Breast cancer
cg15309862	X	Hyper	<i>FOXO4, MED12</i>	In cartilage from the mice with surgically induced OA, <i>FOXO4</i> gene was activated independent of <i>ADAMTS-5</i> activity. (Bateman <i>et al.</i> , 2013); <i>MED12</i> -hypothyroidism, mental diseases
cg16684846	19	Hypo	<i>ZNF841, ZNF432</i>	Both involved in the transcriptional regulation. Previously study shows one zinc finger transcription factor (ZFP60) a negative regulator of cartilage differentiation.
cg16790416	10	Hypo	<i>RBM20</i>	Cardiomyopathy
cg23281382	6	Hypo	<i>RPS6KA2</i>	Aging of bone marrow-derived mesenchymal stem cell (bmMSC) may play a role in age-related skeletal diseases. <i>RPS6KA2</i> has been reported as a tumour suppressor gene and likely to decrease the proliferation rate of human bmMSC.
cg26694386	1	Hypo	<i>DES12, AX747555</i>	<i>DES12</i> - malaria, adenocarcinoma
cg06568260	X	Hypo	<i>SYAP1, TXLNG</i>	<i>SYAP1</i> - breast cancer, hepatocellular carcinoma; <i>TXLNG</i> - scarlet fever
cg05086956	11	Hypo	<i>ROBO3</i>	<i>ROBO3</i> was found to expressed in synovial fibroblasts of both osteoarthritis (weak expressed) and rheumatoid arthritis patients. In the study it also suggest that deregulation of the <i>ROBO3</i> receptor in synovial fibroblasts in OA (& RA) correlates with aggressiveness of the fibroblasts.
cg05119316	6	Hypo	<i>HLA-F-AS1</i>	SLE, lupus
cg19692149	1	Hypo	<i>SYDE2</i>	Neuron diseases
cg25465065	1	Hyper	<i>PMF1-BGLAP, PMF1</i>	<i>PMF1-BGLAP</i> - the locus represents the read-through transcription between <i>PMF1</i> and <i>BGLAP</i> . <i>BGLAP</i> is previous found to be associated with osteoporosis, bone loss, and OA. This join locus might as well have function on OA. <i>PMF1</i> -Parkinson's disease
cg03713666	10	Hypo	<i>INPP5A</i>	Carcinoma, SZ
cg14201544	9	Hypo	<i>NELFB, NRARP</i>	Cholangiocarcinoma, breast carcinoma
cg05338731	22	Hyper	<i>RAB36, RTDR1</i>	Rhabdoid tumours;
cg25140773	5	Hypo	<i>ISOC1, MIR4633</i>	Uterine fibroid

OA-dir: "Hyper" means methylation levels are higher in the OA subjects than their co-twin; "Hypo" means methylation levels in healthy subjects are higher than OA subjects; <sup>2</sup>Gene: genes with bold are associated with OA from previous studies; genes with underlines are potential candidate genes for OA and other bone diseases.

### A.3.1.2 OA case-control design

In the case-control design, methylation levels were compared between cases (N = 16) and controls (N = 30). At P value < 0.001, 24 probes were differentially methylated with OA and their details shown in Table A3. Among these genes, the knock-out of *MMP14* gene (highlight in bold in Table A3) in mice has been reported to induce the arthritis-like symptoms (Holmbeck *et al.*, 1999), and the gene expression levels of both *MMP14* and *CALCR* were higher in OA subjects (Holmbeck *et al.*, 1999; Zupan *et al.*, 2012). Some of butyrophilin-like genes, such as *BTNL3*, *BTNL8*, and *BTNL9* were identified however no evidence shows how it associates with OA and furthermore, these genes are paralogs with the previously identified OA candidate genes *BTNL2*. Other genes, *SUPV3L1* and *BCL3*, were associated with spinal disc degeneration and cartilage remodelling.

**Table A3. Top 24 OA-DMPs found in case-control analysis**

IlmnID	CHR	OA-dir <sup>1</sup>	Gene <sup>2</sup>	Related disease/phenotype
cg24758392	5	Hypo	<i>BTNL3</i>	BTNL3 is paralog for BTNL2, which has previously identified as an OA candidate gene.
cg13409216	10	Hyper	<i>SUPV3L1</i>	Up-regulated gene associated with loss of cells and abnormalities of matrix in degenerated discs (Y. G. Zhang <i>et al.</i> , 2010)
cg18413710	14	Hyper	<b><i>MMP14</i></b> , <i>MRPL52</i> , <i>SLC7A7</i>	MMP14- deletion of MMP14 cause arthritis-like symptoms in mouse (Holmbeck <i>et al.</i> , 1999) In HUVECs, MMP14 expression found to be <i>highly simulated</i> by both OA and shear stress (P. Wang <i>et al.</i> , 2013); SLC7A7- lysinuric protein intolerance, osteoporosis, carcinoma
cg23994061	2	Hypo	<i>COLEC11</i>	Hepatitis, 3MC syndrome type 2
cg12242345	10	Hypo	-	
cg16094954	19	Hyper	<i>BCL3</i>	BCL3 can be induced by IL-1 $\beta$ then activate matrix metalloproteinase-1, which is known to enable the degradation of type II collagen. (S. F. Elliott <i>et al.</i> , 2002) Because it's involved in the cartilage remodelling, it might associate with OA. (Palmer & Chen, 2008)
cg02352685	5	Hypo	<i>BTNL8</i>	BTNL8 is a paralog for BTNL2, which has previously identified as an OA candidate gene.
cg25690715	17	Hyper	<i>SEPT9</i>	Neuralgic amyotrophy, ovarian neoplasms; FAM65B- prostatitis
cg04356381	6	Hyper	<i>FAM65B</i>	
cg03422651	16	Hyper	<i>TBC1D24</i> , <i>NTN3</i> , <i>ATP6V0C</i>	TBC1D24- neuronitis, focal epilepsy; NTN3- TB, leukaemia, neuronitis; ATP6V0C- osteopetrosis, kidney disease
cg13095704	7	Hypo	<b><i>CALCR</i></b> , <i>GNGT1</i>	<i>Higher gene expression</i> of CALCR in OA compare to osteoporosis subjects (human bone tissues) (Zupan <i>et al.</i> , 2012); GNGT1- cancer, Huntington's disease
cg15025536	7	Hypo	<i>CARD11</i>	Immune disease, lymphoma
cg13695075	14	Hypo	<i>C14ORF25</i> , <i>FOXA1</i>	FOXA1- osteoporosis, multiple cancers
cg00003722	10	Hypo	-	
cg06129556	15	Hyper	<i>CSNK1G1</i>	Cell growth



**Table A3. Top 24 OA-DMPs found in case-control analysis (continued)**

IlmnID	CHR	OA-dir <sup>1</sup>	Gene <sup>2</sup>	Related disease/phenotype
cg11081186	5	Hyper	LINC00847, HEIH, MGAT1	MGAT1- muscle disease, neuronitis, insulin resistance, obesity
cg27395288	20	Hyper	MAPRE1	Multiple carcinoma and cancer, neuronitis
cg17239761	22	Hyper	PI4KA, SNAP29	PI4KA- hepatitis, SZ; SNAP29- neuronitis, SZ
cg22860643	10	Hyper	SHOC2, BBIP1	SHOC2- Noonan syndrome-like disorder with loose anagen hair; BBIP1- Bardet-Biedl syndrome
cg25366315	5	Hypo	<u>BTNL3</u> , <u>BTNL9</u>	Both BTNL3 and BTNL9 are paralogs for BTNL2, which has previously identified as an OA candidate gene.
cg26089220	11	Hyper	LOC440040	
cg04211927	7	Hypo	ATP6V0A4	ATP6V0C- osteopetrosis, kidney disease
cg26219797	6	Hyper	GTF3C6, RPF2	
cg06826283	7	Hypo	PRKAR1B	Multiple cancers, SLE

<sup>1</sup>OA-dir: “Hyper” means methylation levels are higher in the OA subjects than that in their co-twin; “Hypo” means methylation levels in healthy subjects are higher than OA subjects; <sup>2</sup>Gene: genes with bold are found to be associated with OA from the previous studies; genes with underlines are potential candidate genes for OA and other bone diseases.

### A.3.2 OA-differentially expressed genes from both designs

For genes identified from the OA-DMP analysis, their expression levels were also compared using the discordant twin design and case-control design. TableA4 shows the significant (P value < 0.05) genes that are differentially expressed with OA among all three tissues. The top two OA-DMPs *BTNL3* and *FOXO3* shows expression differences in multiple tissues.

**Table A4. Differentially expressed genes in different tissues (P < 0.05)**

IlmnID	Chr	Gene	ProbeID	Design	Tissue	Dir	Dir_Meth
cg24758392	22	<i>BTNL3</i>	ILMN_2355786	Twin	LCL	Hypo	Hypo
cg25366315			ILMN_1660446	CaCo	SKIN	Hyper	
cg12801619	6	<i>FOXO3</i>	ILMN_1712515	CaCo	SKIN	Hypo	Hypo
cg09391898	11	<i>TIRAP</i>	ILMN_1776703	Twin	LCL	Hyper	Hypo
			ILMN_1812432	CaCo	FAT	Hyper	
cg20550677	22	<i>TLLI</i>	ILMN_2372795	Twin	FAT	Hyper	Hypo
				CaCo	FAT	Hyper	
cg03713666	10	<i>INPP5A</i>	ILMN_1664608	Twin	LCL	Hypo	Hypo
cg18413710	14	<i>MRPL52</i>	ILMN_1713966	Twin	LCL	Hyper	Hyper
cg18413710	14	<i>SLC7A7</i>	ILMN_1810275	Twin	FAT	Hyper	Hyper
cg25140773	5	<i>ISOC1</i>	ILMN_1764861	CaCo	LCL	Hyper	Hypo
cg22849672	6	<i>BRP44L</i>	ILMN_1666967	CaCo	LCL	Hyper	Hypo

**Table A4. Differentially expressed genes in different tissues (P < 0.05) (continued)**

IlmnID	CHR	Gene	ProbeID	Design	Tissue	Dir	Dir_Meth
cg27584097	19	RZR1	ILMN_1682062	CaCo	LCL	Hypo	Hyper
			ILMN_2411781	CaCo	SKIN	Hypo	
cg06129556	15	CSNK1G1	ILMN_1704713	CaCo	LCL	Hyper	Hyper
			ILMN_1740549	CaCo	FAT	Hyper	
cg14201544	9	NRARP	ILMN_1697666	CaCo	FAT	Hypo	Hypo
cg05338731	22	RAB36	ILMN_1733045	CaCo	FAT	Hyper	Hyper
cg03422651	16	TBC1D24	ILMN_2060212	CaCo	FAT	Hypo	Hyper
cg23281382	6	<u>RPS6KA2</u>	ILMN_1790801	CaCo	FAT	Hypo	Hypo
cg12909732	7	RAMP3	ILMN_2065745	CaCo	FAT	Hypo	Hypo

### A.3.3 meQTL test on top OA-DMP genes

For top 59 OA-DMPs, none of them were found to be cis-meQTL locus.

## A4. Discussion

In this study, we have analysed the large-scale epigenome-wide DNA methylation profile from human white blood cells using discordant MZ twin design and case-control design. Our result indicates that several genes previously associated with OA, cartilage differentiation, and other bone diseases also show differential methylation. The methylation changes in the white blood cells could potentially serve as effective biomarker for OA detection.

From our OA discordant analysis, we found two forkhead box class O (*FOXO*) genes, *FOXO3* (or *FOXO3a*) and *FOXO4* to be differentially methylated. The *FOXO* family members associate with longevity, cardiovascular disease, neurodegenerative disease, and multiple cancers. One potential pathway for *FOXO* members to be involved in OA may be through the inflammation process and neutrophil apoptosis. In particular, there have been few and debatable reports about the role of apoptosis of neutrophils in OA. One study (A. L. Bell *et al.*, 1995; Ivanovska, 2012) shows that neutrophil survival was inhibited in the synovial fluid of OA patients, and as *FOXO3* can regulate the apoptosis of neutrophils, it may associate with OA.

Another interesting finding is the cartilage related gene *COL9A2* (collagen, TYPE IX, Alpha 2). Candidate gene studies and linkage analysis have shown they encode for structural proteins of cartilage ECM, for example, type II collagen gene *COL2A1*, type IX collagen gene *COL9A1* (hip OA) (Mustafa *et al.*, 2000), and type XI collagen genes *COL11A1* and *COL11A2* (Vikkula *et al.*, 1995; Richards *et al.*, 1996). *COL9A2* is associated with many bone diseases, such as multiple epiphyseal dysplasia in two familial-based linkage study (van Mourik *et al.*, 1998; Holden *et al.*, 1999) and intervertebral disc disease in two case-control and one linkage study (Annunen *et al.*, 1999; Jim *et al.*, 2005). From a MRI (magnetic resonance imaging)-based OA study that identified 99 candidate SNPs using 345 twins, the *COL9A2* and *COL10A1* are revealed to be associated with hip OA as a predisposing factor (Nakki *et al.*, 2011). These studies implicate *COL9A2* is involved in multiple bone diseases and as an OA candidate gene. Furthermore, *BET3L* (paralog for *TRAPPC3*) are linked to two bone disease related collagen genes *COL8A2* and *COL10A1* orthologues in an animal study, suggesting it might also play an role (Aldea *et al.*, 2013).

In a case-control design, osteoclast specific gene *CALCR* (calcitonin receptor) is reported to be hypo-methylated with OA, while in one study shows a higher gene expression of *CALCR* in the bone tissue of 31 OA patients (Zupan *et al.*, 2012). The potential role for *CALCR* in OA is through calcitonin (*CT*). One study shows calcitonin to have direct protective effects on articular cartilage. It works via *CALCR* to activate the cyclic AMP (cAMP) and protect *COL2A* degradation and joint degenerative disease (Z. Lin *et al.*, 2008). In a more recent study, *CALCR* is identified in the human OA articular cartilage, which supports the assumption that *CT* can have a direct anabolic effect on articular cartilage (Segovia-Silvestre *et al.*, 2011).

In this analysis, we also see many Butyrophilin-Like family genes, such as *BTNL3*, *BTNL8*, and *BTNL9*. The function of these Butyrophilin-Like family genes yet known, but they are all the paralogs for *BTNL2*, that may inactive T cells then contribute to the chronic inflammation in OA (Valdes *et al.*, 2011). From the analysis, there are many genes involved in multiple carcinoma and cancers, which might due to the same genes that are also involved in the inflammatory response or pain response.

However, not many of our top OA-DMPs overlapped with gene expressions. The reason may be two fold. Firstly, the target tissues we used here are LCL, fat, and skin tissues,

which are not the classic tissues used for identifying the OA expressed genes. Secondly, the discordant twins and case-controls we have included in the expression analysis differ from those in the methylation analysis, which might already have the baseline differences compared to the original subjects. Ideally, if we were to assume that methylation down-regulates gene expression to influence the OA occurrence, we should directly compare the gene expression levels to the methylation levels in those OA subjects. However, due to the small overlaps between the two datasets, we are not able to perform such analysis.

The positive strength of this study is that we controlled for most of the genetic contributions using the OA discordant MZ twin pairs, which was the covariates using unrelated subjects. This study is also provides the whole epigenome-wide scans on the methylation profile in twin and case-control designs. One limitation in our study is the restrictive sample size. In spite of the discordant MZ twin study is often considered as the ideal methylation design because it can control the genetic effects to the disease, there is a trade-off to the study power due to lower sample size. Another important point is the medication intakes for OA treatment. The medication could influence methylation levels in blood sample, and minimizes difference of methylation between OA subjects and the healthy subjects.

In summary, we have identified several OA differentially methylated genes in the human while blood cells that are associated with bone diseases. Further studies shall include more OA discordant MZ twins and simultaneously perform a direct comparison between the methylation from MeDIP-seq data and RNA-seq data in the same OA subjects.

# Appendix B: Publications Related to My PhD work

---

Here I list four first-authored publications related to the work presented in this thesis. My major contributions for these publications involved in data analysis and draft writing.

1. Bell, J. T.\*, Tsai, P. C.\*, Yang, T. P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S. Y., Dempster, E. L., Murray, R. M., Grundberg, E., Hedman, A. K., Nica, A., Small, K. S., Dermitzakis, E. T., McCarthy, M. I., Mill, J., Spector, T. D., & Deloukas, P. (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet*, 8(4), e1002629. doi: 10.1371/journal.pgen.1002629

\*My contribution to this work is analysing the differential methylation to age and age-related phenotypes. Related work: Chapter 4.

---

2. Tsai, P. C., Spector, T. D., & Bell, J. T. (2012). Using epigenome-wide association scans of DNA methylation in age-related complex human traits. *Epigenomics*, 4(5), 511-526. doi: 10.2217/epi.12.45

\*I discussed the main considerations for conducting EWASs and compared the age differential findings from 6 studies. Related work: Chapter 1 & Chapter 4.

---

3. Tsai, P. C., & Bell, J. T. (2015). Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int J Epidemiol*. doi: 10.1093/ije/dyv041

\*I performed the permutation-based power estimation for EWAS studies and discuss the key factors that impact EWAS power. Related work: Chapter 2.

---

4. Tsai, P. C., van Dongen, J., Tan, Q., Willemsen, G., Christiansen, L., Boomsma, D. I., Spector, T. D., Valdes, A., & Bell, J. T. (2015). DNA methylation changes in the IGF1R gene in birth weight discordant adult monozygotic twins. *Twin Res Hum Genet*, (In press).

\* I conducted the birth weight EWAS in discordant MZ twins and replicate the top results with other two MZ twin cohorts. Related work: Chapter 5.

---

# References

---

- Adkins, R. M., Thomas, F., Tylavsky, F. A., & Krushkal, J. (2011). Parental ages and levels of DNA methylation in the newborn are correlated. *BMC Med Genet*, *12*, 47. doi: 10.1186/1471-2350-12-47
- Adkins, R. M., Tylavsky, F. A., & Krushkal, J. (2012). Newborn umbilical cord blood DNA methylation and gene expression levels exhibit limited association with birth weight. *Chem Biodivers*, *9*(5), 888-899. doi: 10.1002/cbdv.201100395
- Albert, M. & Peters, A. H. (2009). Genetic and epigenetic control of early mouse development. *Curr Opin Genet Dev*, *19*(2), 113-121. doi: 10.1016/j.gde.2009.03.004
- Aldea, D., Hanna, P., Munoz, D., Espinoza, J., Torrejon, M., Sachs, L., . . . Marcellini, S. (2013). Evolution of the vertebrate bone matrix: An expression analysis of the network forming collagen paralogues in amphibian osteoblasts. *J Exp Zool B Mol Dev Evol*, *320*(6), 375-384. doi: 10.1002/jez.b.22511
- Alegria-Torres, J. A., Baccarelli, A., & Bollati, V. (2011). Epigenetics and lifestyle. *Epigenomics*, *3*(3), 267-277. doi: 10.2217/epi.11.22
- Alisch, R. S., Barwick, B. G., Chopra, P., Myrick, L. K., Satten, G. A., Conneely, K. N., & Warren, S. T. (2012). Age-associated DNA methylation in pediatric populations. *Genome Res*, *22*(4), 623-632. doi: 10.1101/gr.125187.111
- Annunen, S., Paassilta, P., Lohiniva, J., Perala, M., Pihlajamaa, T., Karppinen, J., . . . Ala-Kokko, L. (1999). An allele of COL9A2 associated with intervertebral disc disease. *Science*, *285*(5426), 409-412.
- Antequera, F. & Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A*, *90*(24), 11995-11999.
- Avner, P. & Heard, E. (2001). X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet*, *2*(1), 59-67. doi: 10.1038/35047580
- Baker, J. L., Olsen, L. W., & Sorensen, T. I. (2008). Weight at birth and all-cause mortality in adulthood. *Epidemiology*, *19*(2), 197-203. doi: 10.1097/EDE.0b013e31816339c6
- Ball, M. P., Li, J. B., Gao, Y., Lee, J. H., LeProust, E. M., Park, I. H., . . . Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol*, *27*(4), 361-368. doi: 10.1038/nbt.1533
- Ballestar, E. (2010). Epigenetics lessons from twins: prospects for autoimmune disease. *Clin Rev Allergy Immunol*, *39*(1), 30-41. doi: 10.1007/s12016-009-8168-4
- Banister, C. E., Koestler, D. C., Maccani, M. A., Padbury, J. F., Houseman, E. A., & Marsit, C. J. (2011). Infant growth restriction is associated with distinct patterns of DNA methylation in human placentas. *Epigenetics*, *6*(7), 920-927. doi: 10.4161/epi.6.7.16079
- Barker, D. J. (1992). Fetal growth and adult disease. *Br J Obstet Gynaecol*, *99*(4), 275-276.
- Barker, D. J. (2004). The developmental origins of adult disease. *J Am Coll Nutr*, *23*(6 Suppl), 588S-595S.
- Barlow, D. P., Stoger, R., Herrmann, B. G., Saito, K., & Schweifer, N. (1991). The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature*, *349*(6304), 84-87. doi: 10.1038/349084a0
- Barr, M. L. & Bertram, E. G. (1949). A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature*, *163*(4148), 676.

- Barter, M. J., Bui, C., & Young, D. A. (2012). Epigenetic mechanisms in cartilage and osteoarthritis: DNA methylation, histone modifications and microRNAs. *Osteoarthritis Cartilage*, 20(5), 339-349. doi: 10.1016/j.joca.2011.12.012
- Bartolomei, M. S., Zemel, S., & Tilghman, S. M. (1991). Parental imprinting of the mouse H19 gene. *Nature*, 351(6322), 153-155. doi: 10.1038/351153a0
- Bateman, J. F., Rowley, L., Belluoccio, D., Chan, B., Bell, K., Fosang, A. J., & Little, C. B. (2013). Transcriptomics of wild-type mice and mice lacking ADAMTS-5 activity identifies genes involved in osteoarthritis initiation and cartilage destruction. *Arthritis Rheum*, 65(6), 1547-1560. doi: 10.1002/art.37900
- Battaglia, F. C. & Lubchenco, L. O. (1967). A practical classification of newborn infants by weight and gestational age. *J Pediatr*, 71(2), 159-163.
- Bell, A. C., West, A. G., & Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3), 387-396.
- Bell, A. L., Magill, M. K., McKane, R., & Irvine, A. E. (1995). Human blood and synovial fluid neutrophils cultured in vitro undergo programmed cell death which is promoted by the addition of synovial fluid. *Ann Rheum Dis*, 54(11), 910-915.
- Bell, J. T., Loomis, A. K., Butcher, L. M., Gao, F., Zhang, B., Hyde, C. L., . . . Spector, T. D. (2014). Differential methylation of the TRPA1 promoter in pain sensitivity. *Nat Commun*, 5, 2978. doi: 10.1038/ncomms3978
- Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., . . . Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol*, 12(1), R10. doi: 10.1186/gb-2011-12-1-r10
- Bell, J. T. & Saffery, R. (2012). The value of twins in epigenetic epidemiology. *Int J Epidemiol*, 41(1), 140-150. doi: 10.1093/ije/dyr179
- Bell, J. T. & Spector, T. D. (2011). A twin approach to unraveling epigenetics. *Trends Genet*, 27(3), 116-125. doi: 10.1016/j.tig.2010.12.005
- Bell, J. T., Tsai, P. C., Yang, T. P., Pidsley, R., Nisbet, J., Glass, D., . . . Deloukas, P. (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet*, 8(4), e1002629. doi: 10.1371/journal.pgen.1002629
- Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Test. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289-300.
- Benowitz, N. L. (2008). Clinical pharmacology of nicotine: implications for understanding, preventing, and treating tobacco addiction. *Clin Pharmacol Ther*, 83(4), 531-541. doi: 10.1038/clpt.2008.3
- Besingi, W. & Johansson, A. (2014). Smoke-related DNA methylation changes in the etiology of human disease. *Hum Mol Genet*, 23(9), 2290-2297. doi: 10.1093/hmg/ddt621
- Bestor, T. H. (2000). The DNA methyltransferases of mammals. *Hum Mol Genet*, 9(16), 2395-2402.
- Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., . . . Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4), 288-295. doi: 10.1016/j.ygeno.2011.07.007
- Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., & Gunderson, K. L. (2009). Genome-wide DNA methylation profiling using Infinium(R) assay. *Epigenomics*, 1(1), 177-200. doi: 10.2217/epi.09.14
- Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., . . . Fan, J. B. (2006). High-throughput DNA methylation profiling using universal bead arrays. *Genome Res*, 16(3), 383-393. doi: 10.1101/gr.4410706

- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1), 6-21. doi: 10.1101/gad.947102
- Bjornsson, H. T., Sigurdsson, M. I., Fallin, M. D., Irizarry, R. A., Aspelund, T., Cui, H., . . . Feinberg, A. P. (2008). Intra-individual change over time in DNA methylation with familial clustering. *JAMA*, 299(24), 2877-2883. doi: 10.1001/jama.299.24.2877
- Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nat Rev Genet*, 13(10), 705-719. doi: 10.1038/nrg3273
- Bocklandt, S., Lin, W., Sehl, M. E., Sanchez, F. J., Sinsheimer, J. S., Horvath, S., & Vilain, E. (2011). Epigenetic predictor of age. *PLoS One*, 6(6), e14821. doi: 10.1371/journal.pone.0014821
- Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., . . . Fox, J. (2011). OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika*.
- Boks, M. P., Derks, E. M., Weisenberger, D. J., Strengman, E., Janson, E., Sommer, I. E., . . . Ophoff, R. A. (2009). The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS One*, 4(8), e6767. doi: 10.1371/journal.pone.0006767
- Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2), 185-193.
- Breitling, L. P., Yang, R., Korn, B., Burwinkel, B., & Brenner, H. (2011). Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet*, 88(4), 450-457. doi: 10.1016/j.ajhg.2011.03.003
- Breton, C. V., Siegmund, K. D., Joubert, B. R., Wang, X., Qui, W., Carey, V., . . . Asthma, B. c. (2014). Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. *PLoS One*, 9(6), e99716. doi: 10.1371/journal.pone.0099716
- Brooks, A. M., Byrd, R. S., Weitzman, M., Auinger, P., & McBride, J. T. (2001). Impact of low birth weight on early childhood asthma in the United States. *Arch Pediatr Adolesc Med*, 155(3), 401-406.
- Brown, A. A., Buil, A., Vinuela, A., Lappalainen, T., Zheng, H. F., Richards, J. B., . . . Durbin, R. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. *Elife*, 3, e01381. doi: 10.7554/eLife.01381
- Buganim, Y., Faddah, D. A., & Jaenisch, R. (2013). Mechanisms and models of somatic cell reprogramming. *Nat Rev Genet*, 14(6), 427-439. doi: 10.1038/nrg3473
- Buil, A., Brown, A. A., Lappalainen, T., Vinuela, A., Davies, M. N., Zheng, H. F., . . . Dermitzakis, E. T. (2015). Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet*, 47(1), 88-91. doi: 10.1038/ng.3162
- Burgess, D. J. (2013). Human epigenetics: Showing your age. *Nat Rev Genet*, 14(1), 6. doi: 10.1038/nrg3391
- Buro-Auriemma, L. J., Salit, J., Hackett, N. R., Walters, M. S., Strulovici-Barel, Y., Staudt, M. R., . . . Crystal, R. G. (2013). Cigarette smoking induces small airway epithelial epigenetic changes with corresponding modulation of gene expression. *Hum Mol Genet*, 22(23), 4726-4738. doi: 10.1093/hmg/ddt326
- Cassidy, S. B., Dykens, E., & Williams, C. A. (2000). Prader-Willi and Angelman syndromes: sister imprinted disorders. *Am J Med Genet*, 97(2), 136-146.
- Castano Betancourt, M. C., Cailotto, F., Kerkhof, H. J., Cornelis, F. M., Doherty, S. A., Hart, D. J., . . . van Meurs, J. B. (2012). Genome-wide association and functional studies identify the DOT1L gene to be involved in cartilage thickness and hip osteoarthritis. *Proc Natl Acad Sci U S A*, 109(21), 8218-8223. doi: 10.1073/pnas.1119899109



- Chang, T. K., Chen, J., Pillay, V., Ho, J. Y., & Bandiera, S. M. (2003). Real-time polymerase chain reaction analysis of CYP1B1 gene expression in human liver. *Toxicol Sci*, *71*(1), 11-19.
- Chong, S. & Whitelaw, E. (2004). Epigenetic germline inheritance. *Curr Opin Genet Dev*, *14*(6), 692-696. doi: 10.1016/j.gde.2004.09.001
- Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., . . . Kelsey, K. T. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet*, *5*(8), e1000602. doi: 10.1371/journal.pgen.1000602
- Christensen, B. C. & Marsit, C. J. (2011). Epigenomics in environmental health. *Front Genet*, *2*, 84. doi: 10.3389/fgene.2011.00084
- Christensen, K., Vaupel, J. W., Holm, N. V., & Yashin, A. I. (1995). Mortality among twins after age 6: fetal origins hypothesis versus twin method. *BMJ*, *310*(6977), 432-436.
- Cleary-Goldman, J. & D'Alton, M. E. (2008). Growth abnormalities and multiple gestations. *Semin Perinatol*, *32*(3), 206-212. doi: 10.1053/j.semperi.2008.02.009
- Cuozzo, C., Porcellini, A., Angrisano, T., Morano, A., Lee, B., Di Pardo, A., . . . Avvedimento, E. V. (2007). DNA damage, homology-directed repair, and DNA methylation. *PLoS Genet*, *3*(7), e110. doi: 10.1371/journal.pgen.0030110
- Dabney, A., Storey, J. D., & Warnes, a. w. a. f. G. R. qvalue: Q-value estimation for false discovery rate control. R package version 1.36.0.
- Daxinger, L. & Whitelaw, E. (2012). Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat Rev Genet*, *13*(3), 153-162. doi: 10.1038/nrg3188
- Day, K., Waite, L. L., Thalacker-Mercer, A., West, A., Bamman, M. M., Brooks, J. D., . . . Absher, D. (2013). Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol*, *14*(9), R102. doi: 10.1186/gb-2013-14-9-r102
- Day-Williams, A. G., Southam, L., Panoutsopoulou, K., Rayner, N. W., Esko, T., Estrada, K., . . . Zeggini, E. (2011). A variant in MCF2L is associated with osteoarthritis. *Am J Hum Genet*, *89*(3), 446-450. doi: 10.1016/j.ajhg.2011.08.001
- DeChiara, T. M., Robertson, E. J., & Efstratiadis, A. (1991). Parental imprinting of the mouse insulin-like growth factor II gene. *Cell*, *64*(4), 849-859.
- Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., & Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, *3*(6), 771-784. doi: 10.2217/epi.11.105
- Delgado-Calle, J., Fernandez, A. F., Sainz, J., Zarrabeitia, M. T., Sanudo, C., Garcia-Renedo, R., . . . Riancho, J. A. (2013). Genome-wide profiling of bone reveals differentially methylated regions in osteoporosis and osteoarthritis. *Arthritis Rheum*, *65*(1), 197-205. doi: 10.1002/art.37753
- Dempster, E., Pidsley, R., Schalkwyk, L., Toulopoulou, T., Picchioni, M., Kravariti, E., . . . Mill, J. (2010a). Methylomic profiling in twins discordant for major psychosis. *Twin Res Hum Genet*, *13*(3), 253.
- Dempster, E., Pidsley, R., Schalkwyk, L. C., Toulopoulou, T., Picchioni, M., Kravariti, E., . . . Mill, J. (2010b). Methylomic profiling in twins discordant for major psychosis (Abstracts for the 13th International Congress on Twin Studies Seoul, South Korea, June 4-7, 2010) *Twin Research and Human Genetics*, *13*(3), 253.
- Dempster, E. L., Pidsley, R., Schalkwyk, L. C., Owens, S., Georgiades, A., Kane, F., . . . Mill, J. (2011). Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Hum Mol Genet*, *20*(24), 4786-4796. doi: 10.1093/hmg/ddr416

- Derom, C. A., Vlietinck, R. F., Thiery, E. W., Leroy, F. O., Fryns, J. P., & Derom, R. M. (2006). The East Flanders Prospective Twin Survey (EFPTS). *Twin Res Hum Genet*, 9(6), 733-738. doi: 10.1375/183242706779462723
- DeVeale, B., van der Kooy, D., & Babak, T. (2012). Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS Genet*, 8(3), e1002600. doi: 10.1371/journal.pgen.1002600
- Di, Y. P., Zhao, J., & Harper, R. (2012). Cigarette smoke induces MUC5AC protein expression through the activation of Sp1. *J Biol Chem*, 287(33), 27948-27958. doi: 10.1074/jbc.M111.334375
- Dogan, M. V., Shields, B., Cutrona, C., Gao, L., Gibbons, F. X., Simons, R., . . . Philibert, R. A. (2014). The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. *BMC Genomics*, 15, 151. doi: 10.1186/1471-2164-15-151
- Doi, A., Park, I. H., Wen, B., Murakami, P., Aryee, M. J., Irizarry, R., . . . Feinberg, A. P. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet*, 41(12), 1350-1353. doi: 10.1038/ng.471
- Drong, A. W., Nicholson, G., Hedman, A. K., Meduri, E., Grundberg, E., Small, K. S., . . . Lindgren, C. M. (2013). The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. *PLoS One*, 8(2), e55923. doi: 10.1371/journal.pone.0055923
- Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., . . . Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, 38(12), 1378-1385. doi: 10.1038/ng1909
- Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10, 48. doi: 10.1186/1471-2105-10-48
- Elliott, H. R., Tillin, T., McArdle, W. L., Ho, K., Duggirala, A., Frayling, T. M., . . . Relton, C. L. (2014). Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin Epigenetics*, 6(1), 4. doi: 10.1186/1868-7083-6-4
- Elliott, S. F., Coon, C. I., Hays, E., Stadheim, T. A., & Vincenti, M. P. (2002). Bcl-3 is an interleukin-1-responsive gene in chondrocytes and synovial fibroblasts that activates transcription of the matrix metalloproteinase 1 gene. *Arthritis Rheum*, 46(12), 3230-3239. doi: 10.1002/art.10675
- Engel, S. M., Joubert, B. R., Wu, M. C., Olshan, A. F., Haberg, S. E., Ueland, P. M., . . . London, S. J. (2014). Neonatal genome-wide methylation patterns in relation to birth weight in the Norwegian Mother and Child Cohort. *Am J Epidemiol*, 179(7), 834-842. doi: 10.1093/aje/kwt433
- Evangeliou, E., Valdes, A. M., Kerkhof, H. J., Styrkarsdottir, U., Zhu, Y., Meulenbelt, I., . . . Spector, T. D. (2011). Meta-analysis of genome-wide association studies confirms a susceptibility locus for knee osteoarthritis on chromosome 7q22. *Ann Rheum Dis*, 70(2), 349-355. doi: 10.1136/ard.2010.132787
- Ezzati, M. & Lopez, A. D. (2003). Estimates of global mortality attributable to smoking in 2000. *Lancet*, 362(9387), 847-852. doi: 10.1016/S0140-6736(03)14338-3
- Fagerberg, B., Bondjers, L., & Nilsson, P. (2004). Low birth weight in combination with catch-up growth predicts the occurrence of the metabolic syndrome in men at late middle age: the Atherosclerosis and Insulin Resistance study. *J Intern Med*, 256(3), 254-259. doi: 10.1111/j.1365-2796.2004.01361.x
- Falconer, D., MacKay TFC. (1996). *Introduction to Quantitative Genetics, 4th Ed*: Longmans Green, Harlow, Essex, UK.

- Feil, R. (2006). Environmental and nutritional effects on the epigenetic regulation of genes. *Mutat Res*, 600(1-2), 46-57. doi: 10.1016/j.mrfmmm.2006.05.029
- Feil, R. & Fraga, M. F. (2011). Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet*, 13(2), 97-109. doi: 10.1038/nrg3142
- Feinberg, A. P. & Irizarry, R. A. (2010). Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A*, 107 Suppl 1, 1757-1764. doi: 10.1073/pnas.0906183107
- Fendri, K., Patten, S. A., Kaufman, G. N., Zaouter, C., Parent, S., Grimard, G., . . . Moldovan, F. (2013). Microarray expression profiling identifies genes with altered expression in Adolescent Idiopathic Scoliosis. *Eur Spine J*, 22(6), 1300-1311. doi: 10.1007/s00586-013-2728-2
- Fernandez-Tajes, J., Soto-Hermida, A., Vazquez-Mosquera, M. E., Cortes-Pereira, E., Mosquera, A., Fernandez-Moreno, M., . . . Blanco, F. J. (2013). Genome-wide DNA methylation analysis of articular chondrocytes reveals a cluster of osteoarthritic patients. *Ann Rheum Dis*. doi: 10.1136/annrheumdis-2012-202783
- Ferreira, J. C., Choufani, S., Grafodatskaya, D., Butcher, D. T., Zhao, C., Chitayat, D., . . . Weksberg, R. (2011). WNT2 promoter methylation in human placenta is associated with low birthweight percentile in the neonate. *Epigenetics*, 6(4), 440-449.
- Filiberto, A. C., Maccani, M. A., Koestler, D., Wilhelm-Benartzi, C., Avissar-Whiting, M., Banister, C. E., . . . Marsit, C. J. (2011). Birthweight is associated with DNA promoter methylation of the glucocorticoid receptor in human placenta. *Epigenetics*, 6(5), 566-572.
- Fisher, R. A. (1954). *Statistical Methods for Research Workers (Twelfth ed.)*. Edinburgh: Oliver and Boyd.
- Florath, I., Butterbach, K., Muller, H., Bewerunge-Hudler, M., & Brenner, H. (2014). Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Hum Mol Genet*, 23(5), 1186-1201. doi: 10.1093/hmg/ddt531
- Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., . . . Esteller, M. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A*, 102(30), 10604-10609. doi: 10.1073/pnas.0500398102
- Freathy, R. M., Mook-Kanamori, D. O., Sovio, U., Prokopenko, I., Timpson, N. J., Berry, D. J., . . . McCarthy, M. I. (2010). Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. *Nat Genet*, 42(5), 430-435. doi: 10.1038/ng.567
- Fryer, A. A., Emes, R. D., Ismail, K. M., Haworth, K. E., Mein, C., Carroll, W. D., & Farrell, W. E. (2011). Quantitative, high-resolution epigenetic profiling of CpG loci identifies associations with cord blood plasma homocysteine and birth weight in humans. *Epigenetics*, 6(1), 86-94. doi: 10.4161/epi.6.1.13392
- Fuks, F., Hurd, P. J., Wolf, D., Nan, X., Bird, A. P., & Kouzarides, T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J Biol Chem*, 278(6), 4035-4040. doi: 10.1074/jbc.M210256200
- Gamazon, E. R., Badner, J. A., Cheng, L., Zhang, C., Zhang, D., Cox, N. J., . . . Liu, C. (2013). Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry*, 18(3), 340-346. doi: 10.1038/mp.2011.174
- Garagnani, P., Bacalini, M. G., Pirazzini, C., Gori, D., Giuliani, C., Mari, D., . . . Franceschi, C. (2012). Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*, 11(6), 1132-1134. doi: 10.1111/accel.12005
- Gardiner-Garden, M. & Frommer, M. (1987). CpG islands in vertebrate genomes. *J Mol Biol*, 196(2), 261-282.

- Gertz, J., Varley, K. E., Reddy, T. E., Bowling, K. M., Pauli, F., Parker, S. L., . . . Myers, R. M. (2011). Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet*, *7*(8), e1002228. doi: 10.1371/journal.pgen.1002228
- Gervin, K., Vigeland, M. D., Mattingsdal, M., Hammero, M., Nygard, H., Olsen, A. O., . . . Lyle, R. (2012). DNA methylation and gene expression changes in monozygotic twins discordant for psoriasis: identification of epigenetically dysregulated genes. *PLoS Genet*, *8*(1), e1002454. doi: 10.1371/journal.pgen.1002454
- Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S. L., . . . Singleton, A. B. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet*, *6*(5), e1000952. doi: 10.1371/journal.pgen.1000952
- Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, *128*(4), 635-638. doi: 10.1016/j.cell.2007.02.006
- Gonzalez-Zulueta, M., Bender, C. M., Yang, A. S., Nguyen, T., Beart, R. W., Van Tornout, J. M., & Jones, P. A. (1995). Methylation of the 5' CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing. *Cancer Res*, *55*(20), 4531-4535.
- Gordon, L., Joo, J. E., Powell, J. E., Ollikainen, M., Novakovic, B., Li, X., . . . Saffery, R. (2012). Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Res*, *22*(8), 1395-1406. doi: 10.1101/gr.136598.111
- Gronniger, E., Weber, B., Heil, O., Peters, N., Stab, F., Wenck, H., . . . Lyko, F. (2010). Aging and chronic sun exposure cause distinct epigenetic changes in human skin. *PLoS Genet*, *6*(5), e1000971. doi: 10.1371/journal.pgen.1000971
- Groom, A., Elliott, H. R., Embleton, N. D., & Relton, C. L. (2011). Epigenetics and child health: basic principles. *Arch Dis Child*, *96*(9), 863-869. doi: 10.1136/adc.2009.165712
- Grossniklaus, U., Kelly, W. G., Ferguson-Smith, A. C., Pembrey, M., & Lindquist, S. (2013). Transgenerational epigenetic inheritance: how important is it? *Nat Rev Genet*, *14*(3), 228-235. doi: 10.1038/nrg3435
- Grundberg, E., Meduri, E., Sandling, J. K., Hedman, A. K., Keildson, S., Buil, A., . . . Deloukas, P. (2013). Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet*, *93*(5), 876-890. doi: 10.1016/j.ajhg.2013.10.004
- Grundberg, E., Small, K. S., Hedman, A. K., Nica, A. C., Buil, A., Keildson, S., . . . Spector, T. D. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet*, *44*(10), 1084-1089. doi: 10.1038/ng.2394
- Guida, F., Sandanger, T. M., Castagne, R., Campanella, G., Polidoro, S., Palli, D., . . . Chadeau-Hyam, M. (2015). Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Genet*, *24*(8), 2349-2359. doi: 10.1093/hmg/ddu751
- Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A., . . . Dermitzakis, E. T. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife*, *2*, e00523. doi: 10.7554/eLife.00523
- Han, L., Lin, I. G., & Hsieh, C. L. (2001). Protein binding protects sites on stable episomes and in the chromosome from de novo methylation. *Mol Cell Biol*, *21*(10), 3416-3424. doi: 10.1128/MCB.21.10.3416-3424.2001
- Handunnetthi, L., Handel, A. E., & Ramagopalan, S. V. (2010). Contribution of genetic, epigenetic and transcriptomic differences to twin discordance in multiple sclerosis. *Expert Rev Neurother*, *10*(9), 1379-1381. doi: 10.1586/ern.10.116

- Hannan, L. M., Jacobs, E. J., & Thun, M. J. (2009). The association between cigarette smoking and risk of colorectal cancer in a large prospective cohort from the United States. *Cancer Epidemiol Biomarkers Prev*, *18*(12), 3362-3367. doi: 10.1158/1055-9965.EPI-09-0661
- Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sada, S., . . . Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*, *49*(2), 359-367. doi: 10.1016/j.molcel.2012.10.016
- Harlid, S., Xu, Z., Panduri, V., Sandler, D. P., & Taylor, J. A. (2014). CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the sister study. *Environ Health Perspect*, *122*(7), 673-678. doi: 10.1289/ehp.1307480
- Harper, K. N., Peters, B. A., & Gamble, M. V. (2013). Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiol Biomarkers Prev*, *22*(6), 1052-1060. doi: 10.1158/1055-9965.EPI-13-0114
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., . . . Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, *22*(9), 1760-1774. doi: 10.1101/gr.135350.111
- Harvey, B. G., Heguy, A., Leopold, P. L., Carolan, B. J., Ferris, B., & Crystal, R. G. (2007). Modification of gene expression of the small airway epithelium in response to cigarette smoking. *J Mol Med (Berl)*, *85*(1), 39-53. doi: 10.1007/s00109-006-0103-z
- Hashimoto, K., Oreffo, R. O., Gibson, M. B., Goldring, M. B., & Roach, H. I. (2009). DNA demethylation at specific CpG sites in the IL1B promoter in response to inflammatory cytokines in human articular chondrocytes. *Arthritis Rheum*, *60*(11), 3303-3313. doi: 10.1002/art.24882
- Hasler, R., Feng, Z., Backdahl, L., Spehlmann, M. E., Franke, A., Teschendorff, A., . . . Rosenstiel, P. (2012). A functional methylome map of ulcerative colitis. *Genome Res*, *22*(11), 2130-2137. doi: 10.1101/gr.138347.112
- Heijmans, B. T. & Mill, J. (2012). Commentary: The seven plagues of epigenetic epidemiology. *Int J Epidemiol*, *41*(1), 74-78. doi: 10.1093/ije/dyr225
- Heijmans, B. T., Tobi, E. W., Stein, A. D., Putter, H., Blauw, G. J., Susser, E. S., . . . Lumey, L. H. (2008). Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A*, *105*(44), 17046-17049. doi: 10.1073/pnas.0806560105
- Hemberger, M., Dean, W., & Reik, W. (2009). Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nat Rev Mol Cell Biol*, *10*(8), 526-537. doi: 10.1038/nrm2727
- Henry, I., Bonaiti-Pellie, C., Chehensse, V., Beldjord, C., Schwartz, C., Utermann, G., & Junien, C. (1991). Uniparental paternal disomy in a genetic cancer-predisposing syndrome. *Nature*, *351*(6328), 665-667. doi: 10.1038/351665a0
- Herman, J. G., Merlo, A., Mao, L., Lapidus, R. G., Issa, J. P., Davidson, N. E., . . . Baylin, S. B. (1995). Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer Res*, *55*(20), 4525-4530.
- Hernandez, D. G., Nalls, M. A., Gibbs, J. R., Arepalli, S., van der Brug, M., Chong, S., . . . Singleton, A. B. (2011). Distinct DNA methylation changes highly correlated with chronological age in the human brain. *Hum Mol Genet*, *20*(6), 1164-1172. doi: 10.1093/hmg/ddq561
- Heyn, H., Carmona, F. J., Gomez, A., Ferreira, H. J., Bell, J. T., Sayols, S., . . . Esteller, M. (2013). DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker. *Carcinogenesis*, *34*(1), 102-108. doi: 10.1093/carcin/bgs321

- Heyn, H. & Esteller, M. (2012). DNA methylation profiling in the clinic: applications and challenges. *Nat Rev Genet*, 13(10), 679-692. doi: 10.1038/nrg3270
- Holden, P., Canty, E. G., Mortier, G. R., Zabel, B., Spranger, J., Carr, A., . . . Briggs, M. D. (1999). Identification of novel pro-alpha2(IX) collagen gene mutations in two families with distinctive oligo-epiphyseal forms of multiple epiphyseal dysplasia. *Am J Hum Genet*, 65(1), 31-38. doi: 10.1086/302440
- Holliday, R. (1994). Epigenetics: an overview. *Dev Genet*, 15(6), 453-457. doi: 10.1002/dvg.1020150602
- Holliday, R. & Pugh, J. E. (1975). DNA modification mechanisms and gene activity during development. *Science*, 187(4173), 226-232.
- Holmbeck, K., Bianco, P., Caterina, J., Yamada, S., Kromer, M., Kuznetsov, S. A., . . . Birkedal-Hansen, H. (1999). MT1-MMP-deficient mice develop dwarfism, osteopenia, arthritis, and connective tissue disease due to inadequate collagen turnover. *Cell*, 99(1), 81-92.
- Horikoshi, M., Yaghooskar, H., Mook-Kanamori, D. O., Sovio, U., Taal, H. R., Hennig, B. J., . . . Freathy, R. M. (2013). New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nat Genet*, 45(1), 76-82. doi: 10.1038/ng.2477
- Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol*, 14(10), R115. doi: 10.1186/gb-2013-14-10-r115
- Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., . . . Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13, 86. doi: 10.1186/1471-2105-13-86
- Howlett, S. K. & Reik, W. (1991). Methylation levels of maternal and paternal genomes during preimplantation development. *Development*, 113(1), 119-127.
- Hoyo, C., Fortner, K., Murtha, A. P., Schildkraut, J. M., Soubry, A., Demark-Wahnefried, W., . . . Murphy, S. K. (2012). Association of cord blood methylation fractions at imprinted insulin-like growth factor 2 (IGF2), plasma IGF2, and birth weight. *Cancer Causes Control*, 23(4), 635-645. doi: 10.1007/s10552-012-9932-y
- Huang, K. & Fan, G. (2010). DNA methylation in cell differentiation and reprogramming: an emerging systematic view. *Regen Med*, 5(4), 531-544. doi: 10.2217/rme.10.35
- Irizarry, R. A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S. A., Jeddloh, J. A., . . . Feinberg, A. P. (2008). Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res*, 18(5), 780-790. doi: 10.1101/gr.7301508
- Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., . . . Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet*, 41(2), 178-186. doi: 10.1038/ng.298
- Ito, K., Lim, S., Caramori, G., Chung, K. F., Barnes, P. J., & Adcock, I. M. (2001). Cigarette smoking reduces histone deacetylase 2 expression, enhances cytokine expression, and inhibits glucocorticoid actions in alveolar macrophages. *FASEB J*, 15(6), 1110-1112.
- Ivanovska, P. D. a. N. (2012). How Important are Innate Immunity Cells in Osteoarthritis Pathology, Principles of Osteoarthritis- Its Definition, Character, Derivation and Modality-Related Recognition, Dr. Bruce M. Rothschild (Ed.).
- Ivorra, C., Fraga, M. F., Bayon, G. F., Fernandez, A. F., Garcia-Vicent, C., Chaves, F. J., . . . Lurbe, E. (2015). DNA methylation patterns in newborns exposed to tobacco in utero. *J Transl Med*, 13, 25. doi: 10.1186/s12967-015-0384-5
- Jaenisch, R. & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*, 33 Suppl, 245-254. doi: 10.1038/ng1089

- Jarvelin, M. R., Sovio, U., King, V., Lauren, L., Xu, B., McCarthy, M. I., . . . Elliott, P. (2004). Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. *Hypertension*, *44*(6), 838-846. doi: 10.1161/01.HYP.0000148304.33869.ee
- Javierre, B. M., Fernandez, A. F., Richter, J., Al-Shahrour, F., Martin-Subero, J. I., Rodriguez-Ubreva, J., . . . Ballestar, E. (2010). Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res*, *20*(2), 170-179. doi: 10.1101/gr.100289.109
- Jim, J. J., Noponen-Hietala, N., Cheung, K. M., Ott, J., Karppinen, J., Sahraravand, A., . . . Chan, D. (2005). The TRP2 allele of COL9A2 is an age-dependent risk factor for the development and severity of intervertebral disc degeneration. *Spine (Phila Pa 1976)*, *30*(24), 2735-2742.
- Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V., & Jordan, I. K. (2012). On the presence and role of human gene-body DNA methylation. *Oncotarget*, *3*(4), 462-474.
- Johansson, S., Iliadou, A., Bergvall, N., de Faire, U., Kramer, M. S., Pawitan, Y., . . . Cnattingius, S. (2008). The association between low birth weight and type 2 diabetes: contribution of genetic factors. *Epidemiology*, *19*(5), 659-665.
- Jones, P. L., Veenstra, G. J., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., . . . Wolffe, A. P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet*, *19*(2), 187-191. doi: 10.1038/561
- Joubert, B. R., Haberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., . . . London, S. J. (2012). 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect*, *120*(10), 1425-1431. doi: 10.1289/ehp.1205412
- Kadonaga, J. T., Carner, K. R., Masiarz, F. R., & Tjian, R. (1987). Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. *Cell*, *51*(6), 1079-1090.
- Kaikkonen, M. U., Lam, M. T., & Glass, C. K. (2011). Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc Res*, *90*(3), 430-440. doi: 10.1093/cvr/cvr097
- Kaminsky, Z., Petronis, A., Wang, S. C., Levine, B., Ghaffar, O., Floden, D., & Feinstein, A. (2008). Epigenetics of personality traits: an illustrative study of identical twins discordant for risk-taking behavior. *Twin Res Hum Genet*, *11*(1), 1-11. doi: 10.1375/twin.11.1.1
- Kaminsky, Z. A., Tang, T., Wang, S. C., Ptak, C., Oh, G. H., Wong, A. H., . . . Petronis, A. (2009). DNA methylation profiles in monozygotic and dizygotic twins. *Nat Genet*, *41*(2), 240-245. doi: 10.1038/ng.286
- Kang, H. J., Kim, E. J., Kim, B. G., You, C. H., Lee, S. Y., Kim, D. I., & Hong, Y. S. (2012). Quantitative analysis of cancer-associated gene methylation connected to risk factors in Korean colorectal cancer patients. *J Prev Med Public Health*, *45*(4), 251-258. doi: 10.3961/jpmph.2012.45.4.251
- Kasai, A., Hiramatsu, N., Hayakawa, K., Yao, J., Maeda, S., & Kitamura, M. (2006). High levels of dioxin-like potential in cigarette smoke evidenced by in vitro and in vivo biosensing. *Cancer Res*, *66*(14), 7143-7150. doi: 10.1158/0008-5472.CAN-05-4541
- Kass, R. E. & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773-795.
- Kelsey, G. & Bartolomei, M. S. (2012). Imprinted genes ... and the number is? *PLoS Genet*, *8*(3), e1002601. doi: 10.1371/journal.pgen.1002601
- Kerkhof, H. J., Lories, R. J., Meulenbelt, I., Jonsdottir, I., Valdes, A. M., Arp, P., . . . van Meurs, J. B. (2010). A genome-wide association study identifies an osteoarthritis susceptibility locus on chromosome 7q22. *Arthritis Rheum*, *62*(2), 499-510. doi: 10.1002/art.27184

- Kleinman, C. L. & Majewski, J. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, 335(6074), 1302; author reply 1302. doi: 10.1126/science.1209658
- Koch, C. M., Suschek, C. V., Lin, Q., Bork, S., Goergens, M., Jousen, S., . . . Wagner, W. (2011). Specific age-associated DNA methylation changes in human dermal fibroblasts. *PLoS One*, 6(2), e16679. doi: 10.1371/journal.pone.0016679
- Koch, C. M. & Wagner, W. (2011). Epigenetic-aging-signature to determine age in different tissues. *Aging (Albany NY)*, 3(10), 1018-1027.
- Kwon, G. S., Viotti, M., & Hadjantonakis, A. K. (2008). The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Dev Cell*, 15(4), 509-520. doi: 10.1016/j.devcel.2008.07.017
- Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet*, 11(3), 191-203. doi: 10.1038/nrg2732
- Lee, J., Inoue, K., Ono, R., Ogonuki, N., Kohda, T., Kaneko-Ishino, T., . . . Ishino, F. (2002). Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells. *Development*, 129(8), 1807-1817.
- Lee, K. W., Richmond, R., Hu, P., French, L., Shin, J., Bourdon, C., . . . Pausova, Z. (2015). Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. *Environ Health Perspect*, 123(2), 193-199. doi: 10.1289/ehp.1408614
- Leek, J. T. & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9), 1724-1735. doi: 10.1371/journal.pgen.0030161
- Leeson, C. P., Kattenhorn, M., Morley, R., Lucas, A., & Deanfield, J. E. (2001). Impact of low birth weight and cardiovascular risk factors on endothelial function in early adult life. *Circulation*, 103(9), 1264-1268.
- Lewis, J. D., Meehan, R. R., Henzel, W. J., Maurer-Fogy, I., Jeppesen, P., Klein, F., & Bird, A. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*, 69(6), 905-914.
- Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754-1760. doi: 10.1093/bioinformatics/btp324
- Lin, W., Piskol, R., Tan, M. H., & Li, J. B. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, 335(6074), 1302; author reply 1302. doi: 10.1126/science.1210624
- Lin, Z., Pavlos, N. J., Cake, M. A., Wood, D. J., Xu, J., & Zheng, M. H. (2008). Evidence that human cartilage and chondrocytes do not express calcitonin receptor. *Osteoarthritis Cartilage*, 16(4), 450-457. doi: 10.1016/j.joca.2007.08.003
- Lister, R., Pelizzola, M., Downen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., . . . Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), 315-322. doi: 10.1038/nature08514
- Liu, Q., Liu, L., Zhao, Y., Zhang, J., Wang, D., Chen, J., . . . Liu, Z. (2011). Hypoxia induces genomic DNA demethylation through the activation of HIF-1alpha and transcriptional upregulation of MAT2A in hepatoma cells. *Mol Cancer Ther*, 10(6), 1113-1123. doi: 10.1158/1535-7163.MCT-10-1010
- Lokk, K., Modhukur, V., Rajashekar, B., Martens, K., Magi, R., Kolde, R., . . . Tonisson, N. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol*, 15(4), R54. doi: 10.1186/gb-2014-15-4-r54
- Lokk, K., Vooder, T., Kolde, R., Valk, K., Vosa, U., Roosipuu, R., . . . Tonisson, N. (2012). Methylation markers of early-stage non-small cell lung cancer. *PLoS One*, 7(6), e39813. doi: 10.1371/journal.pone.0039813



- Lyon, M. F. (1961). Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature*, *190*, 372-373.
- Maksimovic, J., Gordon, L., & Oshlack, A. (2012). SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol*, *13*(6), R44. doi: 10.1186/gb-2012-13-6-r44
- Markunas, C. A., Xu, Z., Harlid, S., Wade, P. A., Lie, R. T., Taylor, J. A., & Wilcox, A. J. (2014). Identification of DNA Methylation Changes in Newborns Related to Maternal Smoking during Pregnancy. *Environ Health Perspect*. doi: 10.1289/ehp.1307892
- Martino, D., Loke, Y. J., Gordon, L., Ollikainen, M., Cruickshank, M. N., Saffery, R., & Craig, J. M. (2013). Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. *Genome Biol*, *14*(5), R42. doi: 10.1186/gb-2013-14-5-r42
- Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, *7*, 29-59. doi: 10.1146/annurev.genom.7.080505.115623
- Mayer, W., Niveleau, A., Walter, J., Fundele, R., & Haaf, T. (2000). Demethylation of the zygotic paternal genome. *Nature*, *403*(6769), 501-502. doi: 10.1038/35000654
- McCormick, M. (1985). The Contribution of Low Birth Weight to Infant Mortality and Childhood Morbidity. *New England Journal of Medicine*, *312*, 82-90.
- McGrath, J. & Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, *37*(1), 179-183.
- McIntire, D. D., Bloom, S. L., Casey, B. M., & Leveno, K. J. (1999). Birth weight in relation to morbidity and mortality among newborn infants. *N Engl J Med*, *340*(16), 1234-1238. doi: 10.1056/NEJM199904223401603
- McLemore, T. L., Adelberg, S., Liu, M. C., McMahon, N. A., Yu, S. J., Hubbard, W. C., . . . et al. (1990). Expression of CYP1A1 gene in patients with lung cancer: evidence for cigarette smoke-induced gene expression in normal lung tissue and for altered gene regulation in primary pulmonary carcinomas. *J Natl Cancer Inst*, *82*(16), 1333-1339.
- Mensaert, K., Denil, S., Trooskens, G., Van Criekinge, W., Thas, O., & De Meyer, T. (2014). Next-generation technologies and data analytical approaches for epigenomics. *Environ Mol Mutagen*, *55*(3), 155-170. doi: 10.1002/em.21841
- Mercer, B. A., Wallace, A. M., Brinckerhoff, C. E., & D'Armiento, J. M. (2009). Identification of a cigarette smoke-responsive region in the distal MMP-1 promoter. *Am J Respir Cell Mol Biol*, *40*(1), 4-12. doi: 10.1165/rcmb.2007-0310OC
- Merlo, A., Herman, J. G., Mao, L., Lee, D. J., Gabrielson, E., Burger, P. C., . . . Sidransky, D. (1995). 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat Med*, *1*(7), 686-692.
- Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., . . . Kent, W. J. (2013). The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res*, *41*(Database issue), D64-69. doi: 10.1093/nar/gks1048
- Miltenberger, R. J., Mynatt, R. L., Wilkinson, J. E., & Woychik, R. P. (1997). The role of the agouti gene in the yellow obese syndrome. *J Nutr*, *127*(9), 1902S-1907S.
- Moayyeri, A., Hammond, C. J., Valdes, A. M., & Spector, T. D. (2013). Cohort Profile: TwinsUK and healthy ageing twin study. *Int J Epidemiol*, *42*(1), 76-85. doi: 10.1093/ije/dyr207
- Monick, M. M., Beach, S. R., Plume, J., Sears, R., Gerrard, M., Brody, G. H., & Philibert, R. A. (2012). Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. *Am J Med Genet B Neuropsychiatr Genet*, *159B*(2), 141-151. doi: 10.1002/ajmg.b.32021
- Morgan, H. D., Santos, F., Green, K., Dean, W., & Reik, W. (2005). Epigenetic reprogramming in mammals. *Hum Mol Genet*, *14 Spec No 1*, R47-58. doi: 10.1093/hmg/ddi114

- Morgan, H. D., Sutherland, H. G., Martin, D. I., & Whitelaw, E. (1999). Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet*, 23(3), 314-318. doi: 10.1038/15490
- Morris, T. J., Butcher, L. M., Feber, A., Teschendorff, A. E., Chakravarthy, A. R., Wojdacz, T. K., & Beck, S. (2014). ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics*, 30(3), 428-430. doi: 10.1093/bioinformatics/btt684
- Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M. C., & Leonhardt, H. (2005). Recruitment of DNA methyltransferase I to DNA repair sites. *Proc Natl Acad Sci U S A*, 102(25), 8905-8909. doi: 10.1073/pnas.0501034102
- Mulligan, C. J., D'Errico, N. C., Stees, J., & Hughes, D. A. (2012). Methylation changes at NR3C1 in newborns associate with maternal prenatal stress exposure and newborn birth weight. *Epigenetics*, 7(8), 853-857. doi: 10.4161/epi.21180
- Murphy, R., Ibanez, L., Hattersley, A., & Tost, J. (2012). IGF2/H19 hypomethylation in a patient with very low birthweight, precocious pubarche and insulin resistance. *BMC Med Genet*, 13, 42. doi: 10.1186/1471-2350-13-42
- Mustafa, Z., Chapman, K., Irven, C., Carr, A. J., Clipsham, K., Chitnavis, J., . . . Loughlin, J. (2000). Linkage analysis of candidate genes as susceptibility loci for osteoarthritis-suggestive linkage of COL9A1 to female hip osteoarthritis. *Rheumatology (Oxford)*, 39(3), 299-306.
- Naeem, H., Wong, N. C., Chatterton, Z., Hong, M. K., Pedersen, J. S., Corcoran, N. M., . . . Macintyre, G. (2014). Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics*, 15, 51. doi: 10.1186/1471-2164-15-51
- Nakajima, M., Takahashi, A., Kou, I., Rodriguez-Fontenla, C., Gomez-Reino, J. J., Furuichi, T., . . . Ikegawa, S. (2010). New sequence variants in HLA class II/III region associated with susceptibility to knee osteoarthritis identified by genome-wide association study. *PLoS One*, 5(3), e9723. doi: 10.1371/journal.pone.0009723
- Nakki, A., Videman, T., Kujala, U. M., Suhonen, M., Mannikko, M., Peltonen, L., . . . Saarela, J. (2011). Candidate gene association study of magnetic resonance imaging-based hip osteoarthritis (OA): evidence for COL9A2 gene as a common predisposing factor for hip OA and lumbar disc degeneration. *J Rheumatol*, 38(4), 747-752. doi: 10.3899/jrheum.100080
- Nan, X., Campoy, F. J., & Bird, A. (1997). MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell*, 88(4), 471-481.
- Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., & Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, 393(6683), 386-389. doi: 10.1038/30764
- Nardone, S., Sams, D. S., Reuveni, E., Getselter, D., Oron, O., Karpuj, M., & Elliott, E. (2014). DNA methylation analysis of the autistic brain reveals multiple dysregulated biological pathways. *Transl Psychiatry*, 4, e433. doi: 10.1038/tp.2014.70
- Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., . . . Spector, T. D. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet*, 7(2), e1002003. doi: 10.1371/journal.pgen.1002003
- Nicholls, R. D., Knoll, J. H., Butler, M. G., Karam, S., & Lalande, M. (1989). Genetic imprinting suggested by maternal heterodisomy in nondeletion Prader-Willi syndrome. *Nature*, 342(6247), 281-285. doi: 10.1038/342281a0
- Numata, S., Ye, T., Hyde, T. M., Guitart-Navarro, X., Tao, R., Wininger, M., . . . Lipska, B. K. (2012). DNA methylation signatures in development and aging of the human prefrontal cortex. *Am J Hum Genet*, 90(2), 260-272. doi: 10.1016/j.ajhg.2011.12.020
- Ohlsson, R., Renkawitz, R., & Lobanenkov, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet*, 17(9), 520-527.

- Ohno, S., Kaplan, W. D., & Kinosita, R. (1959). Formation of the sex chromatin by a single X-chromosome in liver cells of *Rattus norvegicus*. *Exp Cell Res*, *18*, 415-418.
- Ollikainen, M., Smith, K. R., Joo, E. J., Ng, H. K., Andronikos, R., Novakovic, B., . . . Craig, J. M. (2010). DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Hum Mol Genet*, *19*(21), 4176-4188. doi: 10.1093/hmg/ddq336
- Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., . . . Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. *Curr Biol*, *10*(8), 475-478.
- Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K., & Gilad, Y. (2011). A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet*, *7*(2), e1001316. doi: 10.1371/journal.pgen.1001316
- Palmer, S. & Chen, Y. H. (2008). Bcl-3, a multifaceted modulator of NF-kappaB-mediated gene transcription. *Immunol Res*, *42*(1-3), 210-218. doi: 10.1007/s12026-008-8075-4
- Petronis, A., Gottesman, II, Kan, P., Kennedy, J. L., Basile, V. S., Paterson, A. D., & Pependikyte, V. (2003). Monozygotic twins exhibit numerous epigenetic differences: clues to twin discordance? *Schizophr Bull*, *29*(1), 169-178.
- Pfeifer, G. P., Denissenko, M. F., Olivier, M., Tretyakova, N., Hecht, S. S., & Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*, *21*(48), 7435-7451. doi: 10.1038/sj.onc.1205803
- Philibert, R. A., Beach, S. R., Lei, M. K., & Brody, G. H. (2013). Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. *Clin Epigenetics*, *5*(1), 19. doi: 10.1186/1868-7083-5-19
- Phillips, D. I., Davies, M. J., & Robinson, J. S. (2001). Fetal growth and the fetal origins hypothesis in twins--problems and perspectives. *Twin Res*, *4*(5), 327-331. doi: 10.1375/1369052012669
- Pickrell, J. K., Gilad, Y., & Pritchard, J. K. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, *335*(6074), 1302; author reply 1302. doi: 10.1126/science.1210484
- Pidsley, R., CC, Y. W., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, *14*, 293. doi: 10.1186/1471-2164-14-293
- Ponting, C. P., Oliver, P. L., & Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell*, *136*(4), 629-641. doi: 10.1016/j.cell.2009.02.006
- Portela, A. & Esteller, M. (2010). Epigenetic modifications and human disease. *Nat Biotechnol*, *28*(10), 1057-1068. doi: 10.1038/nbt.1685
- Poschl, E., Fidler, A., Schmidt, B., Kallipolitou, A., Schmid, E., & Aigner, T. (2005). DNA methylation is not likely to be responsible for aggrecan down regulation in aged or osteoarthritic cartilage. *Ann Rheum Dis*, *64*(3), 477-480. doi: 10.1136/ard.2004.022509
- Price, M. E., Cotton, A. M., Lam, L. L., Farre, P., Emberly, E., Brown, C. J., . . . Kober, M. S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin*, *6*(1), 4. doi: 10.1186/1756-8935-6-4
- Qiu, J. (2006). Epigenetics: unfinished symphony. *Nature*, *441*(7090), 143-145. doi: 10.1038/441143a
- Rakyan, V. K., Beyan, H., Down, T. A., Hawa, M. I., Maslau, S., Aden, D., . . . Leslie, R. D. (2011). Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genet*, *7*(9), e1002300. doi: 10.1371/journal.pgen.1002300

- Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nat Rev Genet*, *12*(8), 529-541. doi: 10.1038/nrg3000
- Rakyan, V. K., Down, T. A., Maslau, S., Andrew, T., Yang, T. P., Beyan, H., . . . Spector, T. D. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res*, *20*(4), 434-439. doi: 10.1101/gr.103101.109
- Rasmussen, K. M. (2001). The "fetal origins" hypothesis: challenges and opportunities for maternal and child nutrition. *Annu Rev Nutr*, *21*, 73-95. doi: 10.1146/annurev.nutr.21.1.73
- Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, *447*(7143), 425-432. doi: 10.1038/nature05918
- Reik, W., Dean, W., & Walter, J. (2001). Epigenetic reprogramming in mammalian development. *Science*, *293*(5532), 1089-1093. doi: 10.1126/science.1063443
- Reik, W. & Walter, J. (2001a). Evolution of imprinting mechanisms: the battle of the sexes begins in the zygote. *Nat Genet*, *27*(3), 255-256. doi: 10.1038/85804
- Reik, W. & Walter, J. (2001b). Genomic imprinting: parental influence on the genome. *Nat Rev Genet*, *2*(1), 21-32. doi: 10.1038/35047554
- Relton, C. L. & Davey Smith, G. (2010). Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med*, *7*(10), e1000356. doi: 10.1371/journal.pmed.1000356
- Richards, A. J., Yates, J. R., Williams, R., Payne, S. J., Pope, F. M., Scott, J. D., & Snead, M. P. (1996). A family with Stickler syndrome type 2 has a mutation in the COL11A1 gene resulting in the substitution of glycine 97 by valine in alpha 1 (XI) collagen. *Hum Mol Genet*, *5*(9), 1339-1343.
- Richmond, R. C., Simpkin, A. J., Woodward, G., Gaunt, T. R., Lyttleton, O., McArdle, W. L., . . . Relton, C. L. (2015). Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum Mol Genet*, *24*(8), 2201-2217. doi: 10.1093/hmg/ddu739
- Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet*, *14*(1), 9-25.
- Risch, N. (1990a). Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet*, *46*(2), 222-228.
- Risch, N. (1990b). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am J Hum Genet*, *46*(2), 229-241.
- Risch, N. (1990c). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet*, *46*(2), 242-253.
- Risnes, K. R., Vatten, L. J., Baker, J. L., Jameson, K., Sovio, U., Kajantie, E., . . . Bracken, M. B. (2011). Birthweight and mortality in adulthood: a systematic review and meta-analysis. *Int J Epidemiol*, *40*(3), 647-661. doi: 10.1093/ije/dyq267
- Roach, H. I., Yamada, N., Cheung, K. S., Tilley, S., Clarke, N. M., Oreffo, R. O., . . . Bronner, F. (2005). Association between the abnormal expression of matrix-degrading enzymes by human osteoarthritic chondrocytes and demethylation of specific CpG sites in the promoter regions. *Arthritis Rheum*, *52*(10), 3110-3124. doi: 10.1002/art.21300
- Rosenbloom, K. R., Sloan, C. A., Malladi, V. S., Dreszer, T. R., Learned, K., Kirkup, V. M., . . . Kent, W. J. (2013). ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res*, *41*(Database issue), D56-63. doi: 10.1093/nar/gks1172
- Santos, F. & Dean, W. (2004). Epigenetic reprogramming during early development in mammals. *Reproduction*, *127*(6), 643-651. doi: 10.1530/rep.1.00221

- Satta, R., Maloku, E., Zhubi, A., Pibiri, F., Hajos, M., Costa, E., & Guidotti, A. (2008). Nicotine decreases DNA methyltransferase 1 expression and glutamic acid decarboxylase 67 promoter methylation in GABAergic interneurons. *Proc Natl Acad Sci U S A*, *105*(42), 16356-16361. doi: 10.1073/pnas.0808699105
- Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*, *103*(5), 1412-1417. doi: 10.1073/pnas.0510310103
- Schembri, F., Sridhar, S., Perdomo, C., Gustafson, A. M., Zhang, X., Ergun, A., . . . Spira, A. (2009). MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium. *Proc Natl Acad Sci U S A*, *106*(7), 2319-2324. doi: 10.1073/pnas.0806383106
- Schilling, E., El Chartouni, C., & Rehli, M. (2009). Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. *Genome Res*, *19*(11), 2028-2035. doi: 10.1101/gr.095562.109
- Schubeler, D. (2012). Molecular biology. Epigenetic islands in a genetic ocean. *Science*, *338*(6108), 756-757. doi: 10.1126/science.1227243
- Segovia-Silvestre, T., Bonnefond, C., Sondergaard, B. C., Christensen, T., Karsdal, M. A., & Bay-Jensen, A. C. (2011). Identification of the calcitonin receptor in osteoarthritic chondrocytes. *BMC Res Notes*, *4*, 407. doi: 10.1186/1756-0500-4-407
- Seki, Y., Williams, L., Vuguin, P. M., & Charron, M. J. (2012). Minireview: Epigenetic programming of diabetes and obesity: animal models. *Endocrinology*, *153*(3), 1031-1038. doi: 10.1210/en.2011-1805
- Selamat, S. A., Chung, B. S., Girard, L., Zhang, W., Zhang, Y., Campan, M., . . . Laird-Offringa, I. A. (2012). Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res*, *22*(7), 1197-1211. doi: 10.1101/gr.132662.111
- Shenker, N. S., Ueland, P. M., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., . . . Vineis, P. (2013). DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology*, *24*(5), 712-716. doi: 10.1097/EDE.0b013e31829d5cb3
- Siedlinski, M., Klanderma, B., Sandhaus, R. A., Barker, A. F., Brantly, M. L., Eden, E., . . . Demeo, D. L. (2012). Association of cigarette smoking and CRP levels with DNA methylation in alpha-1 antitrypsin deficiency. *Epigenetics*, *7*(7), 720-728. doi: 10.4161/epi.20319
- Skidmore, P. M., Hardy, R. J., Kuh, D. J., Langenberg, C., & Wadsworth, M. E. (2004). Birth weight and lipids in a national birth cohort study. *Arterioscler Thromb Vasc Biol*, *24*(3), 588-594. doi: 10.1161/01.ATV.0000116692.85043.ef
- Slatkin, M. (2009). Epigenetic inheritance and the missing heritability problem. *Genetics*, *182*(3), 845-850. doi: 10.1534/genetics.109.102798
- Snieder, H. W., X., MacGregor, A. J. (2010). Twin Methodology.
- Souren, N. Y., Lutsik, P., Gasparoni, G., Tierling, S., Gries, J., Riemenschneider, M., . . . Walter, J. (2013). Adult monozygotic twins discordant for intra-uterine growth have indistinguishable genome-wide DNA methylation profiles. *Genome Biol*, *14*(5), R44. doi: 10.1186/gb-2013-14-5-r44
- Spector, T. D., Cicuttini, F., Baker, J., Loughlin, J., & Hart, D. (1996). Genetic influences on osteoarthritis in women: a twin study. *BMJ*, *312*(7036), 940-943.
- Spector, T. D. & MacGregor, A. J. (2004). Risk factors for osteoarthritis: genetics. *Osteoarthritis Cartilage*, *12 Suppl A*, S39-44.
- Steegers-Theunissen, R. P., Obermann-Borst, S. A., Kremer, D., Lindemans, J., Siebel, C., Steegers, E. A., . . . Heijmans, B. T. (2009). Periconceptional maternal folic acid use of 400 microg per day is related to increased methylation of the IGF2 gene in the very young child. *PLoS One*, *4*(11), e7845. doi: 10.1371/journal.pone.0007845

- Stein, R. A. (2012). Epigenetics and environmental exposures. *J Epidemiol Community Health*, 66(1), 8-13. doi: 10.1136/jech.2010.130690
- Storey, J. D. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16), 9440-9445. doi: 10.1073/pnas.1530509100
- Stower, H. (2014). Epigenetics: Reprogramming with TET. *Nat Rev Genet*, 15(2), 66. doi: 10.1038/nrg3659
- Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., . . . Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. *Nat Struct Mol Biol*, 16(5), 564-571. doi: 10.1038/nsmb.1594
- Sun, Y. V., Smith, A. K., Conneely, K. N., Chang, Q., Li, W., Lazarus, A., . . . Kardia, S. L. (2013). Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. *Hum Genet*, 132(9), 1027-1037. doi: 10.1007/s00439-013-1311-6
- Surani, M. A., Barton, S. C., & Norris, M. L. (1984). Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, 308(5959), 548-550.
- Suter, M., Ma, J., Harris, A., Patterson, L., Brown, K. A., Shope, C., . . . Aagaard-Tillery, K. M. (2011). Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. *Epigenetics*, 6(11), 1284-1294. doi: 10.4161/epi.6.11.17819
- Suter, M. A., Anders, A. M., & Aagaard, K. M. (2013). Maternal smoking as a model for environmental epigenetic changes affecting birthweight and fetal programming. *Mol Hum Reprod*, 19(1), 1-6. doi: 10.1093/molehr/gas050
- Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., & Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 29(2), 189-196. doi: 10.1093/bioinformatics/bts680
- Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., . . . Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res*, 20(4), 440-446. doi: 10.1101/gr.103606.109
- Thompson, C., Syddall, H., Rodin, I., Osmond, C., & Barker, D. J. (2001). Birth weight and the risk of depressive disorder in late life. *Br J Psychiatry*, 179, 450-455.
- Thun, M. J., DeLancey, J. O., Center, M. M., Jemal, A., & Ward, E. M. (2010). The global burden of cancer: priorities for prevention. *Carcinogenesis*, 31(1), 100-110. doi: 10.1093/carcin/bgp263
- Tobi, E. W., Lumey, L. H., Talens, R. P., Kremer, D., Putter, H., Stein, A. D., . . . Heijmans, B. T. (2009). DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum Mol Genet*, 18(21), 4046-4053. doi: 10.1093/hmg/ddp353
- Torche, F. & Echevarria, G. (2011). The effect of birthweight on childhood cognitive development in a middle-income country. *Int J Epidemiol*, 40(4), 1008-1018. doi: 10.1093/ije/dyr030
- Touleimat, N. & Tost, J. (2012). Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3), 325-341. doi: 10.2217/epi.12.21
- Tsai, P. C. & Bell, J. T. (2015). Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int J Epidemiol*. doi: 10.1093/ije/dyv041
- Tsai, P. C., Spector, T. D., & Bell, J. T. (2012). Using epigenome-wide association scans of DNA methylation in age-related complex human traits. *Epigenomics*, 4(5), 511-526. doi: 10.2217/epi.12.45

- Tsai, P. C., Van Dongen, J., Tan, Q., Willemsen, G., Christiansen, L., Boomsma, D. I., . . . Bell, J. T. (2015). DNA methylation changes in the IGF1R gene in birth weight discordant adult monozygotic twins. *Twin Res Hum Genet*, (*In press*).
- Tsaprouni, L. G., Yang, T. P., Bell, J., Dick, K. J., Kanoni, S., Nisbet, J., . . . Deloukas, P. (2014). Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics*, *9*(10), 1382-1396. doi: 10.4161/15592294.2014.969637
- Tserel, L., Limbach, M., Saare, M., Kisand, K., Metspalu, A., Milani, L., & Peterson, P. (2014). CpG sites associated with NRP1, NRXN2 and miR-29b-2 are hypomethylated in monocytes during ageing. *Immun Ageing*, *11*(1), 1. doi: 10.1186/1742-4933-11-1
- Tucker, K. L., Beard, C., Dausmann, J., Jackson-Grusby, L., Laird, P. W., Lei, H., . . . Jaenisch, R. (1996). Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. *Genes Dev*, *10*(8), 1008-1020.
- Turan, N., Ghalwash, M. F., Katari, S., Coutifaris, C., Obradovic, Z., & Sapienza, C. (2012). DNA methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease? *BMC Med Genomics*, *5*, 10. doi: 10.1186/1755-8794-5-10
- Valdes, A. M., Loughlin, J., Timms, K. M., van Meurs, J. J., Southam, L., Wilson, S. G., . . . Spector, T. D. (2008). Genome-wide association scan identifies a prostaglandin-endoperoxide synthase 2 variant involved in risk of knee osteoarthritis. *Am J Hum Genet*, *82*(6), 1231-1240. doi: 10.1016/j.ajhg.2008.04.006
- Valdes, A. M. & Spector, T. D. (2011). Genetic epidemiology of hip and knee osteoarthritis. *Nat Rev Rheumatol*, *7*(1), 23-32. doi: 10.1038/nrrheum.2010.191
- Valdes, A. M., Styrkarsdottir, U., Doherty, M., Morris, D. L., Mangino, M., Tamm, A., . . . Arden, N. K. (2011). Large scale replication study of the association between HLA class II/BTNL2 variants and osteoarthritis of the knee in European-descent populations. *PLoS One*, *6*(8), e23371. doi: 10.1371/journal.pone.0023371
- van Dongen, J., Slagboom, P. E., Draisma, H. H., Martin, N. G., & Boomsma, D. I. (2012). The continuing value of twin studies in the omics era. *Nat Rev Genet*, *13*(9), 640-653. doi: 10.1038/nrg3243
- van Leeuwen, D. M., van Aagen, E., Gottschalk, R. W., Vlietinck, R., Gielen, M., van Herwijnen, M. H., . . . van Delft, J. H. (2007). Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs. *Carcinogenesis*, *28*(3), 691-697. doi: 10.1093/carcin/bgl199
- van Mourik, J. B., Hamel, B. C., & Mariman, E. C. (1998). A large family with multiple epiphyseal dysplasia linked to COL9A2 gene. *Am J Med Genet*, *77*(3), 234-240.
- Vikkula, M., Mariman, E. C., Lui, V. C., Zhidkova, N. I., Tiller, G. E., Goldring, M. B., . . . et al. (1995). Autosomal dominant and recessive osteochondrodysplasias associated with the COL11A2 locus. *Cell*, *80*(3), 431-437.
- Waddington, C. H. (1942). *The epigenotype*: Endeavour, *1*, 18-20.
- Waddington, C. H. (1957). *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*.: London: George Allen & Unwin.
- Walter, E. C., Ehlenbach, W. J., Hotchkin, D. L., Chien, J. W., & Koepsell, T. D. (2009). Low birth weight and respiratory disease in adulthood: a population-based case-control study. *Am J Respir Crit Care Med*, *180*(2), 176-180. doi: 10.1164/rccm.200901-0046OC
- Wan, E. S., Qiu, W., Baccarelli, A., Carey, V. J., Bacherman, H., Rennard, S. I., . . . Demeo, D. L. (2012). Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum Mol Genet*, *21*(13), 3073-3082. doi: 10.1093/hmg/dds135

- Wang, P., Guan, P. P., Guo, C., Zhu, F., Konstantopoulos, K., & Wang, Z. Y. (2013). Fluid shear stress-induced osteoarthritis: roles of cyclooxygenase-2 and its metabolic products in inducing the expression of proinflammatory cytokines and matrix metalloproteinases. *FASEB J*. doi: 10.1096/fj.13-234542
- Wang, S. (2011). Method to detect differentially methylated loci with case-control designs using Illumina arrays. *Genet Epidemiol*, 35(7), 686-694. doi: 10.1002/gepi.20619
- Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., . . . Yuan, Y. C. (2013). COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Res*, 41(11), e117. doi: 10.1093/nar/gkt242
- Waterland, R. A. & Jirtle, R. L. (2003). Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol Cell Biol*, 23(15), 5293-5300.
- Waterland, R. A., Kellermayer, R., Laritsky, E., Rayco-Solon, P., Harris, R. A., Travisano, M., . . . Prentice, A. M. (2010). Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. *PLoS Genet*, 6(12), e1001252. doi: 10.1371/journal.pgen.1001252
- Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Paabo, S., Rebhan, M., & Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*, 39(4), 457-466. doi: 10.1038/ng1990
- Whitelaw, N. C., Chong, S., & Whitelaw, E. (2010). Tuning in to noise: epigenetics and intangible variation. *Dev Cell*, 19(5), 649-650. doi: 10.1016/j.devcel.2010.11.001
- Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17), 2190-2191. doi: 10.1093/bioinformatics/btq340
- Williams, S. & Poulton, R. (1999). Twins and maternal smoking: ordeals for the fetal origins hypothesis? A cohort study. *BMJ*, 318(7188), 897-900.
- Woenckhaus, M., Klein-Hitpass, L., Grepmeier, U., Merk, J., Pfeifer, M., Wild, P., . . . Dietmaier, W. (2006). Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers. *J Pathol*, 210(2), 192-204. doi: 10.1002/path.2039
- Wojcik, W., Lee, W., Colman, I., Hardy, R., & Hotopf, M. (2013). Foetal origins of depression? A systematic review and meta-analysis of low birth weight and later depression. *Psychol Med*, 43(1), 1-12. doi: 10.1017/S0033291712000682
- Wolff, G. L., Kodell, R. L., Moore, S. R., & Cooney, C. A. (1998). Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. *FASEB J*, 12(11), 949-957.
- Wong, C. C., Caspi, A., Williams, B., Craig, I. W., Houts, R., Ambler, A., . . . Mill, J. (2010). A longitudinal study of epigenetic variation in twins. *Epigenetics*, 5(6), 516-526.
- Wu, A. H., McKean-Cowdin, R., & Tseng, C. C. (2011). Birth weight and other prenatal factors and risk of breast cancer in Asian-Americans. *Breast Cancer Res Treat*, 130(3), 917-925. doi: 10.1007/s10549-011-1640-x
- Wu, Y. W., Xing, G., Fuentes-Afflick, E., Danielson, B., Smith, L. H., & Gilbert, W. M. (2011). Racial, ethnic, and socioeconomic disparities in the prevalence of cerebral palsy. *Pediatrics*, 127(3), e674-681. doi: 10.1542/peds.2010-1656
- Xue, F., Willett, W. C., Rosner, B. A., Hankinson, S. E., & Michels, K. B. (2011). Cigarette smoking and the incidence of breast cancer. *Arch Intern Med*, 171(2), 125-133. doi: 10.1001/archinternmed.2010.503
- Yae, T., Tsuchihashi, K., Ishimoto, T., Motohara, T., Yoshikawa, M., Yoshida, G. J., . . . Nagano, O. (2012). Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell. *Nat Commun*, 3, 883. doi: 10.1038/ncomms1892
- Yamazaki, Y., Mann, M. R., Lee, S. S., Marh, J., McCarrey, J. R., Yanagimachi, R., & Bartolomei, M. S. (2003). Reprogramming of primordial germ cells begins before



- migration into the genital ridge, making these cells inadequate donors for reproductive cloning. *Proc Natl Acad Sci U S A*, 100(21), 12207-12212. doi: 10.1073/pnas.2035119100
- Yang, H. J., Qin, R., Katusic, S., & Juhn, Y. J. (2013). Population-based study on association between birth weight and risk of asthma: a propensity score approach. *Ann Allergy Asthma Immunol*, 110(1), 18-23. doi: 10.1016/j.anai.2012.10.010
- Zeggini, E., Panoutsopoulou, K., Southam, L., Rayner, N. W., Day-Williams, A. G., Lopes, M. C., . . . Loughlin, J. (2012). Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. *Lancet*, 380(9844), 815-823. doi: 10.1016/S0140-6736(12)60681-3
- Zeilinger, S., Kuhnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., . . . Illig, T. (2013). Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One*, 8(5), e63812. doi: 10.1371/journal.pone.0063812
- Zemach, A., McDaniel, I. E., Silva, P., & Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 328(5980), 916-919. doi: 10.1126/science.1186366
- Zhang, B., Xie, Q. G., Quan, Y., & Pan, X. M. (2013). Expression profiling based on graph-clustering approach to determine osteoarthritis related pathway. *Eur Rev Med Pharmacol Sci*, 17(15), 2097-2102.
- Zhang, H., Herman, A. I., Kranzler, H. R., Anton, R. F., Zhao, H., Zheng, W., & Gelernter, J. (2013). Array-based profiling of DNA methylation changes associated with alcohol dependence. *Alcohol Clin Exp Res*, 37 Suppl 1, E108-115. doi: 10.1111/j.1530-0277.2012.01928.x
- Zhang, Y., Yang, R., Burwinkel, B., Breitling, L. P., & Brenner, H. (2014). F2RL3 Methylation as a Biomarker of Current and Lifetime Smoking Exposures. *Environ Health Perspect*, 122(2), 131-137. doi: 10.1289/ehp.1306937
- Zhang, Y. G., Guo, X., Sun, Z., Jia, G., Xu, P., & Wang, S. (2010). Gene expression profiles of disc tissues and peripheral blood mononuclear cells from patients with degenerative discs. *J Bone Miner Metab*, 28(2), 209-219. doi: 10.1007/s00774-009-0120-4
- Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet*, 39(1), 61-69. doi: 10.1038/ng1929
- Ziller, M. J., Muller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., . . . Meissner, A. (2011). Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet*, 7(12), e1002389. doi: 10.1371/journal.pgen.1002389
- Zimmermann, P., Boeuf, S., Dickhut, A., Boehmer, S., Olek, S., & Richter, W. (2008). Correlation of COL10A1 induction during chondrogenesis of mesenchymal stem cells with demethylation of two CpG sites in the COL10A1 promoter. *Arthritis Rheum*, 58(9), 2743-2753. doi: 10.1002/art.23736
- Zou, J., Lippert, C., Heckerman, D., Aryee, M., & Listgarten, J. (2014). Epigenome-wide association studies without the need for cell-type composition. *Nat Methods*. doi: 10.1038/nmeth.2815
- Zudaire, E., Cuesta, N., Murty, V., Woodson, K., Adams, L., Gonzalez, N., . . . Cuttitta, F. (2008). The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers. *J Clin Invest*, 118(2), 640-650. doi: 10.1172/JCI30024
- Zupan, J., Komadina, R., & Marc, J. (2012). The relationship between osteoclastogenic and anti-osteoclastogenic pro-inflammatory cytokines differs in human osteoporotic and osteoarthritic bone tissues. *J Biomed Sci*, 19, 28. doi: 10.1186/1423-0127-19-28
- Zykovich, A., Hubbard, A., Flynn, J. M., Tarnopolsky, M., Fraga, M. F., Kerksick, C., . . . Melov, S. (2014). Genome-wide DNA methylation changes with age in disease-free human skeletal muscle. *Aging Cell*, 13(2), 360-366. doi: 10.1111/accel.12180