Processing and analysis of DNA methylation and statistical integration with genetic and clinical data

Pan, Hong

2017

Pan, H. (2017). Processing and analysis of DNA methylation and statistical integration with genetic and clinical data. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/69957>

<https://doi.org/10.32657/10356/69957>

**PROCESSING AND ANALYSIS OF DNA
METHYLATION AND STATISTICAL
INTEGRATION WITH GENETIC AND CLINICAL
DATA**

PAN HONG

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

2017

Acknowledgement

I would like to thank Agency for Science, Technology and Research, Singapore (A*STAR, Singapore) for the scholarship of Science Staff Development Award (SSDA), and Singapore Institute for Clinical Science (SICS, A*STAR) for supporting me with flexible working time and computer resource.

I thank sincerely to my supervisor, Prof. Kwoh Chee Keong, for his great guidance and encouragement during my study. Prof Kwoh trained me to become a qualified researcher in computer engineering.

I own great appreciation to my supervisor and mentor in SICS, Dr. Joanna Dawn Holbrook. Joanna guided me into epigenetics research area from scratch and taught me to investigate biological data and interpret results. As my role model, she provided support and expertise for me to learn and grow as an individual scientist.

I also own great thanks to my mentor in SICS, Dr. Neerja Karnani, who provided me precious opportunities to study new technology and helped me to accomplish thesis at very tight time schedule.

I would like to thank all my collaborators in research and co-authors in publications, particular thanks to Chen Li, Teh Ai Ling and Lin Xinyi. They are the most prominent people amongst the many people who have taught me about their field of expertise through numerous insightful discussions.

I sincerely thank my beloved son and daughter, Chu Wei and Chu Tong for their love and confidence in me. Last but not least, I would like to thank Yan for love.

Abstract

Epigenetics is genetic regulation that is not directly encoded in the DNA sequence. DNA methylation is an important epigenetic mark and its levels at particular loci have been associated with health and disease states. Better understanding of DNA methylation may lead to development of improved therapeutics for these diseases. Recent technical advances such as high-throughput arrays and next generation sequencing have enabled discovery of epigenetic correlates of clinical phenotypes encompassing hundreds of clinical samples at genome wide level with single base resolution.

Epigenome wide data on large number of samples has necessitated bio-computing with efficient analysis of well-designed experiments, new data analytic pipelines, mathematical models and biological interpretation. This thesis addresses three major challenges. First, it is important to optimize the detection of epigenetic variation and minimize the impact of technical artefacts for a reliable analysis. Second, algorithms must be very efficient because multivariate statistical models are essential in association studies and they require high computational power when applied to genome-wide studies. Third, computational methods should facilitate biological discovery through large scaled data analyses.

Corresponding to the challenges in epigenetics, three major contributions are presented in this thesis. First, a new pipeline has been developed to process and remove technical artefacts in genome-wide DNA methylation data generated by the most commonly used Illumina Infinium HumanMethylation 450K array. Compared to pre-existing algorithms, data processed through this pipeline was in a better agreement with the data obtained from reduced representation bisulfite sequencing on the same clinical samples. This study was further extended to evaluate the emerging next generation sequencing technology, Methyl Capture Sequencing, in buccal clinical samples. This thesis provides a comprehensive comparison of array and sequencing data, across key functional genomic regions in terms of their coverage and concordance of methylation calls and the use in epigenomic wide analysis study.

The second part presents a suite of statistical models developed to study the complicated relationships between Gene, Environment and Methylation. Three major functions GEM_Emodel, GEM_Gmodel and GEM_GxEmodel have been developed

into an R package named GEM. Using matrix based iterative correlation and memory-efficient data analysis, GEM facilitates reliable millions of associations between DNA methylation, genetic variants and environmental factors within minutes, in a standard computational setting. GEM has been validated by comprehensive benchmarking and has now become a part of Bioconductor, an extensively used open source bioinformatics suite.

Lastly, GEM was employed to study the DNA methylation and its integration with genetic variants and environmental influences of multi-ethnicity Asian neonates from a Singapore based birth cohort (Growing Up in Singapore Towards healthy Outcomes, GUSTO) and discover methylation changes associated with sub-optimal health outcomes in early life. In an analysis of 237 GUSTO neonatal methylomes, we found methylation quantitative trait loci were readily detected and the best explanation for 75% of the most variably methylated regions was due to the interaction of genotype with *in utero* environments. This study shed new light on the complex relationship between biological inheritance and individual prenatal experience suggesting the importance of considering both genetic variation and environmental factors in interpreting epigenetic variation. The GEM models were also applicable in finding that *HIF3A* DNA methylation measured in the umbilical cord of 991 newborns can aid understanding the genesis of adiposity at birth.

Author's publications related to this thesis

1. Pan H, Chen L, Dogra S, Teh AL, Tan JH, Lim YI, Lim YC, Jin S, Lee YK, Ng PY, Ong ML, Barton S, Chong YS, Meaney MJ, Gluckman PD, Stunkel W, Ding C, Holbrook JD. "Measuring the methylome in clinical samples: improved processing of the Infinium Human Methylation450 BeadChip Array". *Epigenetics*. 2012, Oct;7(10):1173-87 [1] (IF=4.78, citations=25).
2. Teh AL*, Pan H*, Lin X*, Lim YI, Patro CP, Cheong CY, Gong M, MacIsaac JL, Kwok CK, Meaney MJ *et al*: Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. *Epigenetics* 2016, 11(1):36-48 (* co-first author) [2] (IF=4.78, citation=2).
3. Pan H, Holbrook JD, Karnani N, Kwok CK. Gene, Environment and Methylation (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment. *BMC Bioinformatics*. 2016 Aug 217:299 [3] (IF=2.576)
4. Teh AL*, Pan H*, Chen L, Ong ML, Dogra S, Wong J, MacIsaac JL, Mah SM, McEwen LM, Saw SM *et al*: The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome Research* 2014, 24(7):1064-1074. (* co-first author) [4] (IF=14.63, citations=36).
5. Pan H*, Lin X*, Wu Y, Chen L, Teh AL, Soh SE, Lee YS, Tint MT, Yap, F., Saw, S.M., Kobor, M.S., Meaney, M.J., Godfrey, K.M., Chong, Y-S., Gluckman, P.D., Holbrook, J.D., on behalf of the GUSTO study group "Neonatal DNA methylation of HIF3A associates with birth weight and subsequent adiposity in Asian children". *Epigenomics* 2015 Sep 7(6):937-50. (* co-first author) [5] (IF=4.68, citations = 3).
6. Stünkel W, Pan H, Chew SB, Tng E, Tan JH, Chen L, Joseph R, Cheong CY, Ong ML, Lee YS, Chong YS, Saw SM, Meaney MJ, Kwek K, Sheppard AM, Gluckman PD; GUSTO Study Group, Holbrook JD. "Transcriptome changes affecting Hedgehog and cytokine signaling in the umbilical cord: implications for disease risk". *PLoS One*. 2012;7(7) [6] (IF=3.23, citations = 5).
7. Cheong CY, Chng K, Lim MK, Amrithraj AI, Joseph R, Sukarieh R, Tan YC, Chan L, Tan JH, Chen L, Pan H, Holbrook JD, Meaney MJ, Chong YS, Gluckman PD, Stünkel W. "Alterations to DNA methylation and expression of CXCL14 are associated with suboptimal birth outcomes". *Journal of Human Genetics*. 2014 Aug 7 [7] (IF=2.46, citations = 1).
8. Chen L, Pan H, Tuan TA, Teh AL, MacIsaac , Mah, SM, McEwan, L, Li, Y, Chen H, Broekman BFP, Buschdorf JP, Chong Y-S, Kwek K, Saw SM,

Gluckman PD, Fortier, MV, Rifkin-Graboi, A, Kobor, MS, Qiu, A, Meaney MJ, Holbrook, JD. on behalf of the GUSTO study group, “ Infant BDNF Val66Met influences the association of the DNA methylome with maternal anxiety and neonatal brain volumes”. *Development and Psychopathology*, 2015. Feb 27(1):137-50 [8] (IF=4.89, citations=4).

9. Huang RC, Garratt ES, Pan H, Wu Y, Daviv EA, Barton SJ, Burdge GC, Godfrey KM, Holbrook JH, Lillycrop KA “Genome wide methylation analysis identifies differentially methylated CpG loci associated with severe obesity in childhood”. *Epigenetics* 2015 Nov 2;10(11):995-1005[9] (IF=4.78, citations=1).

Conference Paper

10. Pan H, Lim M et al., Epigenetic Memory of Gestational Diabetes in the Infant, 8th World Congress on Developmental Origins of Health and Disease, Singapore 2013.
11. Chen L, Pan H et al., Epigenetic Memory of Maternal Mental Health in the Infant, 8th World Congress on Developmental Origins of Health and Disease, Singapore 2013.
12. Teh AL and Pan H. et al, Gene x Environment Shapes Neonate Methylomes, 2014 A*STAR Science Conference, Sep25-26, 2014, Singapore. (The best poster award).

(Information for journal impact (IF) is from Google. Information for citation is from NCBI PubMed by Jan 12, 2017.)

Abbreviation

<i>5mC</i>	5-methylcytosine
<i>5mhC</i>	5-hydroxymethylcytosine
<i>AIC</i>	Akaike information criterion
<i>ANOVA</i>	analysis of variance
<i>Autosomes</i>	The chromosomes which are not sex chromosomes. There are 22 pairs of autosomes in humans.
<i>beta value</i>	percentage methylation value (following a beta distribution)
<i>bp</i>	base pair
<i>BIC</i>	Bayesian Information Criterion
<i>cis-</i>	SNP and CpG locate in the same chromosome
<i>CpG</i>	Cytosine phosphate Guanine
<i>ENCODE</i>	Encyclopaedia of DNA Elements Consortium
<i>EWAS</i>	epigenome-wide association study
<i>FDR</i>	False discovery rate
<i>Genome</i>	All the DNA in the nucleus of a cell
<i>GEM</i>	R package, a tool suite for association study for Gene, Environment and Methylation, hosted in BioConductor via http://bioconductor.org/packages/GEM/
<i>GUSTO</i>	Growing Up in Singapore Towards healthy Outcomes, a Singapore based multi-ethnicity birth cohort
<i>HPC</i>	High-Performance Computing
<i>ICGC</i>	International Cancer Genome Consortium
<i>IHEC</i>	International Human Epigenome Consortium
<i>Infinium 27K</i>	Infinium Human Methylation 27 BeadChip Array, or

	27K
<i>Infinium 450K</i>	Infinium Human Methylation 450 BeadChip Array, or 450K in a short format
<i>Infinium 850K</i>	MethylationEPIC BeadChip Infinium, or 850K or EPIC
<i>kb</i>	Kilobase. 1000 base pairs
<i>Locus (plural loci)</i>	In genetics, a locus is the specific location or position of a gene's DNA sequence on a chromosome.
<i>M value</i>	logit transformation of beta value
<i>MC-Seq</i>	Methyl-Capture Sequencing. A next generation sequencing based technology with designed targets for DNA methylation detection.
<i>methQTL</i>	methylation Quantitative Trait Loci. The methylation variant at CpG is influenced by SNP either in <i>cis</i> - or <i>trans</i> - chromosome.
<i>MIC</i>	Maximal Information Coefficient
<i>Phenotype</i>	The observable characteristics or traits of an organism
<i>RRBS</i>	Reduced Representation Bisulfite Sequencing. A next generation sequencing based technology for DNA methylation detection
<i>Sex chromosomes</i>	The X and Y chromosomes that govern sex determination in mammals. Normally, females have two X chromosomes and males have one X and one Y chromosome.
<i>SD</i>	Standard Deviation
<i>SNP</i>	Single Nucleotide Polymorphism
<i>Somatic cells</i>	Cells of the body

<i>TF</i>	Transcription factor
<i>trans-</i>	SNP and CpG locate in different chromosomes
<i>WGBS</i>	Whole Genome Bisulfite Sequencing
<i>DNA</i>	Deoxyribonucleic Acid
<i>RNA</i>	Ribonucleic Acid
<i>PCA</i>	Principal Component Analysis
<i>PCR</i>	Polymerase Chain Reaction

Table of Contents

Acknowledgement	i
Abstract.....	ii
Author's publications related to this thesis	iv
Abbreviation	vi
Table of Contents	ix
1 Introduction.....	1
1.1 Objectives.....	1
1.2 State of the Art: Epigenetics and DNA methylation	2
1.3 DNA methylation measurements and Infinium450K technology	8
1.4 DNA methylation is associated with early experience.....	12
1.5 DNA methylation is influenced by genotype	13
1.6 Understanding methylation and modelling epigenome wide association study by GEM: opportunities and challenges from a computational perspective.....	15
1.7 GUSTO cohort and samples	17
1.8 Overview of the thesis.....	17
2 Improving the process of Infinium450K for measuring the methylome in clinical samples.....	20
2.1 Introduction.....	20
2.2 A novel pipeline to process the Infinium450K	21
2.2.1 Raw data extraction and quality control.....	22
2.2.2 Color adjustment	23

2.2.3	Type I and Type II correction	25
2.3	Concordance study between Infinium450K and RRBS in clinical samples ..	27
2.3.1	Experimental design.....	27
2.3.2	RRBS genomic mapping.....	27
2.3.3	Basic terminologies in next generation sequencing (NGS): depth and coverage ..	28
2.3.4	CpG coverage on heterogeneous clinical samples by RRBS and Infinium450K.....	30
2.3.5	The RRBS read density cut-off was set at 4 or more reads.....	31
2.3.6	DNA methylation by Infinium 450K and RRBS had a bimodal frequency distribution	32
2.3.7	An approximately linear relationship was observed between Infinium 450K and RRBS data, which was stable across samples	33
2.4	Comparison study between processed Infinium 450K and RRBS.....	36
2.4.1	The concordance of Infinium 450K with RRBS are improved by step-wide processing.....	36
2.4.2	Compare the proposed Infinium 450K processing with SWAN.....	37
2.4.3	Removal of batch effects contributes to improved concordance of processed Infinium 450K data and RRBS	39
2.5	Discussion	40
3	Comparing Methyl Capture Seq technology with Infinium450K for methylome analysis in clinical samples	43
3.1	Introduction.....	43
3.2	Material and methods.....	45
3.2.1	Objectives and study design.....	45

3.2.2	Processing of MC-Seq data and Infinium450K data.....	46
3.2.3	Annotation of CpG sites and comparison of genomic function and CpG content for MC-Seq and Infinium450K	47
3.2.4	Evaluation of technical variation and performance using lower quantities of DNA for MC-Seq	48
3.3	Results.....	49
3.3.1	Overview of MC-Seq analysis of human buccal epithelial samples	49
3.3.2	Reproducibility of MC-Seq.....	51
3.3.3	Performance of MC-Seq at 3 µg and 1 µg was similar	52
3.3.4	Methylation values from MC-Seq and Infinium450K were highly correlated and both gave a bimodal distribution.	53
3.3.5	Methylation values from both MC-Seq and Infinium450K distinguish inter-individual variation.	54
3.3.6	MC-Seq provides denser coverage of the epigenome	55
3.4	Discussion	56
3.5	Supplementary data.....	58
4	GEM: modelling the relationships among DNA methylation, genetic variants and environment factors.....	60
4.1	General multiple regression model	60
4.2	Understanding data matrix and terminologies in epigenetic research.....	61
4.2.1	DNA methylation: M	61
4.2.2	Genome-wide SNP measurements: G	63
4.2.3	Environment measurements including birth outcome and maternal body conditions: E	65

4.2.4	Covariates matrix for DNA methylation related study: cvrt	65
4.3	Gene, Environment and Methylation (GEM): A tool suite modelling the associations between genotyping, environmental factors towards methylome.....	65
4.3.1	GEM_Emodel: detecting methylation markers associated with environment factors.....	66
4.3.2	GEM_Gmodel: detecting methylation markers associated with genotype	67
4.3.3	GEM_GxEmodel: interaction of genotype and environment to specify methylation variants.....	68
4.3.4	Exploring the association between DNA methylation, genetic variants and environmental factors	68
4.4	Developing GEM package for efficient EWAS and methQTL analyses	70
4.4.1	GEM implementation.....	71
4.4.2	Benchmarking Results	74
4.4.3	Package availability and the ease of use	78
4.5	Conclusion and discussion	79
5	Modelling the relationships among environmental factors, genetic variants and neonate methylome	80
5.1	Introduction.....	80
5.2	Experimental design and data processing	81
5.2.1	Identification of 1423 Variable Methylation Regions.....	82
5.2.2	Genome-wide SNP measurements	83
5.3	Applying GEM to understand neonate methylome by genetic variants and environment factors.....	83

5.3.1	Ethnicity associates with the first components of genotype but not DNA methylation.....	85
5.3.2	Evidence for methQTLs	86
5.3.3	GxE model best explained the variation in methylation at most VMR-CpGs	88
5.3.4	CpGs with the most variation across samples are most likely to be driven by genotype.....	91
5.4	Discussion and conclusion	91
6	Extending the story of <i>HIF3A</i> association with adiposity to before birth.....	95
6.1	Introduction.....	95
6.2	Study design and data	96
6.2.1	GUSTO participants and methylation/genotype data processing.....	96
6.2.2	Estimation of cell fractions from methylation data	99
6.3	Statistical analysis and results	100
6.3.1	<i>HIF3A</i> methylation in umbilical cords was positively associated with birth weight and adiposity	102
6.3.2	The association of <i>HIF3A</i> methylation and birth weight was independent of the contribution of <i>cis</i> -acting SNPs; but the association was stronger in one genotype group.....	103
6.3.3	None of the prenatal environmental variables examined was consistently associated with <i>HIF3A</i> methylation and birth weight.....	105
6.3.4	<i>HIF3A</i> methylation levels at birth predicted infant weight in the first two years of life.....	106
6.4	Discussion	106
7	Summary and future work.....	108

7.1	What we know now, we didn't know then.....	108
7.2	Summarization and contributions of this thesis	109
7.3	Limitations in this study.....	111
7.3.1	The issue of tissue heterogeneity	111
7.3.2	Limitations in computational analysis.....	112
7.4	Future work.....	114
	Reference	118

List of Tables

Table 1	CpG probe information and quality control.....	23
Table 2	Illustration on the definition of read depth of sequence in next generation sequence.....	29
Table 3	Demographic characteristics of the seven samples.....	30
Table 4	Number of CpGs extracted by RRBS and shared with Infinium 450K across the seven samples.....	31
Table 5	Number of CpGs showing agreement within 5%, 10% and 20% ranges, between Infinium 450K and RRBS ($N_{reads} \geq 4$) data, at different levels of Infinium processing.	36
Table 6	The pseudo codes of peak-based correction (PBC) and SWAN.....	38
Table 7	Number of CpG percentage methylation values showing agreement within 5%, 10% and 20%, between SWAN processed Infinium 450K and RRBS data ($N_{reads} \geq 4$), at different percentage methylation value ranges.	38
Table 8	Comparison of Infinium 450K data processing methods/pipelines.	40
Table 9	Platforms for DNA methylation profiling.....	44
Table 10	DNA methylation processing for MC-seq sequences	47

Table 11 Summary of sequence alignment and duplicate rates for 7 buccal epithelium samples from MC-Seq.	51
Table 12 The coverage of genomic function and CpG content for Infiniu450K and MC-Seq are comparable	56
Table 13 MC-Seq covers CpG islands with a higher density.	56
Table 14 A sample DNA methylation data matrix shows that six CpGs were measured through four samples (S1, S2, S3 and S4). Methylation profiles showed cg00000109, and cg00000658 were hyper-methylated while the other CpGs were hypo-methylated in the four samples.	62
Table 15 A sample genotype data matrix table shows that five SNPs were measured through four samples (S1, S2, S3 and S4). We used additive pseudo codes “0”, “1” and “2” to stand for major allele homozygote, heterozygote and minor allele homozygote.	64
Table 16 Pseudo R code for GEM_Emodel.....	72
Table 17 Pseudo R code to explore methQTLs by Gmodel.....	73
Table 18 Benchmarking time consumption of GEM implementation on Emodel, Gmodel and GxEmodel by comparing it with a R script available in public database and using a standard laptop/HPC setting. (** The time for LM_Gmodel and LM_GxEmodel in standard laptop was computed based on the time cost on 10 CpGs.)	77
Table 19 Ethnicity and sex of subjects.....	82
Table 20 Distribution of VMR-CpGs and SNPs on chromosomes.....	84
Table 21 Comparison of our study design with those from Dick et al [29].	97
Table 22 Characteristics of participants in the GUSTO cohort studied in the analysis.	97
Table 23 Maternal status and maternal conditions.....	99
Table 24 Models in HIF3A analysis	101

Table 25 Association between methylation at three <i>HIF3A</i> CpGs and neonatal anthropometric outcomes at birth and gestational age.	103
Table 26 Association of two <i>cis</i> - SNPs (rs3826795, rs8102595) at <i>HIF3A</i> with methylation at three sites in <i>HIF3A</i>	104
Table 27 Association between SNPs at <i>HIF3A</i> locus and birth weight.....	104
Table 28 Two sided p-values from testing for interaction term between log-transformed birth weight and SNPs, with methylation as outcomes, adjusting for cofounders.	104
Table 29 Comparison of our study findings with those from Dick et al [29].	107

Index of Figures

Figure 1 A schematic depicting of epigenetic marks in chromatin, including but not being limited to DNA methylation, histone modification and non-coding RNAs. The chart is adopted from Fig 1. in [25].....	2
Figure 2 MicroRNA inhibits gene expression by binding to messenger RNA. Usually one strand of the messenger RNA enters into the micro-ribonucleoprotein complex and direct protein assembly [53].....	4
Figure 3 DNMTs catalyse the covalent addition of a methyl group (CH ₃) to cytosine (C5) to generate a methylated 5C, 5-methylcytosine (5mC). (A) the nucleic structure of cytosine. (B) 5-methylcytosine.	6
Figure 4 Charts for DNA methylation contents and patterns, adopted from [81]. The upper block shows methylated and unmethylated CpGs distribute in context of genomic functions, while the lower one shows in context of the density of CpGs (islands) and their neighbors.	8
Figure 5 Bisulfite conversion in detecting 5mC methylation, chart is taken from [97].	10
Figure 6 Density plot of beta values (% methylation) from Infinium450K type I and type II probes. A unique characteristic of Infinium 450K is that it includes both type I and type II assays for measurement of methylation levels at different loci. Due to different chemical design, Type I assay (as shown in black color) measures a wider range of methylation levels, while results from the type II assay (as shown in red color)	

have a narrower dynamic range. The differences between type I and type II assay require DNA methylation profiles from Infinium 450K to be processed carefully.	11
Figure 7 Thesis structure and chapters.	18
Figure 8 A pipeline for Infinium450K data analysis including two major steps of signal adjustment for color channels and correction between type I and type II measurements.	22
Figure 9 Extraction of red and green signal from control profile.	24
Figure 10 The procedure for type I and type II correction. (A) The density plots for beta values of type I and type II, x-axis is beta values, y-axis is density. (B) The converted M values of type I and type II, x-axis is M values, and y-axis is the density. S_U and S_M are the peak maximums of M density, while σ_U and σ_M are distances corresponding to peaks S_U and S_M . (C) The density plots for beta values after type I and type II correction. The left column is for type I and right for type II in all rows. After processing, type I keeps the same and type II is adjusted according to the distribution of type I (chart is adopted from supplementary material in [81].)	26
Figure 11 An overview of RRBS technology.	27
Figure 12 The concordance and comparison study of Infinium 450K and RRBS on same seven GUSTO neonates.	30
Figure 13 Optimal selection of number of reads quality cut-off in RRBS data. (A) Frequency distribution of number of reads in RRBS data. X-axis is truncated at 200 reads but maximum in data are 45,972. (B) Correction between Infinium 450K and RRBS values, estimated using Pearson R (black squares) and Spearman R (black circles). Numbers of CpGs remaining after read cut-off in RRBS are indicated in red stars.	32
Figure 14 Histogram of percentage methylation value frequency in (A) RRBS data (B) Infinium 450K data for the 454,660 CpGs covered by both technologies. Data was plotted into 201 bins stepped by 0.5 between 0% and 100%. For Infinium 450K, type I assay data was indicated in red and type II in blue.	33
Figure 15 The concordance of raw Infinium 450K data vs. RRBS data. (A) A scatter plot of percentage methylation values from RRBS (x-axis) and Infinium 450K (y-axis)	

over a density cloud. (B) Bland-Altman plot for raw Infinium data compared with RRBS data using Nreads $\geq$ 4.	35
Figure 16 Processed Infinium 450K data vs. RRBS data (A) A scatter plot of percentage methylation values from RRBS (x-axis) and Infinium 450K (y-axis) over a density cloud. (B) Bland-Altman plot for processed Infinium 450K data compared with RRBS (Nreads $\geq$ 4).	37
Figure 17 Percentage of assays with methylation values significantly associated with array on which sample was run, gestational age, or gender before (blue bars) and after (red bars) data processing in the set of 72 samples.	40
Figure 18 Study design and samples for comparison of MC-Seq and Infinium450K. .	45
Figure 19 Annotation of CpG content.....	48
Figure 20 Overview of MC-Seq and Infinium450K in terms of number of CpG probes.	50
Figure 21 Overview of MC-Seq from genomic features (A) and CpG content (B) for its target region, 200bp flanking to target region and observed samples.	50
Figure 22 MC-Seq technical comparison between seven samples and four replicates of same quantities of genomic DNA. (A) methylation comparison (B) clustering analysis.	52
Figure 23 MC-Seq technical comparison between samples and replicates of different quantities of genomic DNA (3 μ g and 1 μ g). (A) methylation comparison (B) clustering analysis.....	52
Figure 24 Methylation values from MC-Seq and Infinium450K were highly correlated and both gave a bimodal distribution. (A) Observed number of CpGs (vertical axis) from MC-Seq for one sample at different MC-Seq reads coverage (horizontal axis). (B) Pearson correlation (vertical axis) between methylation values from MC-Seq and Infinium450K at the same CpG sites, at different MC-Seq reads coverage (horizontal axis) for one sample. (C) Scatterplot of methylation values from MC-Seq ($\geq$ 10X, vertical axis) and Infinium450K (horizontal axis) at the same CpG sites for one sample. (D) Distribution of methylation values for all CpGs from MC-Seq ($\geq$ 10X, solid line) and Infinium450K (dotted line) for one sample.....	54

Figure 25 Hierarchical clustering analysis of MC-Seq values showed clinical samples clustered by ethnicity.	55
Figure 26 A typical Infinium 450K design for a probe of 50 base pair nucleotides to measure a CpG in gene PRDM16 region with ID as “cg08007925” in chromosome 1 and position as 3322020.....	62
Figure 27 SNP and major allele homozygote, heterozygote and minor allele homozygote	64
Figure 28 HPC computing settings (n=8).	74
Figure 29 A schematic of the analyses performed within Teh et al [17] . Gmodel sequentially tests the association of 1423 methylation marks with 708,365 genotypes (and covariates), Emodel sequentially tests the association of 1423 methylation marks with 19 environmental factors and GxEmodel sequentially tests that association of 1423 methylation marks with the interaction of every combination of the 708,365 genotypes and 19 environmental factors.....	75
Figure 30 QQplot for p-values from iterations of 255K CpGs with an environment factor. It was produced by GEM_Emodel.....	76
Figure 31 The operation time consumption benchmark on the associations for 250K CpGs and one environmental factor and one covariate for the sequential number of samples from 100 to 1000. X-axis is the number of subjects, and y-axis is the time consumption in seconds. The benchmarking is done in a personal computer (processor=2.2GHz, RAM=8G, system = window7 64bit).	76
Figure 32 The scatter plot to display an example of methylation corresponding to the environment in different genotype groups. AA, AB and BB are pseudo codes for major allele homozygote, heterozygote and minor allele homozygote. Phenotypic values are shown on the x-axis, and methylation values in percentage on the y-axis. The straight lines fit for associations in each group.....	78
Figure 33 GEM graphic user interface. (a) the selection of functions. (b) the file and parameter window.....	78
Figure 34 Emodel benchmarking for methylation matrix containing missing values. <i>pvalue</i> was transformed as $-\log_{10}$. A-axis is <i>pvalue</i> from LM_Emodel, y-axis is from	

GEM_Emodel. Among 255K CpGs that were tested, 18% of them contained at least one missing values. Our results showed pvaluefor CpGs without missing values are perfectly matched, while there were slightly differences between the two implementations when CpG contains missing values.....	80
Figure 35 Genotype, environment, and the interaction between the two influence DNA methylation levels.	81
Figure 36 Data repository for study relationships among genotype, environment and methylation profiles from 237 neonates.....	82
Figure 37 Principal component analysis on (A) genotype and (B) DNA methylation, shows different performance related to ethnicity. Subjects are coloured by self-reported ethnicity.....	86
Figure 38 Genotype is significantly associated with VMR-CpGs. Histogram shows the frequency of p-value distribution from regression analysis of 1423 VMR-CpGs against all SNPs.....	87
Figure 39 Strength of association between genotype and methylation is strongest for disrupting pairs and weakest for <i>trans</i> pairs.	88
Figure 40 Examples of VMR-CpGs whose methylation levels are significantly associated with phenotype in only one genotypic group. Phenotypic values are shown on the x-axis, and methylation value in percentages on the y-axis. (Left) Data for all samples. Samples are colored by their genotypic group (red indicates AA; blue, AB; black, BB), and a straight fit line is fitted for each group. (Right) Genotypic subgroup with the highest R^2	90
Figure 41 VMR-CpGs with larger ranges of methylation values are more likely to be methQTLs. Scatter plot showing the range of methylation values at each CpG across samples (x-axis) compared to the strength of association between the VMR-CpG methylation values and the genotype of the best SNP (y-axis).....	91
Figure 42 Location of methylation probes associated and SNPs in the <i>HIF3A</i> locus reported and studied.	96
Figure 43 Study design for HIF3A: data and modelling.....	101

Figure 44 Association between methylation at cg16672562 and birth weight, stratified by rs3826795 genotypes. (A) Scatter plots of birth weights against methylation. Each panel displays data of each genotype group. (B) Box plot of rs3826795 genotype against methylation. 105

Figure 45 Association of *HIF3A* methylations with child weight measured at 0, 6, 12, 18, 24 months. 106

Figure 46 Schematic diagram illustrating a putative causal pathway for how the effects of genome and environment on phenotype are transmitted through the epigenome and transcriptome. 116

1 Introduction

1.1 Objectives

The first genome in a living species was sequenced in 1995 [10]. Several years later in 2001, the first draft of the human genome sequence published in Science represented a monumental achievement [11]. It was hoped the three billion sequential base pairs composed of just the four nucleotides A,G,C and T would unlock the secrets of human development and disease [12]. Today, however, the task of linking genetic information with underlying biological function remains daunting, in particular, how a single set of genes regulates complex biological networks in over 200 different types of cells.

The Encyclopedia of DNA Elements (ENCODE) project [13] and the National Institutes of Health Roadmap Epigenomics Consortium [14], enabled by the advances in microarray and next generation sequencing technologies [15-20], have evidenced the epigenome and its associated regulatory elements render differentiated and specialized gene expressions in various cell types. Whilst there is one unique DNA sequence per individual, there are hundreds or thousands of different epigenomes in the same individual, changing across cell types and with the impact of time and environmental exposures. Interrogating the interaction of genetic and environmental influences and the relationship of epigenome with development and diseases cannot be elucidated without the extension to multidimensional “omics” data.

The importance and complexity of the epigenome are getting better understood, with the contribution of computational algorithms towards data analytics and statistical integration with genetic and clinical information. However, comprehensive analyses are still relatively scarce. Currently, the computational burden of analyzing the epigenome at a genome-wide scale is beyond the reach of many researchers. Thus, domain-specific computational methodology is in high demand to process, integrate and analyze the epigenome and keep pace with exponential increases in data size and complexity [3]. The epigenetic marks discovered by intensive statistical study can lead with experimental exploration and validation, to personal disease mechanisms.

1.2 State of the Art: Epigenetics and DNA methylation

Epigenetics was first termed by Waddington in his book “An Introduction to Modern Genetics” in 1939 ([21], page 154). Widespread use of the term epigenetics, which refers to “heritable changes that do not involved changes in DNA sequence”, did not occur until the 1990s [22]. “Epigenetics” as a word, its Greek prefix “epi-” means “upon, at, close upon, in addition”, implies epigenetics is “on top of” or “in addition to” genetics. The current well accepted definition of epigenetics is “the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence”[23].

Epigenetic marks, referring to chemical modifications to DNA or molecules that interact with DNA, include, but are not limited to DNA methylation, histone modification and noncoding RNAs [24], as depicted in the schematic figure (Figure 1) [25]. Epigenetic marks are heritable and can be transgenerational (reviewed by Heard and Martienssen [26]). In principle, the term epigenome refers to the global view of sequence-independent processes in genome [24, 27]

Figure 1 A schematic depicting of epigenetic marks in chromatin, including but not being limited to DNA methylation, histone modification and non-coding RNAs. The chart is adopted from Fig 1. in [25].

One of the many known epigenetic marks, DNA methylation is thought to contribute to stable longer-term gene expression regulation and tissue differentiation [28]. Studies in large human cohorts [29-31] have shown DNA methylation more suited for genome-wide assessment in epidemiological studies than others types of epigenetic mark [32] . Nevertheless, it is important to know that gene expression is controlled by the

interaction of many different epigenetic factors, not only DNA methylation[33], including histone modification [34-36], and non-coding RNAs [37]. The ENCODE project [13, 38] taught us that regulator factors work in interaction with each other and DNA tertiary structure [39] and that promoter function as a product of protein binding governed by the combined forces of DNA methylation, histone modification and chromatin accessibility explains the majority of control of transcriptional activity [13, 38] .

Histone modification

Histones are highly alkaline proteins in eukaryotic cell nuclei that package and order DNA into structural units. The post-translational modification of histone proteins has direct and indirect effects on transcription of DNA in a temporal and spatial manner [40]. Histone can obstruct and bend the DNA causing reduced DNA accessibility to transcription factors. Histone proteins can be modified in many different ways, such as acetylation, methylation, phosphorylation, ubiquitination, simulation, ADP ribosylation and deamination [34], and others [41]. In addition to these direct effects on DNA structure, histone modifications often act indirectly by recruiting effector proteins to activate downstream signaling [42] [43], and blocking access to the DNA for remodeling complexes [44]. Thus, gene expression can be controlled, in part, by post-translational modifications of the histones [45].

Genome wide histone modification study in large cohorts is hindered by many difficulties. Firstly, it is hard to measure acetylation of the numerous potential sites of histone modification simultaneously. Histones are composed of two copies of each of four core subunits: H2A, H2B, H3, and H4. Together, these comprise a histone octamer, with each subunit containing between 10 to 20 lysines (such as K9, K27, K119 etc.) that can be targeted for acetylation (such as H3K18ac [46]) or methylation (such as H3K27me [47]) and more than 100 potential sites of acetylation on the complete octamer. Secondly, characterization of the genomic locations of post-translational histone modifications requires fresh tissue samples in order to preserve the native crosslinking DNA and proteins architecture. It is not feasible in patient-derived clinical samples that are generally preserved by freezing and chemical fixation [48]. Thirdly, Chromatin immunoprecipitation (ChIP) followed by a genome-wide sequencing (ChIP-Seq) [49] which is often used for profiling histone modification , is

heavily dependent on the availability of specific antibodies and generally requires a relatively large number of cells (typically $0.5\sim5 \times 10^6$ cells) [50].

So, histone modification study in large scale cohort is still at the very early stage, due to the difficulties of feasibility. Nevertheless, recently we have seen a successful cohort study of H3K27ac in 257 postmortem samples, across subjects with autism spectrum disorder (ASD) and controls, by ChIP-seq platform [51].

Non-coding RNA

Non-coding RNAs (ncRNAs) are RNAs that are transcribed from DNA but are not translated into proteins. In general, ncRNAs regulate gene expression at the transcriptional and post-transcriptional level. Those ncRNAs that appear to be involved in epigenetic processes can be divided into two main groups, the short ncRNAs (<30 nucleotides (nts), like microRNA) and the long ncRNAs (>200 nts). All classes of ncRNA have been shown to play a role in heterochromatin formation, histone modification, DNA methylation targeting, and gene silencing. Recently, circular RNA has been reported as a new member of the non-coding RNA family and it exhibits in tissue/development-stage specific expression and links to phenotypic changes [52]. Figure 2 uses miRNAs as an example of ncRNA affecting transcription.

Figure 2 MicroRNA inhibits gene expression by binding to messenger RNA. Usually one strand of the messenger RNA enters into the micro-ribonucleoprotein complex and direct protein assembly [53].

DNA methylation

In humans, DNA methylation refers to the covalent addition of a methyl functional group (CH₃) onto cytosine residue specifically occurring at a cytosine/guanine dinucleotide, separated by a single phosphate molecule (CpG). Since transcription factors typically bind to the major groove of DNA, addition of this methyl group on cytosine residues is postulated to decrease binding affinities of such proteins while increasing the affinity of other proteins such as methyl-binding proteins which are typically associated with heterochromatin formation and gene repression [54]. Cytosine methylation patterns across the genome are relatively stable and most CpGs found in humans are methylated (70-80%). Nevertheless, global DNA methylation is variable between different cell types – some highly specialized cells, such as spermatozoa, have almost completely methylated DNA, whereas others, such as oocytes, have near baseline globally unmethylated DNA [55].

According to Rakyan [56], large-scale epigenomic profiling to discover disease biomarkers can be feasible only if the epigenetic marks are stable enough to be amenable to high-throughput analysis with easily accessible clinical samples. Currently, of all known epigenetic mechanisms, DNA methylation is the most well-studied and best understood [57] because methylated DNA sequences are amenable to possess those required attributes for large-scale profiling. Thus DNA methylation represents the most studied epigenomic modification in clinical samples [58]. This thesis concentrated particularly on DNA methylation in clinical tissues.

Among the four heterocyclic bases in DNA: adenine (A), guanine (G), cytosine (C) and thymine (T), two of which, cytosine and adenine (A and C), can be methylated. In the mammalian genome, DNA methylation is a covalent modification that occurs predominantly at cytosine residues usually followed by a guanine (Cytosine phosphate Guanine dinucleotide, or CpG as there is only one phosphate linking C and G). The DNA methyl transferase (DNMT) enzymes catalyze the covalent addition of a methyl group (CH₃) to the 5th position of cytosine ring (5C) that gives rise to 5-methylcytosine (5mC) [59], as shown in Figure 3.

Figure 3 DNMTs catalyse the covalent addition of a methyl group (CH₃) to cytosine (C5) to generate a methylated 5C, 5-methylcytosine (5mC). (A) the nucleic structure of cytosine. (B) 5-methylcytosine.

Recently 5-Hydroxymethylcytosine (5hmC) was observed to be abundant in mammalian brain [60], as well as in embryonic stem cells [61], although its role remains unclear. Also, though cytosine methylation is thought to be largely restricted to the CpG dinucleotide, recently studies have been shown to methylate non-CpG cytosine (CpH, H = A/C/T) in human brains [62]. Despite the emerging discovery cell biology and diverse mechanisms in different species, DNA methylation in this thesis is restricted to the measurement of 5-methylcytosine (5mC) at CpG dinucleotide in human clinical samples.

DNA methylation as a chemical modification does not change the DNA sequence but has influence on gene activity. The emerging understanding of the functions of DNA methylation is more nuanced than only gene expression silencing [63], now we understand that the mechanisms of DNA methylation are locus-specific. Methylation can block access of different types of protein factors to DNA and so can activate, repress, modulate or change the splicing pattern of coding genes and non-coding RNAs. It is essential for normal development and plays key roles in diverse biological processes such as X chromosome inactivation [64], transposable element repression [65], genomic imprinting [66], tissue-specific gene expression [67], aging [68] and cancer [69].

Understanding DNA methylation function requires consideration of its genomic context, which is always represented by CpG location with respect to genomic

functional elements, such as transcriptional start sites, gene bodies or intergenic regions [70, 71]. CpGs are generally underrepresented in the genome but tend to be overrepresented at gene promoter regions the top bar in Figure 4 shows the uneven distribution of methylated and unmethylated CpGs within a typical gene. The diverse of CpG density and surroundings are termed into four CpG contents, island, shore, shelf and open sea. The CpG-rich regions are called CpG islands, 2,000 base pairs adjacent to the CpG island are South or North CpG shores (S shore and N shore), with CpG shelves being 2,000 base pairs further and adjacent to the shores (S shelf and N shelf), bottom bar in Figure 4. The remaining CpG sites that spread out across the “open sea” are heavily methylated [72]. The majority of transcription start site, roughly 70%, resides within CpG islands [73].

Of the many established epigenetic marks and mechanisms, DNA methylation is thought to contribute to stable long-term gene expression regulation and tissue differentiation[74], and is ideally suited for genome-wide assessment in large human epidemiological studies [32]. The modifications are highly inheritable through generations [75, 76]. CpG islands, especially those associated with transcription start region (transcription promoters), are found to be highly conserved between mice and humans [71]. However, during development, DNA methylation patterns are dynamic [77, 78] . A stable DNA methylation pattern during cell differentiation regulates tissue-specific transcriptome [79, 80], but there are also a lot of variable DNA methylation regions across individuals and associated with individual phenotypes [72].

Figure 4 Charts for DNA methylation contents and patterns, adopted from [81]. The upper block shows methylated and unmethylated CpGs distribute in context of genomic functions, while the lower one shows in context of the density of CpGs (islands) and their neighbors.

1.3 DNA methylation measurements and Infinium450K technology

There are about 28 million CpGs in the human epigenome [82]. The advent of high-throughput technologies have facilitated DNA methylation measurements that were previously restricted to specific loci in a limited number of genes and can now be performed on a genome-wide scale. It enables epigenome wide association studies (EWAS), and more specifically DNA methylome wide association studies (methWAS) [83] and provides a novel approach for researchers to explore mechanisms of human diseases. Technologies to profile the DNA methylome mainly fall in two categories, microarray and next-generation sequencing (NGS). Several reviews compared these two types of technologies and examined their strengths and weaknesses for DNA methylation profiling [84-87].

Sequencing technologies, for example, Whole Genome Bisulfite Sequencing (WGBS) [88], can in principle assay all of the approximately 28 million CpGs in the human genome [89]. Alternatively, Reduced Representation Bisulfite Sequencing (RRBS) [90], and Methylated DNA Immunoprecipitation Sequencing (MeDIP-seq) [91] were developed in DNA methylation detection and analysis. Recently Methyl-Capture Sequencing (MC-Seq), which can assay about 2-3 million methylation sites, were introduced to study clinical samples [2, 18]. NGS technologies provide single base CpG measurement genome wide. However, taking into consideration on their degrees of technical complexity, methylome coverage and precision in resolving methylation,

NGS technologies are of higher costs, larger data volume, more complex in processing and less sensitivity to subtle methylation changes. These disadvantages limit NGS the application in epigenetics in large data cohort by far.

While advanced sequencing-based methodologies have been applied in epigenetic studies, array-based technologies are more welcome in research labs especially for large sized cohort study because of the balanced genomic coverage, resolution, accuracy, specificity, throughput and cost. The most popular DNA methylation technology is Infinium (The Illumina® Illumina Inc., CA, USA)

HumanMethylation450 BeadChip (Infinium 450K) [15, 92]. The advantages of high-throughput and low cost enabled Infinium450K to become the most popular platform for interrogating DNA methylation in epigenome-wide association studies (EWAS [56]) involving hundreds or even thousands of cohort samples [29, 93] as well as resource-intensive efforts such as the International Cancer Genome Consortium (ICGC) [94], the cancer genome atlas [95], ENCODE [38] and the International Human Epigenome Consortium (IHEC).

Infinium450K, evolved from Infinium27K (HumanMethylation27), can assay 485,777 single-based CpG sites and analyze twelve samples in parallel in one chip and up to 96 samples per run [15]. Genome-widely, it covers 99% of RefSeq genes and all the different epigenetically key genomic regions such as CpG islands, shores and shelves, 5' and 3' UTRs, and promoters and gene bodies. The new Illumina Infinium MethylationEPIC BeadChip, available mid-2016, includes 856,187 sites and retains 90% of the Infinium450K probes while adding 333K probes to gene enhancer regions [96]. The chemistry of the EPIC array is identical to that of the 450K.

The Infinium arrays use bisulfite conversion technologies to detect CpG methylation profiles. Cytosine and 5-methylcytosine (5mC) have different sensitivity in their bisulfite conversion. Under acidic conditions, unmethylated cytosine undergoes conversion to uracil (U), while methylated cytosine (5mC) remains unchanged. The complementary 3'-base is either C (methylated) or T (unmethylated), Figure 5.

Figure 5 Bisulfite conversion in detecting 5mC methylation, chart is taken from [97].

With the discovery of 5-hydroxymethylcytosine (5hmC) some concerns have been raised regarding DNA methylation studies using the bisulfite sequencing technique. 5hmC has been shown to behave like its precursor, 5-methylcytosine (5mC), in bisulfite conversion experiments [15]. Therefore, bisulfite sequencing data may need to be revisited to verify whether the detected modified base is 5mC or 5hmC. In 2012, a technology of TAB-seq [98] was discovered to solve the problems of being detected as 5mC in normal bisulfite conversion experiments using the oxidative properties of the Tet-family of enzymes.

In order to increase the efficiency of signal detection, Infinium450K arrays apply two different types of chemistry with two different types of 50 base pair probes but by different mechanisms. Type I assay employs two bead types per CpG locus. Type II assay utilizes one probe per locus. For each CpG probe, DNA methylation profile is calculated as the *beta* value, as $beta = M / (M + U)$, where *M* is ratio of the intensity of methylated signal over the sum of methylated and unmethylated signals (*U*) [15]. The *beta* value ranges between 0 to 1 and normally presents as a percentage value, so that *beta* value at 0~20% indicates the probe hypo-methylated while 80~100% means hyper-methylated.

Figure 6 Density plot of beta values (% methylation) from Infinium450K type I and type II probes. A unique characteristic of Infinium 450K is that it includes both type I and type II assays for measurement of methylation levels at different loci. Due to different chemical design, Type I assay (as shown in black color) measures a wider range of methylation levels, while results from the type II assay (as shown in red color) have a narrower dynamic range. The differences between type I and type II assay require DNA methylation profiles from Infinium 450K to be processed carefully.

The bimodal distribution of *beta* values derived from type II assays is narrower than that from type I assays, as shown in [81]. Type II signals cover a narrower dynamic range of methylation values than type I signals [99], which is so called type II bias. In addition, batch effects (such as different date of experiments, different chips, different technicians running samples etc.) are a common source of variation in high-throughput experiments. They contaminate the real biological or scientific variables in the study. This suggests that data from the type I and type II assays should be considered separately in any downstream bioinformatics analysis or normalized by sophisticated methodologies. In this thesis, I present an analytic pipeline for Infinim450K data analysis [1] with the validation by concordance study on same clinical samples across microarray and NGS platforms.

Correction for cellular heterogeneity is absolutely necessary in analysis of methylation data from mixed cell type samples [100-102], as DNA methylation profiles will vary across cell types. Cellular heterogeneity correction can be achieved by performing the DNA methylome profiling on homogenous samples or by directly measuring the cell mixture in the DNA sample. However, such measures are not always feasible, for instance, when investigating previously frozen bloods, or tissues that are by nature heterogeneous and not easily fractionated (e.g. placenta). In these cases, cellular

heterogeneity has to be estimated by statistical approaches: either by reference to existing methylation profiles of the constituent cell types, which is called the reference-based method [103]; or, by the reference-free method [104] which can be used to correct cell type effects without a prior estimate of cell types. Both approaches have their limitations. Reference-based estimation requires the use of an appropriate cell-specific methylation reference-panel, while the reference-free approach may over adjust and reduce important variability in the data. In order to improve our corrections for cellular heterogeneity, our lab has started an effort to construct reference-panels by investigating cellular components of the tissues that we are studying.

1.4 DNA methylation is associated with early experience

Efforts are made to achieve unbiased DNA methylation detection and processing, in order to explore how changes in DNA methylation can alter human development, cell differentiation and diseases [105]. Clinical epigenetics tends to identify aberrantly methylated sites [106], which are potentially disease biomarkers. A well-known example is that DNA methylation plays a significant role in cancer [107-110]. Studies have also established the relationships between DNA methylation alternations and diabetes (reviewed in [111, 112]), sclerosis [113], schizophrenia (reviewed in [114]) and ageing (reviewed in [68, 115]). Further, it is also understood DNA methylation marks are susceptible to environmental modification [116], and continue to change well beyond the point of cell differentiation [117]. Therefore, DNA methylation could potentially serve as a molecular mechanism for environmental causes of disease and trans-generational transmission [57, 118].

Environment during development was first addressed as an important factor in disease by Barker [119], who emphasized that early environment, especially fetal life was the “developmental origins of health and disease” (DOHaD) [116, 119]. The “Barker hypothesis”, also known as the “fetal programming hypothesis”, was extended to the “development origins of disease” hypothesis by Gluckman and Hanson and others [118, 120] who proposed environmental exposures during early life can cause biological changes that influence disease susceptibility later in life.

The environmental epigenetic hypothesis [121, 122] for the development origins of disease suggests that alteration of fetal epigenome may have health consequences in later life [123]. Evidence for this hypothesis derives from studies documenting the

relationships among DNA methylation marks, *in utero* environment and later phenotype [60, 93, 124-128]. An increasing number of reports have shown transcriptome [6] and DNA methylation differences responded to *in utero* [129, 130], at birth [129] and early life [130]. Empirical examples included methylation of the glucocorticoid receptor (*NR3C1*) as a result of early life stress [131, 132], fibroblast growth factor receptor 2 (*FGFR2*) methylation is associated with birth weight [131, 133] and methylation of the aryl hydrocarbon receptor repressor (*AHRR*) [134, 135] and coagulation factor II receptor like 3 (*FR2L3*) [136, 137] as a result of smoking exposure. Maternal micronutrients such as folate can modify the offspring's LINE-1 methylation at birth [138] and is associated with birth weight [139] and fetal outcome [140]. Recent reports showed that epigenetic modifications identified from maternal gestational diabetes might be potential epigenetic biomarkers [141].

In addition, some methylation marks act in concert health phenotypes. For example, methylation in hypoxia induced factor 3 (*HIF3A*) is associated with obesity [29]. Some methylation changes can predict later phenotype e.g. methylation in retanoid X receptor alpha (*RXRA*) at birth predicts body composition at six years of age [129]. There are also examples of persistent, environmentally-induced DNA methylation states. For instance variations in maternal care alter the methylation of the glucocorticoid receptors (*NR3C1*) promoter in rats [142] and differential methylation at the same region is associated with childhood in human [131]. Although many aspects of the epigenome such as histone modification are likely to be involved, current evidence concentrates mostly on DNA methylation.

1.5 DNA methylation is influenced by genotype

Epigenetic states can be related to genotype. Genotype is the DNA sequence of the genome of a cell. Single nucleotide polymorphisms (SNPs) are the measurement of the most common type of genetic variants and emerged as important marks for linking genotype and individual diseases and traits in the genome-wide association studies (GWAS) [143]. The outcome of genotype is phenotype, which refers the particular appearances and characteristics of individuals, for example, eye color or certain disease.

Gibbs et al [144] showed that in human brain tissues, there were about 4~5% of the CpGs studied significantly dependent on genotype. Gibbs et al. thus defined the CpGs

dependent on genotypes as methylation quantitative trait loci (methQTL). They also categorized CpG and SNP pairs being within 1Mb chromosomal region as *cis*-methQTL and across chromosomes as *trans*-methQTL. The existence of methQTLs was validated in other studies [145-148].

Ethnicity can be used as a proxy for genotype and has been observed to influence the methylome [149]. African and European individuals have population-specific patterns of DNA methylation at approximately 30% of CpGs measured and, methylation levels at around 50% of these population-specific CpGs could be explained by divergence in allele frequencies at *cis*-acting SNPs between populations [150]. Studies of methylation differences using the Infinium450K methylation array data from 133 lymphoblastoid cell lines from European and African HapMap samples reveal that approximately 13% of analyzed CpGs showed significant methylation differences between the populations and over 50% of these were in methQTLs with local SNPs [151]. CpGs showing differential in methylation levels across ethnicities are more likely to be driven by genotype than other CpGs (68% methQTLs compared to 13% meth-QTLs) [152]. Nevertheless, these studies also found that some methQTLs are specific to one population with no correlation between genotype and methylation in the other population, suggesting possible gene and environment interactions.

There is evidence for an environmental influence on DNA methylation states as well as an influence of genotype. This suggests that gene $\times$ environment interactions may be common. Recent studies have provided evidence for gene $\times$ environment interactions (G $\times$ E, pronounced as “G times E”) on DNA methylation. Yousefi et al. [153] showed that *LEPR* genotype interacted with maternal smoking to predict methylation of *LEPR* in the offspring. A SNP within the *IL4R* gene combined with methylation at a CpG site within the same gene predicts risk of childhood asthma [154]. Klengel et al. [155] described that genotype of the *FKBP5* gene in interaction with early childhood trauma affects methylation of *FKBP5* intron 7, *FKBP5* expression and subsequent deregulation of glucocorticoid receptor signaling. It is unknown what relative portion of inter-individual variation in methylomes is driven by genotype, environment or an interaction of gene and environment [156].

All these findings are consistent with the model that Feinberg and his collaborators proposed in early 2004, in which epigenetics, genetics, and environmental health could be integrated to create a new field of epigenetic epidemiology [157]. Recently,

Holbrook [158] suggested that the epigenome is the product of the integration of genetic variation and environmental exposure.

1.6 Understanding methylation and modelling epigenome wide association study by GEM: opportunities and challenges from a computational perspective

Despite the success of genome-wide association studies (GWAS) in identifying single polymorphism nucleotides (SNP) associated with diseases and health outcomes, studying the epigenetic variations genome-widely and in large-scale cohorts, termed Epigenome-Wide Association Studies (EWAS) [159], presents novel opportunities but also creates new challenges that are not encountered in GWAS.

In EWAS, the relationships between methylation, genotyping and environment are complicated. Understanding DNA methylation changes related to environmental exposures and developmental trajectories to disease is highly desirable [158] and their discovery is the aim of many EWAS [160, 161]. However, what has really pushed EWAS studies to the brink of what is computationally possible is the realization that DNA methylation levels are not just specified by extrinsic factors but also are influenced by genotype. Polymorphisms close to CpGs in the same chromosome (*cis*-) often form methylation quantitative trait loci (methQTLs) with nearby CpGs [144-146], or blocks of *cis*- polymorphisms associated with a cluster of methylation quantitative trait loci, named GeMES (15, 19). Furthermore, it is now apparent that genotype can work in interaction with environment (G x E) to influence specific DNA methylation levels [4, 31] and these can be linked to phenotypes [5, 8].

EWAS is very challenging. Unlike genetics which deals with one unique copy of genetic sequences for each individual, there are many versions of DNA methylation profiles for one individual in different cell types and at different ages. For example, gestational age is correlated with inter-individual variation in the DNA methylation profiles of neonate's cord blood [5] and the DNA methylome of older people are very different from those of neonates [162]. Environmental factors are vital components in EWAS, however missing values and unknown factors make large-scaled study very difficult. An EWAS study on gestational environment and birth outcome could be hampered due to missing values of maternal lifestyle and socio-economic status, for example, mother's smoking history before pregnancy [163]. Demands for measuring

environment factors make large scale sampling very costly. Literature survey by end of 2016 showed only four EWAS studies performed on large cohorts with 1000+ samples [29, 164-166], while GWAS has been extended to the scale of millions of samples [38].

Computational burden is another issue in EWAS, especially when the interaction of genotype and environment is included. High-throughput technology can detect methylation changes at 450,000 loci by Infinium450K and 1,000,000 genetic changes by Illumina Omniexpress+exome genotyping arrays. Creating a genome wide methQTL map requires assessing the correlation of genotype at millions of SNPs with thousands to millions of CpG methylation states, by billions of linear iterative regressions. Sun 2014 [167] surveyed methQTL studies between year 2010~2014 and found that most of methQTL studies were restricted to screens of *cis*- SNP-CpG pairs, while some were even restricted to the 50 Kb ~1Mb regions flanking to each SNP. However SNPs far from the CpG or in different chromosome (*trans*-) were also reported to be associated with CpG. *Trans*- methQTLs have been detected to be relevant to normal or disease states in many studies [168]. With the availability of multiple “omics” data for the same samples and quick expanding environment factors or exposures, exponentially increasing computation requests optimized operation and efficient methodologies.

In this thesis, a tool suite to explore the associations of Gene, Environment and Methylation was created and named as GEM. It provides three statistical models denoted as GEM_Emodel, GEM_Gmodel and GEM_GxEmodel to facilitate analyses in EWAS. The GEM_Emodel tests the association of methylome marks and environmental factors; the GEM_Gmodel creates a methQTL genome-wide map; finally, the GEM_GxEmodel tests the ability of gene and environmental interaction models to predict DNA methylation levels. In this thesis we designed and implemented GEM package for efficient operation to deal with large scaled modeling in EWAS and methQTLs genome widely in a standard computer within minutes [3]. GEM has been successfully applied in epigenetic study from neonate methylomes [4] and the discovery of epigenetic biomarkers [5].

1.7 GUSTO cohort and samples

Studies in this thesis are from clinical samples from **G**rowing **U**p in **S**ingapore **T**owards healthy **O**utcomes (GUSTO) birth cohort [169]. GUSTO is a Singaporean birth cohort which was carefully designed to collect biological specimens and prenatal and postnatal data from parents and their infants. More than 1,200 umbilical cords were collected at birth from babies born in Singapore's two major public maternity units, National University Hospital (NUH) and KK Women's and Children's Hospital (KKH). GUSTO recruits the babies from homogeneous parental ethnic background, a very unique cohort comprised of three different Asian ethnicities, Chinese (57.2%), Malay (25.5%) and Indian (17.3%). Data and specimens are collected from children at birth and throughout childhood till 9 years old and one aim of the cohort is to examine of the roles of fetal development and epigenetic factors in pathways to metabolic and neurocognitive diseases. By 2016, the eldest GUSTO child is 8+ years old.

The GUSTO birth cohort provided a comprehensive data set including clinical data for gestational environment, epigenetic profiles and genotyping from mother and child. The multi-dimensional 'omics' data and comprehensive clinical measurements provide the possibility for epigenetic study in tackling the associations among genotype, epigenomics and phenotype on samples from neonates and their parents. Upon the availability of GUSTO data, the main objective of this thesis is to explore the epigenetic mechanisms of development and diseases on Asian neonates by meaningful and relevant computational analysis.

1.8 Overview of the thesis

Early in the year 2005, Barker suggested the role of epigenetics in transmitting the influence of early life events to affect adult health outcome [170], however, genome wide study of epigenetic profiles was not started till 2011 [56]. Epigenomics research is new but has posed exciting opportunities for understanding the relations between human health and environment. However, current limitations in analytic techniques constrain the questions that can be asked of complex data and impair the efficiency of hypothesis generation and validation. Comprehensive data analysis algorithms are needed to process fast developing array and sequence technology. We need to develop of mathematical models enabling extraction of robust patterns from complex data, and to understand the fundamental principles behind such patterns. This thesis reports

research work lying at the intersection of multiple disciplines, including data modeling, mathematical analysis and biological discovery. Most importantly, this thesis demonstrates the improvement in computational operations and modeling can greatly benefit biomedical research.

The thesis is organized into seven chapters, as shown in Figure 7 , which are arranged based in the order of “DATA-Methodologies-Application”. After the overview of epigenetics and presentation of the motivations and challenges for this project in Chapter 1, Chapter 2 describes the most widely-adopted epigenetic profiling technology, Infinium 450K, followed by a novel statistical and computational pipeline for improving Infinium 450K data processing. The proposed method for Infinium 450K was validated with a concordance study on the same seven clinical samples across from an independent next generation sequence platform (NGS) [1]. In Chapter 3, the evaluation of epigenetic profiling technologies was extended to a novel and recently developed NGS platform, Methyl Capture Seq (MC-Seq) [2]. Our detailed comparative analysis on MC-Seq and Infinium 450K data showed MC-Seq can be an attractive cost-effective alternative to whole-genome bisulfite sequencing.

Figure 7 Thesis structure and chapters.

In Chapter 4, GEM is contributed to computational analysis, modeling and integration of epigenetic, genetic and clinical data. GEM is a R package [171] with a suite of statistical models studying the interplay among Gene, Environment and Methylation [3]. The three major functions GEM_Emodel, GEM_Gmodel and GEM_GxEmodel work for efficient EWAS, methQTL detection and GxE analysis. Implemented by matrix based iterative correlation and memory-efficient data analysis, within GEM,

millions of association studies can be facilitated on a standard laptop computer within minutes. GEM is validated by comprehensive benchmarking and has been contributed to Bioconductor [172].

In Chapter 5, the DNA methylation analytic pipeline was used for data processing and GEM was applied to an infant-mother cohort study with hundreds of clinical samples. We presented the first genome-wide integration of genetic, clinical and epigenetic data. Our results suggested that the major driver of neonatal DNA methylation inter-individual variation is the interaction of genetics and *in utero* environment [4]. Simultaneously, Dick et al (2014) [29] worked with a much larger adult cohort with more than a thousand obesity cases and controls and found that DNA methylation on the first intron of *HIF3A* is significantly associated with adult BMI [29]. We explored the same *HIF3A* methylation probes and single nucleotide polymorphisms (SNPs) in GUSTO infant cohort and extended the finding to show that the story of *HIF3A* association with adiposity “begins before birth” [5] in Chapter 6. The thesis ends with the summary of the study, discussion of limitations and the plan for future work, in Chapter 7.

2 Improving the process of Infinium450K for measuring the methylome in clinical samples

2.1 Introduction

The Infinium Human Methylation450 BeadChip ArrayTM (Infinium 450K, or 450K in a short) is a microarray technology for human DNA methylation profiling [15]. It interrogates 485,577 methylation sites per sample at single-nucleotide resolution with 99% coverage of known genes. Infinium 450K has been well accepted as DNA methylation measurement especially for large size clinical samples. Data by Aug 2016 showed, there were 576 research results conducting 37624 human samples by Infinium450K contributed to Gene Expression Omnibus (GEO, <http://ncbi.nlm.nih.gov/geo/>), a public functional genomics data repository.

Infinum450K platform is a hybrid of two different assays (Infinium type I and II) in order to cover more CpGs. Type I assay exploits two different probes corresponding to methylated and unmethylated alleles located on two different bead types and the methylated and unmethylated signals are generated in the same color channel. Type II uses only one bead type with a unique probe allowing detections of both alleles. The methylated and unmethylated signals are generated respectively in the green and red channels. It is noticed that methylation profiles obtained from type II probes were less accurate or reproducible than those obtained from type I probes [81], and type II signals are less stable and cover a narrower dynamic range of methylation values than type I signals [99]. The significant differences between the type I and type II assays require very careful and proper bioinformatics solutions in the first step in order to meet unbiased downstream analysis [1]. Besides, batch effect is observed when experimental procedures necessitate the processing of samples in separate groups or on different days. This is another source of technical artifact that has to be minimized.

In this chapter, a novel pipeline was designed to process 450K data in order to overcome its technique disadvantages and disagreements across batches. The data processing pipeline attempted to deal with technical artifacts in 450K by combining color adjustment and type correction step by step. However, given the lack of a “gold standard” in DNA methylation profiling, unbiased assessment of the data processing is difficult.

Treatment of DNA with bisulfite results in the deamination of unmethylated cytosine to uracil, leaving methylated cytosines unaffected, as shown in Figure 5 and [97], and thereby providing a methyl-dependent DNA mutation that can be easily detected by next generation sequencing methods; known as bisulfite sequencing. In order to reduce the costs involved with bisulfite sequencing a restriction enzyme is often employed along with selection of the resulting small fragments of DNA, to enrich for fragments with high CpG content, thereby greatly reducing the percentage of the genome that needs to be sequenced to analyze CGI methylation. This technique is known as Reduced Representation Bisulfite Sequencing (RRBS) [19].

RRBS offers approximately six times CpG coverage and achieves with better accuracy because of NGS technology [19, 173]. In addition, both RRBS and Infinium450K technologies measure DNA methylome in human samples. Both methods utilize sodium bisulfite treatment of DNA to convert unmethylated cytosine to uracil (C to U) [174]. Both assess methylation at single CpG resolution. So a concordance study was conducted on clinical samples for better understanding the two technologies. In order to benchmark the proposed pipeline, the comparison study between pre- and post-processed Infinium450K data and RRBS was conducted to validate the improvement of proposed data processing methodologies.

2.2 A novel pipeline to process the Infinium450K

A novel pipeline to process Infinium450K methylation data was proposed in Figure 8. It includes two major processes, named color adjustment and type I/II correction, which were introduced at the first time as the essential effects to minimize the technical bias and reduce the batch effect as much as possible. The process was implemented in the R programming language [1]. Overall, as shown in Figure 8, raw signals were extracted from GenomeStudio™. With the information of signal detection parameters, probe quality control was the next step and then color adjustment was applied to intensities from red and green color channels. M values were calculated from beta values (β values) and used for the type I and type II correction. After type I/II correction, M values were converted back to beta values for normalization across samples. Finally, the processed beta values were generated.

Figure 8 A pipeline for Infinium450K data analysis including two major steps of signal adjustment for color channels and correction between type I and type II measurements.

2.2.1 Raw data extraction and quality control

Infinium450K signal extraction was performed in GenomeStudio™ Methylation Module on the intensity files (with a .idat suffix) produced by the Illumina iSCAN system. Full raw data were extracted from GenomeStudio™ without using background subtraction or normalization. Basic information for raw data is described in Table 1. For each sample, this data is collected for a total 135,501 type I and 350, 076 type II CpGs. As can be seen, the data collection includes intensities for signal A/B (Signal_A, Signal_B), average number of beads for signal A/B (Avg_NBEADS_A, Avg_NBEADS_B) and detection_pvalues (Detection_pvalue) in columns. CpG probe quality control criteria include, probe should be present on three and more than three beads and the detection_pvalue should be less than 0.05 for methylated and unmethylated signals. CpG that failed the quality control criteria for any sample would be discarded from further analysis.

Table 1 CpG probe information and quality control

CpGID	Signal_A	Signal_B	Avg_NBEADS_A	Avg_NBEADS_B	detection_pvalue
CpG QC criterial	(Avg_NBEADS_A ≥ 3) AND (Avg_NBEADS_B ≥ 3) AND (detection_pvalue < 0.05)				
Signal_A is red channel for unmethylated signal. Signal_B is green channel for methylated signal.					

2.2.2 Color adjustment

There are red and green color signal channels in type II probes. Red color indicates unmethylated probe and the intensity of the red light is Signal_A. Green color indicates methylated probe and the intensity of the green light is Signal_B. The color adjustment utilizes information from control profile (an example file format as in Figure 9) for Infinium450K, which was also the product from GenomeStudio™.

For each sample (eg. 6164655096_R01C01 as the sample identification), the red and green lights have a baseline value, which denoted as “NEGATIVE” in the row Index 5 in Figure 9. We used N_r and N_g for red and green lights respectively, in this example, $N_r = 119.2350$ and $N_g = 239.7917$. The modified unmethylated and methylated signals are:

$$SignalA' = Signal_A - N_r \quad \text{and} \quad (2-1)$$

$$SignalB' = Signal_B - N_g$$

After background subtraction for red and green signals in Equation (2-1), we used control values measured by the GenomeStudio software to balance the green and red signals. The control values for red and green signals are different. For each sample, the control of green signal is the mean value of A and T (C_g), and red signal is the mean value of G and C (C_r). In Figure 9, the rows at index number 7 to10 are normalized intensities of bases A, C, G and T, and we can calculate $C_g = 414.39575$, $C_r = 537.44315$ for the example sample. The ratio of red and green color channel is, $\gamma = C_r/C_g$. In the step of “color adjustment”, the green signals are normalized to the red channel signals by multiplying γ , the ratio of control values between red and green channels, while red signals keep the same. The adjusted methylated and unmethylated signals are:

$$Adj_SignalA = SignalA' \text{ and } Adj_SignalB = SignalB' * \gamma \quad (2-2)$$

		6164655096_R01C01			6164655096_R01C01	
Index	TargetID	Signal_Grn	Signal_Red	Detection Pval	Signal_Grn	Signal_Red
1	BISULFITE CONVERSION	3536.9170	6412.0000	0.0000	3850.4170	7112.0000
2	BISULFITE CONVERSION	981.7500	24901.0000	0.0000	1439.5000	25000.0000
3	EXTENSION	17611.0000	30062.2500	0.0000	15475.0000	30000.0000
4	HYBRIDIZATION	19539.0000	1497.0000	0.0200	17498.0000	1500.0000
5	NEGATIVE	119.2350	239.7917	0.0000	108.8767	200.0000
6	NON-POLYMORPHIC	6533.5000	10243.2500	0.0000	6497.5000	10000.0000
7	NORM_A	577.9063	6277.8130	0.0000	660.7188	9000.0000
8	NORM_C	3665.7710	407.5738	0.0000	5733.0490	4000.0000
9	NORM_G	5037.8750	667.3125	0.0000	6768.2190	6000.0000
10	NORM_T	250.8852	4579.3280	0.0000	309.7541	8000.0000
11	SPECIFICITY I	2371.5830	5579.2500	0.0000	2595.5830	6000.0000
12	SPECIFICITY II	577.6667	22905.3300	0.0000	411.6667	23000.0000
13	STAINING	7154.5000	14208.2500	0.0000	6259.5000	14000.0000
14	TARGET REMOVAL	387.5000	1292.0000	0.0000	314.0000	1300.0000
		AT	CG		AT	
Average of norm ATCG		414.39575	537.44315		485.2365	

Figure 9 Extraction of red and green signal from control profile.

After color adjustment, β -values were calculated based on adjusted methylated and unmethylated signals for further analysis. β -values are the ratio of the methylated probe intensity to the overall intensity, β -value for an i^{th} interrogated CpG site is defined as:

$$\begin{aligned} \beta_i &= \frac{\max(Meth_i, 0)}{\max(Meth_i, 0) + \max(Unmeth_i, 0) + \alpha} \\ &= \frac{\max(Adj_SignalB_i, 0)}{\max(Adj_SignalA_i, 0) + \max(Adj_SignalB_i, 0) + \alpha} \end{aligned} \quad (2-3)$$

where $Meth_i$ and $Unmeth_i$ are the intensities measured by the i^{th} methylated (measured by adjusted green signal, $Adj_SignalB_i$) and unmethylated (measured by adjusted red signal, $Adj_SignalA_i$) probes respectively, averaged over the replicate beads. α is a constant offset, by default 100. Therefore β -values are ranged between 0 and 1. β value equals to 0 representing no methylation and 1 representing full methylation. It is usually presented as percentage value.

2.2.3 Type I and Type II correction

The idea of the “peak based correction” [81] was implemented to rescale the type II range to that of type I, as in Figure 10.

Firstly, the M values, or log ratio of beta values, which were introduced to measure methylation [175] because they are less sensitive to signal noises, were calculated on top of beta values, where $M = \log_2 \frac{\beta}{1-\beta}$ and hypermethylation is $M > 0$ and hypomethylation is $M < 0$.

Secondly the densities of type I and type II were calculated. By the nature of bimodal of the signal distribution, there are two peaks of each

$$\sigma_M = \operatorname{argmax}_{M>0}(\operatorname{density}(M)), \text{ and} \quad (2-4)$$

$$\sigma_U = \operatorname{argmax}_{M<0}(\operatorname{density}(M))$$

$$AdjM_{M>0} = M * \frac{\sigma_{M1}}{\sigma_{M2}} \text{ and } AdjM_{M<0} = M * \frac{\sigma_{U1}}{\sigma_{U2}} \quad (2-5)$$

The corrected *beta* value is $\beta = \frac{2^M}{2^M + 1}$.

Figure 10 The procedure for type I and type II correction. (A) The density plots for beta values of type I and type II, x-axis is beta values, y-axis is density. (B) The converted M values of type I and type II, x-axis is M values, and y-axis is the density. S_U and S_M are the peak maximums of M density, while σ_U and σ_M are distances corresponding to peaks S_U and S_M . (C) The density plots for beta values after type I and type II correction. The left column is for type I and right for type II in all rows. After processing, type I keeps the same and type II is adjusted according to the distribution of type I (chart is adopted from supplementary material in [81].)

2.3 Concordance study between Infinium450K and RRBS in clinical samples

2.3.1 Experimental design

DNA methylation profiles from seven umbilical cord samples were measured in Infinium450K and RRBS methods respectively. The concordance and comparison study design is shown in Figure 12. Two major steps were conducted on the DNA methylation profiles across 450K and RRBS platforms. The first step was the concordance study of the raw data between two platforms. The second step was the comparison study of processed 450K data by the pipeline with the RRBS data.

The seven umbilical cords are part of the GUSTO birth cohort study [176]. The babies were all of Chinese ethnic origin and males. They were conceived without assisted conception techniques. They were within the normal birth weight range (2734g ~ 3565g) and had gestational ages classified as near-term (37 weeks ~ 39 weeks and 6 days). The maternal ages varied between 26-36 years. Sample characteristics were shown in Table 3.

2.3.2 RRBS genomic mapping

In RRBS sequences, typically, genomic DNA was digested by the restriction endonuclease MspI in RRBS technology [19]. As shown in Figure 11, this is followed by end-repair, A-tailing, adaptor ligation and finally bisulfite conversion. Often, the library is also size-selected for fragments between 40 and 220bp in length.

Figure 11 An overview of RRBS technology.

Quality control of sequence reads was performed and methylation calls were calculated as percentage value by the count of cytosine divided by the sum of the count of

cytosine and thymine [54]. Similar to sequence mapping as the other NGS technologies, the major steps for RRBS mapping including below,

1. Quality control (QC) of sequence reads was performed by **FastQC** (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>).
2. Adaptive quality and adapter trimming with **Trim Galore** (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The low-quality base calls were trimmed by the option of “- - phred” score. The option of “- - rrbs” works for RRBS sequences from DNA maternal digested with MspI enzyme. Other options like “- - paired-end” allows paired FastQ files. The current version of Trim Galore wrapped **FastQC** as an option so step1 can also be done within step2 by **Trim Galore**.
3. Sequence read mapping and cytosine methylation status determination is done by **Bismark** (<http://www.bioinformatics.babraham.ac.uk/projects/bismark>). **Bismark** calls **Bowtie/Bowtie2** for sequence alignment and a function of **bismark_methylation_extractor** produced percentage methylation based on the equation that $\% \text{ methylation (context)} = 100 * \text{methylated Cs (context)} / (\text{methylated Cs (context)} + \text{unmethylated Cs (context)})$.

2.3.3 Basic terminologies in next generation sequencing (NGS): depth and coverage

Sequencing depth (also known as number of reads, Nreads) describes the number of times that a given nucleotide in the genome has been read in an experiment. After sequence alignment, one nucleotide can have more than one reads and the number of reads helps the confidence of measurement, as shown in Table 2, position 3 nucleotide “A” was read 3 times, we denoted it as 3X. Our data showed the number of reads can reach beyond 1000X in RRBS and MC-Seq [18] technologies.

Table 2 Illustration on the definition of read depth of sequence in next generation sequence.

Position	1	2	3	4	5	6	7	8	9	10
read 1	G	A	A	T	C	C	G	A	T	C
read 2		A	A	T	C	C	G	A	T	C
read 3					C	C	G	A	T	
read 4			A	T	C	C	G	A		
Depth	1	2	3	3	4	4	4	4	3	2

In many cases, a given NGS technology won't encompass the entirety of a target genome, because certain regions are difficult to sequence and not of interest.

Therefore, the depth of coverage in genome is an important parameter in evaluating the technology. We used percentage to count the coverage using read length divided by size of target genome. For example, there are 28 million CpGs within the human genome (hg19) [89], MC-Seq [18] was designed to assay up to 3.2 million CpGs within the target regions, the coverage of CpG in epigenome is 11%.

2.3.4 CpG coverage on heterogeneous clinical samples by RRBS and Infinium450K

Table 3 Demographic characteristics of the seven samples

Gender	7 males
Median BW (g)	3352.00
Mean BW (+/- sd) (g)	3325.72(+/- 329.33)
Median GA (w)	39.43
Mean GA (+/- sd) (w)	38.88 (+/- 1.20)
Ethnic Group	Chinese

DNA methylation profiles were extracted from 450K machine, and passed through quality control. Only 1.9% of the data which failed in detection $p\text{-value} < 0.05$ was removed, leaving 129,236 Type I probes and 347,123 Type II probes for subsequent analysis. The number of CpGs measured by both methodologies in the seven common samples was compared in Table 4. In RRBS, the count of CpGs varied by the number of read depth, termed as *Nreads*, and the bigger the *Nreads*, the less amount of CpGs

could be extracted. The mean value of CpG count at any read density ($N_{reads} > 0$) in the seven samples by RRBS was 6,052,399 (standard deviation, $SD = 496,344$) and the average number of CpGs shared by both 450K and RRBS was 104,335 ($SD = 8,602$). When the criterion of $N_{reads} \geq 4$ was applied to the RRBS, the average number of CpGs covered was reduced to 2,808,469 ($SD = 167,495$) and the average number of CpGs overlapped by both was 64,951 ($SD = 3,911$).

Table 4 Number of CpGs extracted by RRBS and shared with Infinium 450K across the seven samples.

Sample	Number of CpGs extracted by RRBS	Number of CpGs overlapped with Infinium 450K	Number of CpGs covered by RRBS filtered by $N_{reads} \geq 4$	Number of CpGs overlapped with Infinium 450K filtered by $N_{reads} \geq 4$	Pearson R (RRBS $N_{reads} \geq 4$)	Spearman R (RRBS $N_{reads} \geq 4$)
1	6,715,332	113,571	2,846,600	65,702	0.96	0.82
2	6,658,349	114,336	3,027,004	69,041	0.96	0.83
3	6,250,352	109,113	2,990,478	68,749	0.96	0.84
4	5,838,715	101,765	2,628,619	60,641	0.95	0.83
5	5,462,059	93,562	2,607,467	61,366	0.96	0.82
6	5,817,127	93,693	2,703,504	60,815	0.96	0.81
7	5,624,859	104,305	2,855,611	68,346	0.94	0.82
Average	6,052,399	104,335	2,808,469	64,951		
Stdev	496,344	8,602	167,495	3,911		
Total	42,366,793	73,0345	19,659,283	454,660	0.96	0.83

2.3.5 The RRBS read density cut-off was set at 4 or more reads

RRBS data had a range of read densities from 1-45,972 (Figure 13A). Analysis of the correlation between the raw Infinium 450K values and RRBS measured methylation values at the different read cut-offs suggested concordance climbs quickly between 1-4 reads (Figure 13B). In Figure 13B, the number of CpGs extracted from RRBS were indicated in red stars, which were quickly decreased with the increasing number of reads, as the same as Figure 13A. The correlation between Infinium 450K and RRBS were estimated using Pearson R (black squares in Figure 13B) and Spearman ranked R (black circles in Figure 13B). The cut-off of $N_{reads} = 4$ seemed acceptable as when c , Pearson's $R = 0.96$ and Spearman $R = 0.83$ whilst retaining a dataset of 454,660 CpGs for downstream analysis (the dataset is the aggregate of CpGs measured by both Infinium and RRBS across all seven samples). The four or more reads cut-off

is at the lower end of the range used by other researchers, for instance Bock et al [177] used five or more reads, and Meissner et al (2005) [19] used ten or more reads.

Figure 13 Optimal selection of number of reads quality cut-off in RRBS data. (A) Frequency distribution of number of reads in RRBS data. X-axis is truncated at 200 reads but maximum in data are 45,972. (B) Correlation between Infinium 450K and RRBS values, estimated using Pearson R (black squares) and Spearman R (black circles). Numbers of CpGs remaining after read cut-off in RRBS are indicated in red stars.

2.3.6 DMA methylation by Infinium 450K and RRBS had a bimodal frequency distribution

Both the Infinium 450K and RRBS datasets for the 454,660 shared CpGs had a bimodal frequency distribution, approaching a beta distribution, as shown in Figure 14.

The RRBS distribution was irregular and “spiky” due to tied values produced by the reduced resolution of continuous methylation values at low read depth. In the RRBS frequency distribution the peaks were located at 0-0.25% and 99.75-100%. In the Infinium 450K type I assay data the peaks were at 1.75-2.25% and 97.25-97.75%. The 450K data shows a shorter range for the type II assays with peaks at 5.25-5.75% and 85.75-86.25%. This phenomenon has previously been observed [81, 178].

Figure 14 Histogram of percentage methylation value frequency in (A) RRBS data (B) Infinium 450K data for the 454,660 CpGs covered by both technologies. Data was plotted into 201 bins stepped by 0.5 between 0% and 100%. For Infinium 450K, type I assay data was indicated in red and type II in blue.

2.3.7 An approximately linear relationship was observed between Infinium 450K and RRBS data, which was stable across samples

The raw beta values were compared to the RRBS data with $Nreads \geq 4$. An approximately linear relationship was observed and the slope of the fitted line was 0.83 (Figure 15A). It showed a scatter plot of percentage methylation values from RRBS (x-axis) and Infinium 450K (y-axis) over a density cloud. Density cloud is generated by the smoothed two-dimensional histogram using 50 equally spaced bins in both directions. A random selection of 2000 data from type I probes is plotted as red dots and another random selection of 2000 data from type II probes is plotted as blue dots. As an alternative method to quantify agreement between the methylome measures, Bland-Altman plots [179] were produced in Figure 15B. Bland-Altman plot displays for the average measurement from both 450K and RRBS on the x-axis, the y-axis shows the average differences between the two. Data from type I assay is shown in red, type II is shown in blue. It revealed a greater discordance at intermediate methylation levels i.e. 20-80%, an observation also made by Roessler et al. [178]. Also noticeable in the

plots is an upward trend in the type II assay values suggestive divergence of zero for the two platforms at high methylation values.

Pearson's and Spearman correlation statistics were used to study the concordance of both technologies, however neither of them are appropriate for bimodal data and results could be misleading. Pearson's correlation assumes a joint normal distribution [180], an assumption that is obviously violated by non-linear methylation data and so probably overestimates the correlation. Spearman correlation is sensitive to tied values in the data [181], which is a feature of the "spiky" distribution of RRBS data and so probably underestimates the correlation. Also, both methods explicitly assume a linear relationship between the two variables. Thus, maximal Information Coefficient (MIC) [56] was used to account for the non-linear dependency between 450K and RRBS data. The MIC (strength) value lies between 0 to 1. A perfect deterministic relation has a MIC value of 1, while totally independent relation has a value of 0 and noisy relations lie somewhere in between. The strength of the interdependency between RRBS ($N_{reads} \geq 4$) and 450K data is 0.81. MIC methodology does not assume a joint normal distribution or rank the data points, however the power and robustness under different of the method is under debate [182].

A

B

Figure 15 The concordance of raw Infinium 450K data vs. RRBS data. (A) A scatter plot of percentage methylation values from RRBS (x-axis) and Infinium 450K (y-axis) over a density cloud. (B) Bland-Altman plot for raw Infinium data compared with RRBS data using $N_{reads} \geq 4$.

2.4 Comparison study between processed Infinium 450K and RRBS

2.4.1 The concordance of Infinium 450K with RRBS are improved by step-wise processing

To provide a quantitative measure of agreement between the two technologies, the number of CpGs with values from 450K and RRBS that fall within 20%, 10% and 5% of each other were counted at each stage of the 450K processing. A stepwise improvement in agreement was achieved at all levels in Table 5, with the most dramatic improvement resulting from the type II assay adjustment to the range of type I range.

Table 5 Number of CpGs showing agreement within 5%, 10% and 20% ranges, between Infinium 450K and RRBS ($Nreads \geq 4$) data, at different levels of Infinium processing.

Difference Range (n=454,660)	within 20%		within 10%		within 5%		Spearman's Rank R	Pearson's R^2	Slope
Probe Type	I	II	I	II	I	II			
Raw Data	196,937	217,721	172,003	161,520	132,267	75,144	0.83	0.92	0.83
	93%	90%	81%	67%	62%	31%	p<0.001	p<0.001	
After color adjust-ment	197,089	221,887	173,872	165,158	139,888	73,510	0.83	0.92	0.87
	93%	92%	82%	68%	66%	30%	p<0.001	p<0.001	
After color and Type II adjust-ment	197,089	225,828	173,872	192,004	139,888	144,070	0.83	0.93	0.92
	93%	93%	82%	79%	68%	60%	p<0.001	p<0.001	
After color, Type II and QN	197,444	226,386	175,211	195,114	143,022	149,216	0.83	0.93	0.93
	93%	94%	82%	81%	67%	62%	p<0.001	p<0.001	

Next, to determine if the individual data processing steps were additive in their effects, we applied each independently to the raw data. Type II adjustment alone increases the number of type II assays within 5% of the RRBS data ($Nreads \geq 4$) from 75,144 (31%) to 144,968 (60%). This is by far the greatest improvement in concordance resulting from a single processing step. Type I assays are not affected by type II adjustment. The processing step which most improves the agreement of the type I assays within 5% of the RRBS data ($Nreads \geq 4$) is color adjustment (139,888 (66%) compared to 132,267 (62%) in the raw data).

A

B

Figure 16 Processed Infinium 450K data vs. RRBS data (A) A scatter plot of percentage methylation values from RRBS (x-axis) and Infinium 450K (y-axis) over a density cloud. (B) Bland-Altman plot for processed Infinium 450K data compared with RRBS (Nreads ≥ 4).

2.4.2 Compare the proposed Infinium 450K processing with SWAN

Very recently an alternative solution for typeI/typeII assay distribution adjustment called SWAN (subset-quartile within array normalization) was published [183]. It is an interesting point of difference that whist Dedeurwaerder et al assumed the type I distribution is “correct” and fit the type II value distribution to that of type I, Maksimovic et al adjusted both type I and type II values in SWAN. For this dataset, the SWAN processed data agreed less with RRBS than data processed with the full processing pipeline described in Table 6. From the pseudo codes of both peak-based

correction (PSB) and subset-quartile within array normalization (SWAN) in Table 6, PSB results are “type-I-like” distribution while SWAN returns a more “type-II-like” distribution. The comparison results on both methods in the sample DNA methylation profiles in this study demonstrated that the PSB method is more appropriate than SWAN with the consistent higher levels of agreements in different range of differences, as in Table 7. The possible reason might be because the type I assays are quite accurate, whilst the more numerous type II assays are less so. As Infinium II assays outnumber Infinium I, for example, the portion of Infinium II assays and Infinium I is 73% vs. 27% in the processed probes in this study, type II adjustment results in the greatest part in concordance.

Table 6 The pseudo codes of peak-based correction (PBC) and SWAN

PBC <ul style="list-style-type: none"> • Calculate M values and densities of methylated and unmethylated signals • Correct the density distribution of type II towards that of type I 	SWAN <ul style="list-style-type: none"> • Order the signal into three quartiles: unmethylated, hemi-methylated and methylated. • Replace the signal with the mean values of a subset in each quartile. • Normalize the type I values towards type II.
--	---

Table 7 Number of CpG percentage methylation values showing agreement within 5%, 10% and 20%, between SWAN processed Infinium 450K and RRBS data (Nreads $\geq$ 4), at different percentage methylation value ranges.

Difference Range (n=454,660, Nreads $\geq$ 4)	within 20%		within 10%		within 5%		Spearm an's Rank R	Pears on's R ²	Slope
Probe Type	I	II	I	II	I	II			
Raw Data	93%	90%	81%	67%	62%	31%	0.83	0.92	0.83
After SWAN	93%	93%	81%	73%	60%	38%	0.83	0.93	0.87
After the proposed method	93%	94%	82%	81%	67%	62%	0.83	0.94	0.95

2.4.3 Removal of batch effects contributes to improved concordance of processed Infinium 450K data and RRBS

The proposed pipeline of the Infinium 450K data processing resulted in greater agreement with RRBS data, as shown in Table 5. In order to evaluate batch effects before and after data processing, 72 umbilical cord samples were run across six arrays. This sample set includes the seven samples also studied by RRBS and described above. The number of CpGs with differential methylation values (at $p\text{-value} \leq 0.05$) between the array batches was calculated by ANOVA. In the unprocessed data 163,438 (38%) CpGs were differentially methylated between the arrays. After processing, this number was reduced to 58,532 (14%). After FDR correction at $p\text{-value} \leq 0.2$, the number was 233,650 (55%) before processing and 123,611 (29%) after processing (Figure 17). Therefore, although an array-batch effect was still present and may necessitate batch-effect-removal methodologies such as “COMBAT” [184, 185], it was much reduced by data processing. To determine if processing had reduced variation in all directions (not just batch), the ANOVA test was also applied to gender and gestational age, where there were phenotypes expected to associate with differential methylation ([186, 187] and [188, 189] respectively). For gender, methylation levels at 29,642 (7%) of the CpGs were associated in the raw data and 48,436 (11%) were associated in the processed data. After FDR correction ($p\text{-value} \leq 0.2$) 41,108 (10%) assays were differential before processing and 73,595 (17%) were differential after processing. For gestational age 3,842 (0.9%) CpGs had methylation levels differential for gestational age (<37 weeks or ≥ 37 weeks) in the raw data, compared to 8,454 (2%) probes in the processed data. After FDR correction at $p\text{-value} \leq 0.2$, 104 (0.02%) were differential before processing and 179 (0.04%) were differential after processing. Therefore, one of the factors improving the concordance between the 450K processed data and RRBS data, seems to be the reduction of the batch effects in 450K data. The resulting decrease in batch-related variation appears to have the advantage of increasing power to detect differential methylation associating with biological phenotypes. Further study is warranted on the robustness and biological implications of methylome variation in umbilical cord tissues with these and other phenotypes.

Figure 17 Percentage of assays with methylation values significantly associated with array on which sample was run, gestational age, or gender before (blue bars) and after (red bars) data processing in the set of 72 samples.

2.5 Discussion

After Infinium 450K was introduced in early 2011[81] , studies on this relatively new technology included two major efforts, the comparison between 450K with other technologies or platforms on signal detection and coverage, and the bioinformatics tools attempting to deal with the unique two different array designs in Infinium 450K. The work reported in this thesis was the pioneer work among peers. The concordance study between Infinium 450K and RRBS in clinical samples was new and the idea of combination both color adjustment and two type design correction was novel. After that, the computational challenges in Infinium 450K array attracted a lot of studies with more comprehensive computing packages. A review paper [190] addressed both disadvantages and advantages for existing packages and we also provided the below comparison table, shown in Table 8.

Table 8 Comparison of Infinium 450K data processing methods/pipelines.

Package	SQN[99]	Lumi[191]	BMIQ[192]	minfi[193]	ChAMP[194]	pipeline proposed in this thesis [1]
Raw data processing	N	Y	N	Y	Y	Y
Background / quality control	Y	Y	NA	Y	Y	Y
Color bias adjustment	Y	Y	N	N	N	Y
Type I and II bias correction	Y	N	Y	Y	Y	Y
Across sample normalization	Y	Y	N	Y	Y	Y
N: No. NA: Not applicable. Y: Yes.						

After careful designed evaluation in [195], neither our pipeline or any of the alternatives performed consistently better across different studies. Efforts seem continued with advanced versions of packages such as “mini R package” [193] and integrated analysis pipelines , for example “ChAMP R package” [194]. However, improvement in data preprocessing of 450K array has to be validated across different measurement platforms for same clinical samples. For this perspective, our study excelled among all the others and has been well cited by about twenty research publications in prestigious journals [4, 90, 125, 196-209].

The knowledge about DNA methylation was extended that it is tissue-specific and the effects of heterogeneity of cells in the samples bias the methylation profiles [103]. The extent to small perturbations in cell mixture proportions can register changes in DNA methylation remains, however, largely unknown. Recently there has been increasing interest in statistical methods to exploit the association in cell mixture proportions and DNA methylation variants, in the context of epigenome-wide association studies [102, 104]. However the work may be laborious and expensive as again there is no reference datasets, especially for some tissues such as umbilical cords which the relevant underlying cell types may not be known.

Other challenges include the statistical power in EWAS studies and in-depth biological experiments and replications in much larger data cohorts [159]. The arguments about large sample sizes (for example about 450,000 probes) to detect small effects [210] are compelling for EWAS, and the relevant parameters are even more difficult to predict because DNA methylation based studies are associated with complex environment cofounders. Replication could be potentially more demanding for EWAS because of

the unknown tissue mixture and variable confounding in determining methylation changes, however limited availability of tissue samples and study subjects make the replication harder to achieve. A successful study of *HIF3A* methylation associating with body mass index for western adults [29] firstly demonstrated significant results in 479 samples of a discovery cohort, and then replicated the results in a primary cohort of 339 and the second cohort of 1789 samples.

In conclusion, this study was the first endeavor to show methylome has very good reproducibility between the 450K and RRBS platforms in seven heterogeneous clinical samples. The pipeline combined two major steps, and implemented methodologies proposed by others [15, 81, 191]. It is shown that linearity of 450K data against independent techniques can be vastly improved. After Infinium450K was well studied and data processing was well established, we extended the exploration on a new next generation sequencing technology, Methyl Capture Seq (MC-Seq), in the next chapter.

3 Comparing Methyl Capture Seq technology with Infinium450K for methylome analysis in clinical samples

3.1 Introduction

Methods available to interrogate DNA methylation in clinical samples at single base resolution can be broadly classified into two categories, microarray based and next generation sequencing (NGS) based. These two categories are elaborated in Table 8. Infinium 450K has thus far been the most widely used method in epigenome wide association studies (EWAS) due to its low cost, modest DNA requirement and significantly reduced sample processing time, making high-throughput processing of large numbers of clinical samples possible [159, 189]. However, microarray based platforms are limited by the number of probes designed in genomic regions.

Of the different NGS based technologies, whole genome bisulfite sequencing (WGBS) [211-213] is regarded as the gold standard because it provides the highest genomic coverage; however, its substantial cost and processing time renders it unfeasible in EWAS when large numbers of samples have to be assayed. To reduce the cost and processing time, the other two alternative to WGBS are reduced-representation bisulfite sequencing (RRBS) [214, 215] and methylated DNA immunoprecipitation sequencing (MeDIP-Seq) [216-218]. RRBS platforms use cytosine-methylation-specific restriction enzyme digestion, while MeDIP-Seq utilizes antibodies to extract methylated-cytosine-containing DNA fragments after DNA fragmentation. Compared to RRBS platforms, MeDIP-Seq allows broader coverage of the genome but fails to provide methylation measures at base-pair resolution. Both platforms, however, introduce bias towards CpG-rich repeats [85, 177]. Moreover, both technologies are limited in regions surveyed due to the available restriction digestion enzymes or antibodies. Another affinity-enrichment platform, Methyl Capture Seq (MC-Seq) utilizes target-specific bait sequences, circumventing this restriction through bait design and allows for epigenome-wide interrogation of specific genomic loci of physiological and clinical interest. Hence, MC-Seq is an attractive cost-effective NGS alternative which helps to overcome the limitations of low genome coverage (Infinium 450K), high cost and processing time (WGBS), while avoiding overrepresentation of repeated (RRBS) and methylated regions (MeDIP-Seq) [2]. The MC-Seq platform that

we studied was from SureSelect Human Methyl-Seq panel (Agilent Technologies) in this chapter.

Table 9 Platforms for DNA methylation profiling

Technology		Features (Pros)	Features (Cons)
Microarray based	Infinium® HumanMethylation450 BeadChip (Infinium450K), using bisulfite conversion and hybridization.	Affordable, repeatable, stable, feasible for large samples, the most popular technology in EWAS.	Low CpG coverage
Next generation sequencing based	Whole genome bisulfite sequencing (WGBS)	Best coverage, gold standard	Extremely costly and time consuming, not practical for large samples
	Methylation Dependent ImmunoPrecipitation Sequencing (MeDIP-Seq)	Can interrogate CpGs in repetitive elements	Dependent on quality and cross-reactivity of antibodies. Not base-pair resolution
	Reduced representation bisulfite sequencing (RRBS)	The design of CG-rich parts can reduce the amount of sequencing required	Dependent on enzyme digestion.
	MethylCapture-Seq (MC-Seq) from Agilent Technologies (Sureselect Methyl-Seq) and Roche NimbleGene (SeqCap)	Reduced cost, allowed more samples, more CpGs	With designed targets, not full coverage in genome wide

The performance of MC-Seq has been examined in a few studies [219] [18] [220], however, there are a few major issues that need to be addressed when MC-Seq approach is considered for studies that involve association study of DNA methylation and phenotype. First, no study has yet examined the performance of MC-Seq in investigating inter-individual variation from routinely used clinical samples such as buccal swabs. Second, the minimum amount of DNA required for MC-Seq has not been clearly stated. Finally, there is, as yet, no direct comparison between MC-Seq with Infinium 450K.

The concordance study and comparison analysis between processed 450K and RRBS ensured our understanding of 450K, in the previous chapter. In this chapter, we compared MC-Seq and 450K using seven buccal epithelium samples. We firstly evaluated the technical performance of MC-Seq in human buccal DNA samples. Secondly, we provided a comprehensive comparison of 450K and MC-Seq, across key functional genomic regions in the context of their coverage and concordance of

methylation calls in clinical samples. Lastly, as a proof-of-principle, in a small-sized sample set, we examined the utility of MC-Seq in studying inter-individual variation and demonstrated that methylation values from MC-Seq can distinguish subjects from different ethnic groups [2].

3.2 Material and methods

3.2.1 Objectives and study design

Figure 18 Study design and samples for comparison of MC-Seq and Infinium450K.

Seven buccal epithelium tissues were collected with informed consent from seven volunteers comprising of three different ethnicities as shown in Figure 18. There were two Indian (F1_Ind and F2_Ind), one Caucasian (F3_Cau) and four Chinese subjects (F4_Chi, F5_Chi, F6_Chi and M7_Chi). Among them, six were female (indicated by F in the prefix) and one was male (M7_Chi). DNA methylation was profiled using Infinium 450K and MC-Seq from seven same buccal epithelial DNA samples. 750 ng and 3µg of DNA was used for DNA profiling using Infinium 450k and MC-Seq respectively. We have included four randomly selected samples (F1_Ind, F3_Cau, F4_Chi and M7_Chi) as replicates. Besides, DNA profiling were repeated for two samples (F4_Chi and M7_Chi) using as little as 1 µg DNA samples.

The first comparison, as “*comparison1*” shown in Figure 18 was designed for evaluating the technical performance of MC-Seq in human buccal samples by comparing samples and their replicates. The second comparison was designed to

examine DNA methylation between 450K and MC-Seq, based on the same seven buccal samples, shown as “*comparison2*” in Figure 18. We have also evaluated the performance of different amount of DNA (3µg vs. 1µg) for two other samples, shown as “*comparison3*” in Figure 18.

3.2.2 Processing of MC-Seq data and Infinium450K data

The details of processing of MC-Seq data was shown in Table 10. The latest version of *Trim Galore!* (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) (v0.4.2) wrapped with FastQC function was used to access the basic statistics and quality of the data, and then trim/remove reads with phred score < 30 and/or read length < 70 bases. Quality-trimmed paired-end reads were aligned to the reference human genome (hg19) using *Bismark* (v0.13.0) [221] and *bowtie2* (v2.2.4) [222] using default parameters. Duplicated reads were removed using *Bismark deduplication* tool. Methylation values were called using *Bismark methylation extractor*. Only methylation sites that were on the negative strand were retained. Sites that were on-target were determined using the *bedtools intersect* (v2.24.0) command.

Infinium 450K data analysis was carried out using the proposed pipeline [1], as described in Chapter 2. Methylation beta values were adjusted such that the range of Type II probes is similar to the range of Type 1 probes for downstream comparison [81].

Table 10 DNA methylation processing for MC-seq sequences

Process flowchart	Software and parameter	Source
<pre> graph TD A[Raw Sequences (*fastq files)] --> B([Trim Galore]) B --> C[Trimmed sequence (*fastq files)] C --> D([Sequence alignment by Bismark]) D --> E[Aligned sequences (*.sam file)] E --> F([Duplication removal by Bismark]) F --> G([Methylation call calculation by Bismark (methylation extractor)]) G --> H[methylation calls] </pre>	<p>trim_galore --quality 30 --phred33 --fastqc --length 70 --adapter N --fastqc --e 0 --stringency 2</p>	<p>Trim Galore!: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/</p>
	<p>bismark --bowtie2 --phred33-quals --path_to_bowtie bowtie2-2.2.4 --old_flag</p>	<p>Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2</p>
	<p>deduplicate_bismark -p</p>	<p>Bismark v0.13.0: http://www.bioinformatics.babraham.ac.uk/projects/bismark</p>
	<p>bismark_methylation_extractor -p --no_overlap - -cytosine_report --remove_spaces --bedGraph</p>	

3.2.3 Annotation of CpG sites and comparison of genomic function and CpG content for MC-Seq and Infinium450K

To annotate CpG sites assayed by MC-Seq (SureSelect Human Methyl-Seq panel), the full list of 28 million CpGs within the human genome (hg19) was downloaded from Saffery *et. al.* [89]. From this list of 28 million CpGs, we determined CpGs that were (1) within the bait design target regions and (2) within 200 bases of bait design target regions. These CpGs, which represent the maximum possible CpGs that can be captured by MC-Seq, were then annotated in terms of their genomic features

(promoter, 5'-UTR, exon, intron, 3'-UTR, TTS and intergenic as shown in Figure 4, Chapter 1) and CpG content (island, shores, shelves, open seas, Figure 4, Chapter 1).

Based on the CpG content definition illustrated in in Figure 4, Chapter 1, genomic coordinates (hg 19) of CpG islands were downloaded from UCSC genome browser. CpG shores, shelves and open seas were generated and the annotation was generated following the steps shown in Figure 19.

Figure 19 Annotation of CpG content.

CpG genomic features (promoter, 5'-UTR, exon, intron, 3'-UTR, TTS and intergenic) were annotated using *Homer annotatePeaks* function (hg19) [223].

We also determined the number of CpG sites located near SNPs for both platforms. The dbsnp142Common table was downloaded from UCSC genome browser and we further restricted the analysis to SNPs with “rs” numbers, of high quality (weight =1) and that were reported by the 1000 genomes project to have a minor allele frequency of at least 1%. We then determined the number of CpGs that had SNPs within 1 bp or 10 bp of the CpG site using *bedtools intersect* command.

3.2.4 Evaluation of technical variation and performance using lower quantities of DNA for MC-Seq

To examine technical performance of MC-Seq on 3µg of genomic DNA (“comparison1” in Figure 18), and to compare the performance of MC-Seq using 3µg and 1µg of genomic DNA (“comparison3” in Figure 18), we first assessed reproducibility of the data with Pearson correlation and scatterplots, using methylation

sites covered with at least 10X reads coverage. Second, we calculated absolute differences in methylation values between two comparable variables as a function of reads coverage. Third, hierarchical clustering analysis was used by methylation sites covered with at least 10X reads coverage.

The comparative analysis of MC-Seq and Infinium450K (“*comparison3*” in Figure 18) were conducted by two metrics, (i) methylation values and (ii) genome coverage.

Firstly, we compared methylation values from Infinium450K and MC-Seq at the same CpG sites over different MC-Seq reads coverage, using Pearson correlation, scatterplots and distribution plots of the methylation values. Secondly, we compared the coverage of the human epigenome by both platforms in terms of the proportions/numbers of regions covered and the density of coverage.

Lastly, to illustrate the utility of both methods for distinguishing inter-individual variation, unsupervised hierarchical clustering analysis was performed, using most variably methylated CpGs (interquartile range > 20%). For Infinium450K, methylation sites from sex chromosomes and probes known to be cross-hybridizing [224, 225] were excluded in the hierarchical clustering analysis. For MC-Seq, only probes from autosomal chromosomes and those with a minimum read coverage of 30X were used. Clustering was performed using Euclidean distance and “ward.D” method in R.

3.3 Results

3.3.1 Overview of MC-Seq analysis of human buccal epithelial samples

There are 28 million CpGs in the human epigenome (hg19). MC-Seq was designed to assay up to 3.2 million CpGs within the target regions (11% of all CpGs in epigenome). By contrast Infinium450K arrays assay only 1.7% of the CpGs present in the human epigenome. MC-Seq platform includes 395,817 (82%) of the 482,421 CpGs on Infinium450K, as described in Figure 20.

Figure 20 Overview of MC-Seq and Infinium450K in terms of number of CpG probes.

MC-Seq was designed to capture 3.2 million CpGs within its baits target region, of which, 25%, 3%, 15%, 34%, 2%, 2% and 19% of the CpGs belonged to promoter, 5'-UTR, exon, intron, 3'-UTR, TTS and intergenic regions (Figure 21A), while 51%, 19%, 5% and 25% belonged to CpG islands, shores, shelves and open seas, respectively (Figure 21B). With 200 bases flanking both sides of the target regions, MC-Seq would capture 4.8 million CpGs, with similar CpG content distribution and similar genomic features distribution. In our buccal epithelium samples, we detected an average of 2.6 million CpGs with at least 10X coverage, the majority of which (89%) were within the target region. The observed distribution (from sample F1_Ind) of CpGs within genomic features was similar to the expected.

Figure 21 Overview of MC-Seq from genomic features (A) and CpG content (B) for its target region, 200bp flanking to target region and observed samples.

The DNA methylome of seven buccal epithelium samples were generated using MC-Seq with a capture-then-bisulfite-convert approach with 3µg of genomic DNA. Mapping efficiency, sequence duplication rates and sequence bait specificities are summarized in Table 11. On an average 38 million pair-end reads were generated per sample, of which 34 million aligned uniquely to the bisulfite-converted human reference genome (hg19/GRCH37). 28 million aligned reads remained after removal of duplicated reads; 93% and 97% of the reads were found within the target region or within 200 bases of the target region, respectively. Bait capture efficiency was high as reads within 200 bases of targeted regions were observed to be >25-fold more abundant than the genomic background.

Table 11 Summary of sequence alignment and duplicate rates for 7 buccal epithelium samples from MC-Seq.

Item	Average
Raw sequence reads	44,103,266
Sequence pairs analysed in total	≈ 38,000,000
Number of paired-end alignments with a unique best hit	≈ 34,000,000
Mapping efficiency (%)	91.7
Duplicate (%)	11.87
Sequence pairs after removing duplicate	≈ 28,000,000
Reads in targeted region (%)	93.09
Reads in targeted regions 200 bp (%)	96.52

3.3.2 Reproducibility of MC-Seq

To investigate the technical variation of MC-Seq in buccal samples, four of the seven samples were investigated in duplicate based the same quantities of 3µg DNA. Mapping efficiency, sequence duplication rates and sequence bait specificities for the four replicate samples of F1_Ind, F3_Cau, F4_Chi and M7_Chi. Sample-based Pearson correlation between replicates was high (R = 0.9812 on sample F1_Ind and its replicate, Figure 22A). Replicates clustered together in hierarchical clustering (Figure 22B).

3.3.3 Performance of MC-Seq at 3 μ g and 1 μ g was similar

To investigate the performance of MC-Seq in buccal samples using lower quantity of genomic DNA, two (F5_Chi and M7_Chi) of the seven samples were repeated with MC-Seq using 1 μ g of genomic DNA. Mapping efficiency, sequence duplication rates and sequence bait specificities for the two samples performed using 1 μ g of genomic DNA are summarized in Table 11 and are similar to those performed using 3 μ g of genomic DNA (Figure 23).

Figure 22 MC-Seq technical comparison between seven samples and four replicates of same quantities of genomic DNA. (A) methylation comparison (B) clustering analysis.

Figure 23 MC-Seq technical comparison between samples and replicates of different quantities of genomic DNA (3 μ g and 1 μ g). (A) methylation comparison (B) clustering analysis.

The comparison of MC-Seq at 1 μ g and 3 μ g (Figure 23) mirrors the comparison between technical replicates at 3 μ g (Figure 22), with similar error rates. Correlation of methylation values derived using 1 μ g and 3 μ g of genomic DNA was high ($R=0.9842$ with minimum of 10X reads coverage, Figure 23A). Furthermore, hierarchical

clustering analysis performed on the four samples (F5_Chi at 1 μ g and 3 μ g and M7_Chi at 1 μ g and 3 μ g) showed that the corresponding 1 μ g and 3 μ g samples clustered together (Figure 23B).

3.3.4 Methylation values from MC-Seq and Infinium450K were highly correlated and both gave a bimodal distribution.

To compare MC-Seq and Infinium450K, we accessed the data distribution and calculated the sample-based correlation between the methylation values. Figure 24 A shows that the number of CpGs detected decreased significantly (Number of CpGs = 2.6, 1.3 and 0.6 million at 10X, 30X and 50X, respectively) with increasing coverage of reads in MC-Seq, while the sample-based correlation between methylation values from MC-Seq and Infinium450K was high and increased slightly with increasing coverage ($R=0.9776$, 0.9819 and 0.9840 at 10X, 30X and 50X, respectively, as in Figure 24 B and C. The distribution of methylation values from both platforms followed a bimodal distribution (Figure 24 D).

Figure 24 Methylation values from MC-Seq and Infinium450K were highly correlated and both gave a bimodal distribution. (A) Observed number of CpGs (vertical axis) from MC-Seq for one sample at different MC-Seq reads coverage (horizontal axis). (B) Pearson correlation (vertical axis) between methylation values from MC-Seq and Infinium450K at the same CpG sites, at different MC-Seq reads coverage (horizontal axis) for one sample. (C) Scatterplot of methylation values from MC-Seq ($\geq 10X$, vertical axis) and Infinium450K (horizontal axis) at the same CpG sites for one sample. (D) Distribution of methylation values for all CpGs from MC-Seq ($\geq 10X$, solid line) and Infinium450K (dotted line) for one sample.

3.3.5 Methylation values from both MC-Seq and Infinium450K distinguish inter-individual variation.

A key interest in EWAS is to identify methylation sites that show appreciable inter-individual variation with the outcome/phenotype of interest. For methylation array data, the comparison of hierarchical and non-hierarchical clustering methods did not show any single method consistently outperformed others [226]. As a proof-of-principle to demonstrate that MC-Seq can detect significant inter-individual variation that is related to phenotype, we clustered the seven multi-ethnic samples using unsupervised hierarchical clustering. Both Infinium450K and MC-Seq achieved the same clustering results which were based on ethnicity group (shown in Figure 25). As it has been previously reported that the most variable CpGs can be influenced by genotype [4], we performed clustering analysis using only the most variable CpGs (CpG sites with interquartile range $> 20\%$). We found, for the 7880 most variable CpGs in MC-Seq, 4393 (56%) and 4712 (60%) were located within 1bp and 10 bp of SNP.

Figure 25 Hierarchical clustering analysis of MC-Seq values showed clinical samples clustered by ethnicity.

3.3.6 MC-Seq provides denser coverage of the epigenome

We further compared MC-Seq and Infinium450K in the context of their coverage of the epigenome (Table 12). Both platforms are similar in the percentage of promoter regions covered and provided relatively comparable coverage in CpG islands, shores and shelf.

Table 12 The coverage of genomic function and CpG content for Infinium450K and MC-Seq are comparable

Genomic Function	Infinium450K	MC-Seq	Observed sample from MC-Seq
3Utr	40%	40%	38%
5Utr	35%	45%	38%
Coding	13%	23%	22%
Exon	21%	29%	27%
Promoter	85%	88%	86%
TTS	87%	87%	86%
Intergenic	74%	71%	70%
Intron	29%	39%	38%
CpG Content			
Island	93%	91%	86%
Shore	87%	86%	84%
Shelf	79%	60%	57%

Table 13 MC-Seq covers CpG islands with a higher density.

MC-Seq		Infinium450K	
Coverage	Frequency	Coverage	Frequency
No coverage	973	No coverage	2029
1-24	0	1	1045
25 to 49	4817	2 to 4	13309
50 to 99	8021	5 to 9	8480
>=100	14880	10 to 24	3680
		25 to 49	132
		50 to 99	15
		>=100	0

We observed, MC-Seq covered the function regions at a much higher density, i.e. MC-Seq assays more CpGs for the same promoter region when compared to Infinium450K. Likewise, MC-Seq generally assays other genomic regions at much higher density. For example, in Table 13, for the total 28,691 CpG islands genome-widely, MC-Seq captures most of them and the number of CpGs in 14,880 islands are bigger than 100. In contrast, most islands are only measured by 2-4 CpGs in Infinium 450K.

3.4 Discussion

High-throughput methylome profiling of human samples is an important and evolving approach to investigate the role of epigenetic alterations in health. We established that MC-Seq with a capture-then-bisulfite-convert approach is an attractive alternative to

the widely used 450K array. We demonstrated that the performance of MC-Seq at 1 μ g and 3 μ g were similar, making these two technologies comparable in terms of the amount of DNA required (450K also requires about 1 μ g of starting DNA). We also compared the two platforms in terms of their coverage of the epigenome and showed that MC-Seq detected a higher density of CpGs than Infinium450K in key functional genomic regions and provided improved coverage of the human epigenome. Additionally MC-Seq also allowed detection of 20.8 million methylation sites within the non-CpG context (CHG + CHH), which is a magnitude higher than the 450K, which offers only 3091 such sites. Finally, we demonstrated as proof-of-principle, in a small sample, that methylation values from MC-Seq could distinguish subjects from different ethnic groups.

There are five practical considerations in choosing a platform (MC-Seq vs. 450K) to employ in an EWAS. These factors include

1. desired genomic coverage (MC-Seq has higher density of CpGs, and hence higher coverage of variable sites),
2. expected inter-group differences in methylation or sensitivity desired (450K has higher sensitivity for subtle variations in methylome),
3. cost (450K has lower cost),
4. coverage of methylation sites in non-CpG context (MC-Seq has a higher capacity),
5. customized and specified methylation sites of interest can be only done through bait design in MC-Seq [219] [220] .

Another practical consideration in employing MC-Seq is the amount of sequencing required and minimum sequencing depth to filter methylation sites for analysis. This decision should be guided by statistical power calculations and would depend on (i) technical variation of MC-Seq, (ii) expected inter-group differences and (iii) study sample size. For detecting large inter-group differences, the required amount of sequencing and sequencing depth would be less than that for detecting small inter-group differences. Nevertheless, we find that at 10X reads coverage, 84% of the detected CpGs had technical variation within 10%. When read coverage was increased to 30X, 91% of the detected CpGs had technical variation within 10%. At 50X, the number was 94%. It was a clear fact that further increases in reads coverage did not decrease technical variation significantly.

This study has several strengths. First, we conducted an extensive analysis of MC-Seq in buccal samples, a non-invasive clinical sample ubiquitously collected across cohorts and popular among those studying a paediatric population. Previous investigations of MC-Seq were in mouse [18, 219], maize [18], cell lines [18] and adipose tissue [18, 220]. Second, we demonstrated that MC-Seq could be used effectively with a capture-then-bisulfite-convert approach using only 1 µg of DNA. Other studies have used a capture-then-bisulfite-convert approach with 3 µg of DNA [18, 219, 220] which has a number of limitations. Third, it is the first study investigating the utility of MC-Seq in studying inter-individual variation from buccal samples from different ethnic groups. The only other study investigating inter-individual variation with MC-Seq was done on adipose tissue [220]. Last, we provided a comprehensive evaluation between MC-Seq and Infinium450K comparing the strengths and weaknesses of each platform for future EWAS. We demonstrated that MC-Seq with a capture-then-bisulfite-convert approach could be employed with 1 µg of DNA. Li et al [18] showed that a bisulfite-convert-then-capture assay could be used with as little as 500 ng of DNA. It will be of interest to investigate if MC-Seq approach can be used at lower quantities of DNA (≤ 500 ng) in future.

In summary, we found that MC-Seq is an attractive alternative platform to 450K, for interrogating DNA methylation at single base resolution in large number of clinical samples. Both platforms can be deployed with 1µg of DNA. MC-Seq provides denser coverage of the epigenome but the use of MC-Seq in EWAS with small effect sizes will only be feasible if the inter-group differences exceed the technical variation.

In the next chapters, a series of tools bundled in a package called GEM was developed based on comprehensive statistical models and data analysis, for the purpose to understand DNA methylome and their association with genotype, environmental factors and the interaction of genotype and environmental factors. GEM has been successfully applied in EWAS and is able to identified epigenetic biomarkers.

3.5 Supplementary data

Supplementary Table 1. Summary of MC-Seq sequence alignment and duplicate rates for four technical replicates at 3 µg

	F1_Ind	F3_Cau	F4_Chi	M7_Chi
Raw sequence reads	93798516	96096291	92769553	82161548
Sequence pairs analysed in total	83536052	83171725	82577165	74770924
Number of paired-end alignments with a unique best hit	69907181	70538589	68090749	58664868
Mapping efficiency (%)	83.7	84.8	82.5	78.5
Duplicate (%)	47.48	39.53	52.6	42.68
Sequence pairs after removing duplicate	36712773	42658112	32276416	33625981
Reads in targeted region (%)	89.43	88.81	89.56	90.82
Reads in targeted regions 200 bp (%)	92.91	92.32	92.84	93.35

Supplementary Table 2 Summary of sequence alignment and duplicate rates for two buccal epithelium samples using 1µg DNA from MC-Seq.

1µg DNA	F5_Chi	M7_Chi	Average
Raw sequence reads	82669270	46087480	64378375
Sequence pairs analysed in total	73443192	41243404	45866620
Number of paired-end alignments with a unique best hit	68026518	37965457	42481376
Mapping efficiency (%)	92.6	92.1	92.77
Duplicate (%)	29.66	12.44	16.73
Sequence pairs after removing duplicate	47848930	33244243	33603102
Reads in targeted region (%)	93.07	93.4	93.42
Reads in targeted regions 200 bp (%)	96.77	96.81	96.85

4 GEM: modelling the relationships among DNA methylation, genetic variants and environment factors

4.1 General multiple regression model

A general multiple-regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \dots + \beta_k x_{i_k} + \epsilon_i, \quad i = 1, \dots, n \quad (4-1)$$

Where β_0 is a coefficient for consistent variable 1 and $\beta_1, \beta_2 \dots \beta_k$ are k coefficients for k independent variables $x_{i_1}, x_{i_2}, \dots x_{i_k}$. The basic assumptions for errors ϵ_i are (1) mean (ϵ_i) = 0, (2) errors are not correlated to variables, (3) they are identically distributed.

In matrix form, the model can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (4-2)$$

Taking into consideration of the variables are combined of interesting factor $\mathbf{X}$ and covariates such as ($cvrt1, cvrt2, \dots$), the Equation (4-2) can be written into

$$\mathbf{Y} = \alpha + \beta \mathbf{X} + \gamma \mathbf{cvrt} + \epsilon \quad (4-3)$$

where $\mathbf{X}$ is the predictor factor and $\mathbf{Y}$ is the response. $\mathbf{cvrt}$ represents the covariates. ϵ indicates errors which satisfy the basic assumptions and we test β .

In order to understand DNA methylation changes related to environmental exposures and genetic variants, epigenome wide association studies (EWAS) employ general

multiple regression model to study the relationships among DNA methylation, genotypes and environment factors by the adjustment of cofounders from individual differences. Three major questions interrogated included:

1. How methylation changes correspond to particular environment factor or disease trait?
2. How genotypes affect the changes in DNA methylation?
3. Can the interaction between genotype and environment factors be able to explain DNA methylation changes better?

Representing methylation data matrix as **M**, genotype data as **G**, and environment matrix as **E**, a suite of statistical models to study the complicated relationships among of Gene, Environment and Methylation was developed and named as **GEM**.

4.2 Understanding data matrix and terminologies in epigenetic research

4.2.1 DNA methylation: M

The genomic DNA extracted from biological specimens and as discussed in previous chapters, DNA methylation cross the genome can be assayed using either microarray (eg. Infinium450K) or specialized next generation sequencing protocols (eg. RRBS or MC-Seq). In order to process Infinium 450K data with minimum bias, a novel data analysis pipeline was proposed and validated by comparison studies on same samples from RRBS. I also introduced the latest MC-Seq technology and reviewed the data analysis with real clinical samples. All these efforts are crucial to ensure the quality of methylation data with minimum bias from factors related to sample processing and avoid confounding experimental variables with non-biological batch effects.

DNA methylation typically occurs in a CpG (Cytosine followed by a Guanine) dinucleotide context and DNA methylation profile is a percentage measurement for cytosine methylation. In the Infinium 450K, each CpG is measured by a 50 base pair (bp) probe centered to targeted cytosine in CpG dinucleotide, as shown in

Figure 26.

AGGTGAATGAAGAGACTAATGGGAGTGGCTTGCAAGCCAGGTACTGT
GCCCTACAACTCT[CG]CGCTGCCACAAAAAGCCAATTAACGAAGCAAA
CTGTTTTATCTTTTCCTCTCCCCAGAAG

Figure 26 A typical Infinium 450K design for a probe of 50 base pair nucleotides to measure a CpG in gene PRDM16 region with ID as “cg08007925” in chromosome 1 and position as 3322020.

A methylation profile for a CpG is a percentage measurement of the number of cytosine methylated. For each CpG, DNA methylation profile is calculated as the *beta* value, as $beta = \frac{Methylated}{(Methylated+Unmethylated)}$, where *beta* is ratio of the intensity of methylated signals over the sum of methylated and unmethylated signals [15]. The *beta* value ranges between 0 to 1 and normally presents as a percentage value, so that *beta* value at 0~20% indicates the probe hypo-methylated while 80~100% means hyper-methylated.

Table 14 A sample DNA methylation data matrix shows that six CpGs were measured through four samples (S1, S2, S3 and S4). Methylation profiles showed cg00000109, and cg00000658 were hyper-methylated while the other CpGs were hypo-methylated in the four samples.

CpG	S1	S2	S3	S4
cg00000029	0.1741	0.1293	0.0961	0.1688
cg00000109	0.8210	0.8878	0.8415	0.8597
cg00000289	0.5682	0.7400	0.7493	0.6609
cg00000292	0.6949	0.7543	0.7118	0.7377
cg00000363	0.0747	0.0847	0.0644	0.1249
cg00000658	0.8531	0.8991	0.8365	0.8607

We use matrix to represent DNA methylation $\mathbf{M}$, each row representing one CpG cross all samples, and each column for all CpGs for one sample, such as

$$\mathbf{M} = \begin{bmatrix} CpG_{11} & CpG_{12} & \cdots & CpG_{1j} \\ \vdots & \vdots & \ddots & \vdots \\ CpG_{i1} & CpG_{i2} & \cdots & CpG_{ij} \end{bmatrix},$$

where $i = 1, \dots, number_CpGs$ and $j = 1, \dots, number_Samples$ and the matrix $\mathbf{M}$ is a methylation matrix with a dimension of $number_CpGs \times num_Samples$. For DNA methylation from Infinium 450K platform, the *number_CpGs* is about 450,000 and in a typical EWAS study, the *number_Samples* is about hundreds and thousands.

For data from Table 14, the data matrix is,

$$M = \begin{bmatrix} 0.1741 & 0.1293 & 0.0961 & 0.1688 \\ 0.8210 & 0.8878 & 0.8415 & 0.8597 \\ 0.5682 & 0.7400 & 0.7493 & 0.6609 \\ 0.6949 & 0.7543 & 0.7118 & 0.7377 \\ 0.0747 & 0.0847 & 0.0644 & 0.1249 \\ 0.8531 & 0.8991 & 0.8365 & 0.8607 \end{bmatrix}.$$

4.2.2 Genome-wide SNP measurements: G

The genotype is the DNA sequence of the genetic makeup of a cell. SNPs are the bases for the measurement of genetic variants and emerged as important marks for linking genotype and individual diseases and traits in GWAS [143]. In this study, the genotype was interrogated on Illumina Omniexpress+exome genotyping arrays. These arrays measure at least 1,000,000 the most frequent SNPs genome wide (omniexpress) and all of those discovered within exomes. Data are processed in GenomeStudio Genotyping Module™. Genotyping calls were made by the GenCall software which incorporates a clustering algorithm (GenTrain) and a calling algorithm (Bayesian model). GenCall Score of each SNP probe and call rate of each sample are generated. The genotypes with a GenCall score less than 0.15 are not assigned genotypes. A typical genotype data is shown in Table 15, where pseudo codes “0”, “1”, “2” stand for major allele homozygote, heterozygote and minor allele homozygote and the minor allele is presented in an additive way. An allele is defined as a specific genetic variants at a given gene or genetic locus where more than one variants exists. For example, there are two pieces of DNA sequences in

Figure 27, “seq 1” and “seq 2”, “G” is an allele in “seq1” and “A” is an allele in “seq 2”. If a majority of the population present “G” at this same sequence locus, allele “G” is declared the major allele and it follows that “A” as a minor allele. Given humans are diploid organism (contain two copies of all autosomal chromosomes), for any locus there can be two possible alleles, one on each chromosome copy. So “G/G” is termed as a major allele homozygote, “G/A” represents heterozygote and “A/A” is the minor allele homozygote.

Table 15 A sample genotype data matrix table shows that five SNPs were measured through four samples (S1, S2, S3 and S4). We used additive pseudo codes “0”, “1” and “2” to stand for major allele homozygote, heterozygote and minor allele homozygote.

SNPID	S1	S2	S3	S4
rs3094315	0	0	0	0
rs3131972	1	0	0	1
exm2268640	1	0	0	1
rs12562034	2	1	2	2
rs11240777	1	1	0	1

Figure 27 SNP and major allele homozygote, heterozygote and minor allele homozygote

In genetic variants matrix G , each row represents one SNP, and each column represents one sample, such as

$$G = \begin{bmatrix} SNP_{11} & SNP_{12} & \cdots & SNP_{1j} \\ \vdots & \vdots & \ddots & \vdots \\ SNP_{i_1} & SNP_{i_2} & \cdots & SNP_{ij} \end{bmatrix},$$

where $i = 1, \dots, number_SNPs$ and $j = 1, \dots, number_Samples$ and the matrix G is a genotype data with a dimension of $number_SNPs \times num_Samples$.

4.2.3 Environment measurements including birth outcome and maternal body conditions: E

We generally termed all biological conditions, disease status and other phenotypic variables as environmental factors. Environment factors are independent vectors in the format of continuous or categorical values. Due to the need to conduct comprehensive analysis for multi-ethnicity GUSTO birth cohort, we focused on the clinical data from birth outcomes and prenatal conditions, such as birth weight, gestational age and maternal mood etc. In environment factor E , each row represents one factor, while each column represents one sample, such as

$$E = [E_1 \ E_2 \ \cdots \ E_j],$$

where $j = 1, \dots, number_Samples$ and the matrix E is one dimensional matrix or vector.

4.2.4 Covariates matrix for DNA methylation related study: $cvrt$

Epigenetic studies explore the methylation changes in association with genotypic and phenotypic differences in samples. Including covariates pertaining to the subjects, such as gender, ethnicity and socio-economic status [77], and to the biological specimens, such as cellular heterogeneity [79, 80, 230], is important. In consistency with the matrix designed for GEM, matrix $cvrt$ is also a data matrix with each row representing one covariate, and column for each sample.

4.3 Gene, Environment and Methylation (GEM): A tool suite modelling the associations between genotyping, environmental factors towards methylome

The interplay among genetic, environment and epigenetic variation is not fully understood. Advances in high-throughput genotyping methods, high-density DNA methylation detection and well-characterized sample collections, enable epigenetic association studies at the whole genomic and population levels (EWAS). The field has extended to interrogate the interaction of environmental and genetic (G x E) influences on epigenetic variation. Also, the detection of methylation quantitative trait loci

(methQTLs) and their association with health status has enhanced our knowledge of epigenetic mechanisms in disease trajectory.

In order to address these problems we attempted to understand the relationships among **Genotype**, **Environment** and **Methylation**. A tool suite named GEM was developed to address these problems by three major functions GEM_Emodel, GEM_Gmodel and GEM_GxEmodel.

4.3.1 GEM_Emodel: detecting methylation markers associated with environment factors

In most EWAS studies, the primary goal of DNA methylation analysis is to identify the differentially methylated CpGs among environment factor or phenotypic groups.

For a given sample size N , the methylation profile cross at each CpG site is,

$$M_i = (M_1, M_2, \dots, M_N, i = 1, 2, \dots, \text{number_CpGs}).$$

An environmental factor, such as a disease trait or a clinical measurement is fixed for all CpGs and can be presented as,

$$E = (E_1, E_2, \dots, E_N \mid N = \text{number_Samples}).$$

The association of methylation differences and environmental factors can be tested by

$$M_i = \alpha + \beta_E \times E + \gamma \times \text{cvrt} + \epsilon_i \quad (4-4)$$

The model in Equation (4-4) underlying a linear regression analysis is that the explanatory variable E and outcome variable M are linearly related by adjustment of covariates, such that the population mean of the methylation can be explained by environment factor in a linear relationship. It can be simplified as

$$M \sim E + \text{cvrt} \quad (4-5)$$

If association is present, M_i for a particular CpG will be seen more often than expected by chance in an individual carrying the environment factor by computing the standard errors, the t-statistic values and the corresponding p-values. We then compare these p-

values for all CpGs tested to a threshold and make the decision whether the relation can be retained or rejected.

We name Equation (4-5) as GEM_Emodel. GEM_Emodel is performed at the number of *number_CpGs* to determine for methylation profiles among all CpGs measured genome widely (in Infinium 450K, the number of CpGs is about 450,000), whether a methylation variant is associated with an environment factor.

4.3.2 GEM_Gmodel: detecting methylation markers associated with genotype

Epigenetic states are related to genotype. SNPs are the most important genotype marks for genetic variants in genome loci.

Gibbs et al [120] showed that in human brain tissues, there were about 4~5% of the CpGs studied significantly dependent on genotype. Gibbs et al. thus defined the DNA methylation driven by DNA sequence as methylation quantitative trait loci (methQTL). In order to find methQTLs, GEM_Gmodel was developed in Equation (4-6) to test for a methylation profile at particular CpG,

$M_i = (M_1, M_2, \dots, M_N, i = 1, 2, \dots, \text{number_CpGs})$, which SNP can explain CpG changes SNPs are in the format of $G_j = (G_1, G_2, \dots, G_N, j = 1, 2, \dots, \text{number_SNPs})$.

$$M_i = \alpha + \beta_{G_j} \times G_j + \gamma \times cvrt + \epsilon_{ij} \quad (4-6)$$

The above Equation (4-6) can be simplified as

$$M \sim G + cvrt \quad (4-7)$$

We name Equation (4-7) as GEM_Gmodel. GEM_Gmodel tests methylation variants at thousands of CpG sites in relation to genetic changes at millions of SNPs. If association is present and the p-values passed statistical testing, a methQTL is readily detected as a CpG-SNP pair. In GEM_Gmodel, billions of linear regressions are tested (*number_CpGs* $\times$ *number_SNPs*) against thousands of CpGs and millions of SNPs and a genome wide methQTL map is created.

4.3.3 GEM_GxEmodel: interaction of genotype and environment to specify methylation variants

The GEM_GxEmodel tests the ability of gene and environment interaction (GxE) to predict DNA methylation levels. Evidences for an environmental influence on DNA methylation states as well as an influence of genotype to methylation variation suggest that gene $\times$ environment interactions may be common. In epigenetics, the integration of genetics and environment could be integrated to explain the underling mechanism of methylation changes.

In GEM_GxEmodel, methylation is

$M_i = (M_1, M_2, \dots, M_N, i = 1, 2, \dots, number_CpGs)$, genotyping is

$G_j = (G_1, G_2, \dots, G_N, j = 1, 2, \dots, number_SNPs)$ and environment is $E =$

$(E_1, E_2, \dots, E_N, N = number_Samples)$, GEM_GxEmodel tests whether the

interaction of genotype and environment can explain methylation by the below equation:

$$M \sim G \times E + cvrt \quad (4-8)$$

GEM_GxEmodel as illustrated in Equation (4-8) tests the association between methylation variants and environment in the stratification by genotype groups. For a particular environment factor, there are billions of tests ($number_CpGs \times number_SNPs$). If association is present and the p-values passed statistical testing, the result is a triplet as CpG-SNP-Environment. This triplet can be interpreted as, methylation in different genotype groups have significantly different response to the changes in the environment.

4.3.4 Exploring the association between DNA methylation, genetic variants and environmental factors

To explain methylation variants, GEM_Emodel, GEM_Gmodel and GEM_GxEmodel have been developed. Methylation for each CpG, M , has been tested by G , E , and $G \times E$. A further question to be answered is ‘what proportions of CpGs are independently predicted by environment, genotype or GxE?’ In order to profile the

relationships between methylation, genotype and environment, Akaike Information Criterion (AIC) [227] approach was utilized to determine the most fitted model among GEM_Emodel, GEM_Gmodel and GEM_GxEmodel. Models were ranked by their AIC values, and the best model was defined as the one with the minimum AIC value.

Literature shows, both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [228] are two famous and well used measures of ascertain model quality in statistics. The best model is the one with the minimum AIC or BIC value as shown in Equation (4-9) and (4-10). Basically the equations are,

$$AIC = -2 \ln L + 2k \quad (4-9)$$

And

$$BIC = -2 \ln L + k \ln n \quad (4-10)$$

where L is the maximum likelihood of the candidate model, k is the number of estimated parameters and n is the sample size in the candidate model.

The comparison between AIC and BIC in model selection has been studied extensively [229, 230]. Basically both AIC and BIC measure the likelihood estimation of model fitting and each of them has a different penalty function. When the number of estimated parameters k is fixed, BIC increases as sample size increases. In our particular test, as the number of sample size was fixed, both AIC and BIC produced identical selection results though the magnitudes of both values were different. Given the identical results, we decided to use AIC to determine the best fitting model among GEM_Emodel, GEM_Gmodel and GEM_GxEmodel. Models were ranked by their AIC values and the best model was defined as the one with the minimum AIC value. Note that the calculation of AIC and BIC were used by the built-in function in R scripts [171].

For each CpG, corresponding to $M_i, i = 1, \dots, N_{CpG}$, the best CpG~SNP was selected as j th GEM_Gmodel, and k th GEM_GxE model, and the “winner” i.e. the model with the least AIC value, which best explained the M compared to the alternative models was declared in Equation (4-11) to (4-14). The AIC values for models to fit

methylation are AE by Emodel, AG by Gmodel and AGE by GxEmodel and the least of them is the “winner”, for example, for a CpG, if its AE is the least, its methylation profile can be best explained by Emodel.

GEM_Emodel	$AE = AIC(Emodel)$	(4-11)
Winner in GEM_Gmodel	$n = \operatorname{argmin}_j(AIC(Gmodel_j)),$ $j = 1, \dots, number_SNPs$ $AG = AIC(Gmodel_n)$	(4-12)
Winner in GEM_GxEmodel	$l = \operatorname{argmin}_j(\operatorname{argmin}_j(AIC(GxEmodel_j))),$ $j = 1, \dots, number_SNPs$ $AGE = AIC(GxEmodel_l)$	(4-13)
Best of all	$winner = \operatorname{argmin}(AE, AG, AGE)$	(4-14)

4.4 Developing GEM package for efficient EWAS and methQTL analyses

The computational burden for analyzing the ‘omics’ data from EWAS is large due to the need to integrate the high throughput measurements from thousands of CpGs, millions of SNPs, complicated environment factors and cofounders for hundreds or thousands subjects. Hands-on analysis tools in EWAS, especially tools in a normal operational setting that can be easily accessed by biology lab researcher are currently not available. GEM was the first tool suite developed to help expedite epigenomic discovery in thousands of samples [4] and it was well appreciated by the scientific community [4]. The successful application of GEM in exploring the effect of genotype and *in utero* environment on inter individual variation in neonate DNA methylomes [4] was the motivation to develop computationally effective solutions that allow biological researchers to explore EWAS, methQTLs and GxE using standard desktop computers within reduced processing time.

Recently Shabalin [231] developed a R package named MatrixEQTL for the analysis of expression quantitative trait loci (eQTL) analysis. Based on matrix operation, iterative correlation was implemented to achieve computational efficiency, and data was sliced into blocks to achieve memory efficiency. A function in MatrixEQTL that allows inclusion of interaction terms in correlative statistical models, gained our attention, though the author did not highlight it when the package was reported. We

deployed the fast and efficient MatrixEQTL software and implemented “GEM” tool suite into R package.

GEM package implemented three fast linear regression models denoted GEM_Emodel, GEM_Gmodel and GEM_GxEmodel to facilitate analyses in EWAS. The GEM_Emodel tests the association of methylome marks and environmental factors; the GEM_Gmodel creates a methQTL genome-wide map; finally, the GEM_GxEmodel tests the ability of gene and environmental interaction models to predict DNA methylation levels. We benchmarked the performance of the GEM operations on a publically available EWAS dataset generated on the 450K array with concurrent genotyping on the OmniExpress Array and simulated environment and phenotype information on 237 neonates [4]. Our results demonstrated that the GEM package can facilitate reliable EWAS analyses within minutes, in a standard computational setting (processor=2.2GHz, RAM=8G, system = window7 64bit).

4.4.1 GEM implementation

Simplifying the data input into a methylation matrix as M , genetic variants matrix as G , and the environment vector as E , and the matrix for covariates as $cvrt$, and using a R language like pseudo code, we can denote GEM_Emodel (detecting methylation markers associated with environment) function as $lm(M \sim E + cvrt)$, Gmodel (detecting methylation markers associated with genotype i.e. methQTLs) as $lm(M \sim G + cvrt)$ and GxE model (interaction of genotype and environment to specify methylation marks) as $lm(M \sim G \times E + cvrt)$. The genome wide studies for Emodel, Gmodel and GxEmodel can be accomplished by calling R function lm iteratively by millions of times, which were denoted as LM_Emodel (Table 16), LM_Gmodel (Table 17) and LM_GxEmodel.

Shabalin [231] introduced matrix standardization and projection and successfully made an ultra-fast software for expression quantitative trait loci (eQTL). Basically, to quantify the strength of the relationship between x and y controlled by covariates ($cvrt$), a practical regression is,

$$y = \alpha + \beta x + \gamma cvrt + \varepsilon,$$

where α, β, γ and ε are coefficients, β is to be estimated. A standardization method (22) was applied to vector $x, y, cvrt$, then the projections of x and y to $cvrt$ are,

$$\tilde{\mathbf{y}} = \mathbf{y} - \langle \mathbf{y}, \mathbf{cvrt} \rangle \mathbf{cvrt}, \text{ and } \tilde{\mathbf{x}} = \mathbf{x} - \langle \mathbf{x}, \mathbf{cvrt} \rangle \mathbf{cvrt},$$

where $\langle \rangle$ denotes inner product of two matrix. After these operations, the linear regression between $\mathbf{x}$ and $\mathbf{y}$ with covariates $\mathbf{cvrt}$ can be simplified into the calculation of inner product of the projects of $\mathbf{x}$ and $\mathbf{y}$ as $\mathbf{r}_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}} = \langle \tilde{\mathbf{x}}, \tilde{\mathbf{y}} \rangle$ and estimation of the test statistics.

Shabalin [231] also demonstrated the strategy to slice the large matrix into a small “blocks” in the correlation calculation for memory efficiency, which make the software able to handle data matrix with millions of rows and columns feasible in normal computational setting.

GEM tools called the MatrixEQTL[231] library. GEM_Emodel was implemented by calling matrixEQTL with parameter as “modelLINEAR”, replacing gene expression with methylation, and SNP with environmental data. GEM_Gmodel was implemented by calling matrixEQTL with “modelLINEAR”, replacing gene expression with methylation. GEM_GxEmodel was implemented by calling matrixEQTL with “modelLINEAR_CROSS”, replacing gene expression with methylation.

GEM_Emodel finds the association between methylation and environment genome-wide by performing millions of linear regression ($N = \text{number_CpGs}$). The output of Emodel for particular phenotype, environmental factor or disease trait is a list of CpGs that are potential epigenetic biomarkers, as in Table 16.

Table 16 Pseudo R code for GEM_Emodel

Algorithm: exploring the association between methylation and environment factors by Emodel.	
<hr/>	
Input:	
	Methylation matrix M ($\text{number_CpGs} \times \text{number_Samples}$), and each $M(i)$ is the vector for i th CpG cross all samples.
	Environmental vector E (number_Samples),
	Covariate matrix $cvrt$ ($\text{number_covariates} \times \text{number_Samples}$).
Output:	
	A list of CpG-Env pairs
1	Function: LM_Emode ($M, E, cvrt$)
2	For $i=1$ to number_of_CpGs {
3	$\text{fit} \leftarrow \text{summary}(\text{lm}(M(i) \sim E + cvrt))$
4	}
5	To process GEM_Emodel($M, E, cvrt$)
6	{ library (GEM)
7	$\text{result} = \text{GEM_Emodel}(M, E, cvrt)$
8	}

Table 17 Pseudo R code to explore methQTLs by Gmodel.

Algorithm: exploring the methQTL by Gmodel.	
Input:	
	Methylation matrix M (number_CpGs × number_Samples), and each M(i) is the vector for ith CpG cross all samples.
	Genotype matrix G (number_SNPs × number_Samples), and each G(j) is the vector for jth SNP cross all samples.
	Covariate matrix cvrt (number_covariates × number_Samples).
Output:	
	A list of CpG-SNP pairs, where the SNP is the best fit to explain the particular CpG. The significant association between CpG-SNP pair suggests the methylation driven by genotyping variants, which is so called (methQTL).
1	Function LM_Gmode (G, M, cvrt)
2	For i=1 to number_of_CpGs {
3	For j=1 to number_of_SNPs{
4	fit<- summary(lm(M(i) ~ G(j) + cvrt))
5	}
6	result(i) = (M(i), G(K)) by
7	$K = \operatorname{argmin}_{k \in [1, \text{number_SNPs}]} (AIC(fit))$
8	}
9	To process GEM_Gmode(G, M, cvrt)
10	{ library(GEM)
11	result = GEM_Gmodel(G, M, cvrt)
	}

Table 17 demonstrates the pseudo code that used the *lm* function in iterative loops for Gmodel denoted as LM_Gmodel. The best fit is chosen by the minimum AIC values.

Replacing the linear regression equation (line 6) in Table 17 by “fit <- summary (lm(M(i) ~ G(j) * E + covt))”, produces the pseudo code for the implementation of LM_GxE model. The output of GxEmodel is a list of CpG-SNP-Env triplets, indicating the CpG-Env association segregated by genotype. The significant association of each triplet implies the methylation change is determined by the interplay between genotyping and environment. Both implementations indicate the number of linear regressions, $N = \text{number_SNPs} \times \text{number_CpGs}$. N could be in the billions of linear regressions engendering a very substantial computational task. However, using GEM tools, calculation of methQTLs and GxE interactions can be accomplished with much improved computational efficiency.

4.4.2 Benchmarking Results

To benchmark GEM suite, we used the dataset from Teh et al [4]. The standard laptop used for time comparisons had a 2.2GHz processor, 8G RAM, a windows 7 operating system and was 64 bit, which is typical in an academic setting. The high performance computing (HPC) structure had eight parallel processes, each with eight core CPUs as in Figure 28 (n=8).

Figure 28 HPC computing settings (n=8).

In Teh et al. [4], 1423 CpGs were studied from the methylomes of 237 neonates, and their association with 708,365 genetic variants and nineteen environmental factors made up of maternal conditions and birth outcomes. The methylome and genotype data are publically available at the NCBI Gene Expression Omnibus (GEO: <http://www.ncbi.nlm.nih.gov/geo/> under accession numbers GSE53816 and GSE54445.) Environmental factors were simulated.

A schematic of the analyses performed is shown in Figure 29. When the original analyses were conducted, multivariate regression models were applied sequentially in a HPC environment. For the same dataset, we compared the time taken to execute LM_Gmodel, LM_GxEmodel and LM_Emodel in a standard and HPC computational environment with the time taken to run GEM_Gmodel, GEM_GxEmodel and GEM_Emodel on a standard laptop, Table 18.

Figure 29 A schematic of the analyses performed within Teh et al [17] . Gmodel sequentially tests the association of 1423 methylation marks with 708,365 genotypes (and covariates), Emodel sequentially tests the association of 1423 methylation marks with 19 environmental factors and GxEmodel sequentially tests that association of 1423 methylation marks with the interaction of every combination of the 708,365 genotypes and 19 environmental factors.

4.4.2.1 Benchmarking GEM_Emodel

A substantial time saving was achieved using GEM_Emodel compared to standard sequential regression as LM_Emodel as in Table 16 (19 seconds compared to 95 seconds for 19 Emodels on 1423 CpGs). Results achieved were identical. In addition, GEM_Emodel has the option to create Q-Q plot for theoretical distribution and observed distribution on p_values for every environment, e.g. Figure 30. As subject numbers increase, computational time to run sequential models increases geometrically, whilst computational time in GEM_Emodel increases linearly. Figure 31 shows the computational time required for one Emodel on 100-1000 subjects for 250K CpGs.

Figure 30 QQplot for p-values from iterations of 255K CpGs with an environment factor. It was produced by GEM_Emodel.

Figure 31 The operation time consumption benchmark on the associations for 250K CpGs and one environmental factor and one covariate for the sequential number of samples from 100 to 1000. X-axis is the number of subjects, and y-axis is the time consumption in seconds. The benchmarking is done in a personal computer (processor=2.2GHz, RAM=8G, system = window7 64bit).

4.4.2.2 Benchmarking GEM_Gmodel

In the original analysis [4], the regression equation, shown in Equation (4-7) and Table 17, used built-in lm function in R script, which we denoted as LM_Gmodel, was applied to each of the 1423 VMRs, cycling through the 708,365 SNPs, adjusted by sex

as the covariate, resulting in 1008 million regression models. We compared the LM_Gmodel with GEM_Gmodel by the result and computational efficiency in a standard laptop (processor=2.2GHz, RAM=8G, system = window7, 64bit). We also used a HPC structure with eight parallel processes of each with eight core CPUs (denoted as HPC) to benchmark LM_Gmodel as a reference. The computational time on HPC was 3 hours, in a standard computational environment, computational time was estimated to be 61 days. The same data was processed by the GEM_Gmodel. It took 5.2 minutes to accomplish the task on a standard laptop. The results were identical to those reported by Teh et al [4] i.e. 12 disrupting pairs, 828 in *cis*- pairs and 583 in *trans*- pairs.

Table 18 Benchmarking time consumption of GEM implementation on Emodel, Gmodel and GxEmodel by comparing it with a R script available in public database and using a standard laptop/HPC setting. (** The time for LM_Gmodel and LM_GxEmodel in standard laptop was computed based on the time cost on 10 CpGs.)

Dataset: Teh et al, 1423 CpGs, 708,365 SNPs and 19 environments in a standard laptop				
Method	Time cost on standard laptop	Time cost in HPC	Method	Time cost
LM_Emodel	95.1 seconds		GEM_Emodel	18.9 seconds
LM_Gmodel	>= 60days (**)	3 hours	GEM_Gmodel	5.2minutes
LM_GxEmodel	>= 60 days (**)	21 hours	GEM_GxEmodel	1.5 hours

4.4.2.3 Benchmarking GEM_GxEmodel

The same scale of improvement in performance was achieved for the GEM_GxEmodel where each CpG was tested against the interaction of genotype at each of 708,365 SNPs with each of 19 environmental factors. This analysis originally took 21hours in the HPC environment and an estimated >=60 days on a standard laptop by using normal linear regression in R script, denoted as LM_GxEmodel. In the GEM_GxEmodel, it was accomplished in only 1.5 hours. The results were identical between analyses with identical p-values for models containing all winning pairs of SNPs and environments (data not shown).

In addition GEM also has the option to produce a “segregation scatter plot” for methylation corresponding to environment in different genotype groups, for example, Figure 32.

Figure 32 The scatter plot to display an example of methylation corresponding to the environment in different genotype groups. AA, AB and BB are pseudo codes for major allele homozygote, heterozygote and minor allele homozygote. Phenotypic values are shown on the x-axis, and methylation values in percentage on the y-axis. The straight lines fit for associations in each group.

4.4.3 Package availability and the ease of use

GEM package can be downloaded from Bioconductor at <https://www.bioconductor.org/packages/GEM/>. In order to make it more user friendly, we also incorporated a graphic user interface (GUI) in the tool package, see Figure 33.

Figure 33 GEM graphic user interface. (a) the selection of functions. (b) the file and parameter window.

4.5 Conclusion and discussion

The advancements in genome-wide genotyping and DNA methylation assessment methods, coupled with well-characterized biological samples enable epigenetic association studies. GEM was designed to test the millions of hypotheses in epigenetics for the needs to understanding methylation changes responsive to the interaction between genotype and environment. This new method has evolved the analysis approach from simple iterations to matrix based efficient operation. Both of the analysis and benchmarking demonstrated the ease of using GEM with standard computing resources available to nearly all researchers.

GEM includes a GUI for the convenience of researchers and does not require specialist computational knowledge, outside of the widely used R environment.

Missing value is not allowed in matrix based operation so we used the mean values to impute the missing values in GEM's functions, on the contrary, the linear regression function (LM) in R programming treats missing value as "NA". In order to benchmark the impact of mean value-based imputation, 255K CpGs for N=237 samples were tested with an simulated birth weight as the enviornmental factor for both GEM_Emodel and LM_Emodel . In the 255K CpGs, 18% of them contained at least one missing values. Figure 34 demonstrated that there were strong concordance for p-values of the 255K CpGs from the two functions at $R^2 = 0.956$. However further in-depth analysis based on simulated and real data have to be done in order to make a conclusive statement, researchers should access the suitability of this imputation in their particular cases.

Figure 34 Emodel benchmarking for methylation matrix containing missing values. *pvalue* was transformed as $-\log_{10}$. A-axis is *pvalue* from LM_Emodel, y-axis is from GEM_Emodel. Among 255K CpGs that were tested, 18% of them contained at least one missing values. Our results showed *pvalue* for CpGs without missing values are perfectly matched, while there were slightly differences between the two implementations when CpG contains missing values.

5 Modelling the relationships among environmental factors, genetic variants and neonate methylome

5.1 Introduction

This chapter highlights GEM's capacity to test if DNA methylation levels (M) can be modified by genetic variation (G), environmental exposure (E) or the interaction between genetic variation and environment factors (GxE) [158], as shown in Figure 35.

Figure 35 Genotype, environment, and the interaction between the two influence DNA methylation levels.

The aim of this study was to investigate the relative influences of genotype and *in utero* environment on inter-individual differences in DNA methylation of neonates. Genotypic influence on DNA methylation has been evidenced in many studies [144-146, 148] and others have evidenced the influence of *in utero* environment on neonate methylomes [129, 139, 161, 232-234]. We interrogated a unique dataset of neonatal methylomes from the GUSTO birth cohort, along with *in utero* environment and genotype data.

5.2 Experimental design and data processing

DNA methylation and genotypic variation were interrogated on genomic DNA extracted from the umbilical cord tissue of 244 neonates from GUSTO cohort [176]. The birth outcomes included offspring's nationality, gender, birth weight and gestational age. Maternal measurements related to *in utero* environment were collected at 26th weeks of pregnancy and included maternal mood health, maternal oral glucose tolerance (OGTT) and micronutrients levels. In all nineteen environment factors were interrogated in this study. All data were processed based on the best knowledge about the properties of each data. Figure 36 provided the overall data repository of this study.

Data repository: 237 GUSTO neonates

Figure 36 Data repository for study relationships among genotype, environment and methylation profiles from 237 neonates.

Table 19 Ethnicity and sex of subjects

Characteristics		count(percentage)
Ethnicity	Chinese	131(57.81)
	Malay	72 (30.38)
	Indian	34 (14.35)
Gender	Male	133 (56.12)
	Female	104 (43.88)

5.2.1 Identification of 1423 Variable Methylation Regions

The genomic DNA was extracted from frozen umbilical cord specimens and DNA methylation profiles were measured by Infinium 450K arrays. Data was processed as described previously in Chapter 2 and [1]. In quality check (QC) assessment during data preparation, 237 of 244 samples passed QC and were used for downstream analyses. These 237 subjects included 131 Chinese, 72 Malay and 34 Indian neonates (see Table 19, for study subject characteristics).

Batch effect was observed upon principal component analysis (PCA)[235], and the data was converted to M-values [175] and the batch effect corrected via a widely used empirical Bayes methodology (COMBAT) [185]. Sex chromosome and cross-hybridizing CpGs [224, 225] were removed. A total of 301,468 probes (from the 411,107 assayed) remained after QC and were used for downstream analyses.

Methylation levels at the majority of CpGs in the neonate methylomes were very similar across individuals, with median absolute deviation scores (MAD) [126] < 0.1 , so only inter-individual variably methylated regions (VMRs) were studied. VMRs were defined across the 237 individuals using a previously published methodology [162]. A candidate VMR was defined as at least two spatially contiguous probes within 1000 bp of each other and with median absolute deviation (MAD) values more than 95th percentile. Candidate regions were expanded to contain more than two probes, as long as the distance between any two neighboring probes within the region was not larger than 1000 bp. For each VMR, a representative CpG was selected, the VMR-CpG was the CpG with the highest MAD score in the variably methylated region (VMR). This analysis returned 1423 VMR regions. Thus 1423 CpGs were used as the dataset for subsequent analyses.

5.2.2 Genome-wide SNP measurements

The genotype in the GUSTO birth cohort [176] was interrogated on Illumina Omniexpress+exome genotyping arrays. Totally 708,365 SNPs (from the 958,178 assayed) varied in genotypes. These were data G (708365 SNPs x 237 samples) in Figure 36.

5.3 Applying GEM to understand neonate methylome by genetic variants and environment factors

The number of samples was denoted by $N_{sample} = 237$ and the number of VMR-CpGs by $N_{CpG} = 1423$. For each $i = 1, \dots, N_{CpG}$, neonate methylome of VMR-CpG was represented by a vector $\mathbf{M}_i$, with N_{sample} elements of *beta* values ranged in [0,1], as shown in Equation (5-1). The number of N_{CpG} vectors of VMR-CpGs constructed the matrix $\mathbf{M}$ whose dimension was $N_{CpG} \times N_{sample}$ as shown in data repository Figure 36.

Similarly, the number of SNPs was denoted as N_{SNP} . For each $j = 1, \dots, N_{all_SNPs}$, the genotype for SNP was represented by a vector $\mathbf{G}_j$, with N_{sample} elements. Together, N_{all_SNPs} vectors of SNP constructed the matrix $\mathbf{G}$, which represented the genotyping profiles for all N_{sample} samples, in Equation (5-1). As shown in Figure 36, $N_{SNP} = 708,635$. It covered 708,635 SNPs in genome wide scale which was coded as 0 (AA, or major-major allele), 1 (AB, or major-minor allele) and 2 (BB, or minor-minor allele). In GEM_GxE model, CpG was modelled with all SNPs in the same chromosome, and the detailed data distribution is shown in Table 20.

Table 20 Distribution of VMR-CpGs and SNPs on chromosomes.

Chromosome	$N_{CpGs_in_Chromosome}$	$N_{SNPs_in_Chromosome}$
1	114	57603
2	112	55686
3	44	45952
4	73	39468
5	89	41455
6	148	48945
7	98	36217
8	81	36078
9	15	32472
10	79	37718
11	75	36463
12	98	34920
13	50	26268
14	23	22886
15	39	21509
16	54	22308
17	73	20215
18	17	20467
19	74	16044
20	22	17954
21	22	10068
22	23	10518

For each $k = 1, \dots, N_{ENV}$, the environment factor $\mathbf{E}$ was represented by a vector $\mathbf{E}_k$ of N_{sample} elements. Together nineteen measurements of baby outcomes and *in utero* environments were conducted, $N_{ENV} = 19$.

$$\mathbf{M}_i = (M_{i,1}, M_{i,2}, \dots, M_{i,N_{sample}}), \quad (5-1)$$

$$\text{and } i = 1, \dots, N_{CpG}, N_{CpG} = 1423, N_{sample} = 237$$

$$\mathbf{G}_j = (G_{j,1}, G_{j,2}, \dots, G_{j,N_{sample}}), \quad (5-2)$$

$$\text{and } j = 1, \dots, N_{all_SNPs}, N_{all_SNPs} = 708,6357$$

$$\mathbf{E}_k = (E_{k,1}, E_{k,2}, \dots, E_{k,N_{sample}}), \text{ and } k = 1, \dots, 19 \quad (5-3)$$

The control factor **Cvrt** was the category values for sex to differentiate male and female samples, eg. **Cvrt** = (Female, Male).

CpGs with the lowest AIC values in GxEmodel were segregated by three genotypic groups (AA, AB and BB) and regression against phenotype was performed in each genotypic group

5.3.1 Ethnicity associates with the first components of genotype but not DNA methylation

When the genotype data was subjected to principal component analysis (PCA)[235], the samples were cleanly separated by ethnicity. The Indian subjects were clustered away from Chinese and Malay subjects on the 1st component and the Chinese and Malay subjects were separated on the 2nd principal component (Figure 37A). In contrast when the DNA methylation data was subjected to principal component analysis, the samples did not separate well by ethnicity on component 1 or 2 (Figure 37B), suggesting inter-individual methylation is not just a product of genotypic difference.

5.3.2 Evidence for methQTLs

All 1423 VMR-CpG methylation levels were compared to all 708,365 heterogonous SNP genotypes by GEM_Gmodel. The best associated SNP (with lowest p-value) for each VMR-CpG was retained.

The p-values obtained from the regression analysis were binned into 1000 equally spaced bins. Distributions were defined as skewed if the first bin contained more than the expected 708 p-values in each bin that was expected if p-values were distributed evenly. 966 (68%) of the individual VMR-CpGs had a skewed p-value distribution. In addition 1037 (73%) had a $-\log_{10}(p)$ value above 6.5, which was the background

noise level across the genome. The final dataset of 1423 CpG-SNP pairs showed a range of associative p-values skewed towards the low end, suggestive of methQTLs in the dataset Figure 38.

Figure 38 Genotype is significantly associated with VMR-CpGs. Histogram shows the frequency of p-value distribution from regression analysis of 1423 VMR-CpGs against all SNPs.

Twelve of the 1423 CpG-SNP pairs (<1%) included a SNP located within the CpG that either creates or eliminates a CpG site. These instances were denoted as “disrupting pairs”. 828 of the 1423 pairs (58%) included a SNP and CpG from the same chromosome, which were denoted as “*cis* pairs”. The remaining 583 pairs (41%) include a CpG and SNP located on different chromosomes and were denoted as “*trans* pairs”. The proportion of *cis* pairs was much greater than expected by chance (chance would predict pairs to be equally distributed across chromosomes at ~4.5%)

The strengths of association between methylation level and genotype among *cis*-, *trans*- and disrupting CpG-SNP pairs were investigated further (Figure 39). Disrupting pairs had significantly stronger associations than *cis* pairs, which in turn had stronger associations than *trans* pairs (p-value = 7.3×10^{-117} by Kruskal-Wallis test, the `kruskal.test()` function in R [171]).

Figure 39 Strength of association between genotype and methylation is strongest for disrupting pairs and weakest for *trans* pairs.

5.3.3 GxE model best explained the variation in methylation at most VMR-CpGs

The GUSTO study has collected multiple environmental measures that were known to influence or reflect the quality of the *in utero* experience. We used nineteen parameters as surrogate measures of the uterine environment and asked whether methylation at each of 1423 VMR-CpGs was better explained by genotype (G), environment (E) or the interaction between gene and environment (GxE) by competing the AIC values from a total 39 GEM models with 19 GEM_Emodels, 1 GEM_Gmodel and 19 GEM_GxEmodels.

Methylation levels at ~25% of the 1423 VMR-CpGs were best explained by genotype alone, whilst the rest were best explained by GxE models. The models containing environment alone were never the best explanation of methylation at the VMR-CpGs. For the 75% of VMRs best explained by the interaction of genotype with *in utero* environments, the environments included maternal smoking, maternal depression, maternal BMI, infant birth weight, gestational age and birth order. This

finding combined the considerable published evidence for both the effects of environmental factors in pregnancy on both epigenetic states [124, 127, 128, 236-247] and health outcomes [248] and for genotypic influence on methylation. The results in this project suggest that genotype exerts a moderating influence on such environmental effects, and the interplay of genotype and environmental effects influence development.

The GxE model assumes that methylation is associated with environment in each of the genotypic subgroups by different coefficients. However, we observed that in many cases, environment and methylation were very closely associated in one genotype and less so in the other two genotypes. This could be the result of a genotypic group which is more sensitive to environment and methylomes accordingly. To identify CpG-VMRs that fit the sensitive genotype pattern, the subjects were segregated by the genotype in the best GxE model and stratified regressions were run for methylation and environment in the best GxE model. The association of environment and methylation was significant after Bonferroni correction for 50 VMR-CpGs. Examples in Figure 40 showed VMR-CpGs whose methylation levels are significantly associated with phenotype in only one genotypic group. Phenotypic values are shown on the *x*-axis, and methylation values in percentages on the *y*-axis. Samples are colored by their genotypic group (red indicates AA; blue, AB; black, BB) in left column, and a straight fit line is fitted for each group. In the right column, genotypic subgroup with the highest R^2 showed for different environment factors.

One reason why the GxE models might fail to explain the majority of the variance for more VMR-CpGs is that the association between methylation and a specific environmental condition may only be apparent in one genotypic subgroup. This observation is consistent with the emerging view that genotype can determine the degree of environmentally-induced phenotypic plasticity, i.e. so called ‘plasticity genes’ [249, 250] and epigenetic mechanisms serve to maintain environmentally-induced phenotypic variation [121, 251]. This is a particularly important finding for studies of either population differences or the influence of environmental factors on phenotypic outcomes. The finding of GxE interaction effects on the epigenome suggests that the necessary level of interrogation extends beyond simple EWAS analysis to include genotype. The failure to include assessment of genotypic moderation of environment – epigenome relations might result in an underestimation of the potential for environmental impact among sub-populations. Patel et al. (2013)

[252] showed that candidate SNP and CpG loci with marginal associations in GWAS and EWAS respectively can show strong associations with disease (in this case type 2 diabetes) when combined.

Figure 40 Examples of VMR-CpGs whose methylation levels are significantly associated with phenotype in only one genotypic group. Phenotypic values are shown on the x-axis, and methylation value in percentages on the y-axis. (Left) Data for all

samples. Samples are colored by their genotypic group (red indicates AA; blue, AB; black, BB), and a straight fit line is fitted for each group. (Right) Genotypic subgroup with the highest R^2 .

5.3.4 CpGs with the most variation across samples are most likely to be driven by genotype

CpG MAD score across samples or the range of methylation values across samples was positively related to the strength of association between the CpG methylation values and the genotype of the best associated SNP (Figure 41, $p = 3.54 \times 10^{-39}$). This suggested the more variable CpGs are more likely to be within methQTL (Figure 41). The 1423 VMR-CpGs chosen to lie within VMRs, as defining methylation variation by region, improved the specificity of analysis [162].

Figure 41 VMR-CpGs with larger ranges of methylation values are more likely to be methQTLs. Scatter plot showing the range of methylation values at each CpG across samples (x-axis) compared to the strength of association between the VMR-CpG methylation values and the genotype of the best SNP (y-axis).

5.4 Discussion and conclusion

Genotype (G) is accepted to be largely set at conception. In humans, *de novo* single-nucleotide variation rate is estimated to range between $0.82 \sim 1.70 \times 10^{-8}$ mutations per base per generation [253]. A study in 66 healthy monozygotic (MZ) adult twins at

506,786 SNPs found only two *de novo* somatic mutations to be present in white blood cells, resulting in a frequency of 1.2×10^{-7} mutations per nucleotide [254]. Another study using whole-exome sequencing on nine MZ twin pairs with genotypic discordances was unable to show any validated somatic mutation [255]. Copy number variants (CNVs) as a substantial part of genetic variation have relatively high mutation rates, however a genome-wide survey for post-twinning *de novo* CNVs in 1,097 MZ twin pairs identified and validated only 20 post-twinning *de novo* CNVs [256]. These findings provide direct evidence that early somatic point mutations do occur but the rate of mutation is low.

However, methylation (M) is accepted to be mostly set during gestation (after conception) and is available to be changed by genotype (G->M) [257]. In an experiment [258], to define sequence contribution to DNA methylation states, 50 different DNA elements were inserted into mouse stem cells. Results demonstrated that proximal sequence elements are both necessary and sufficient for regulating DNA methylation [258]. Therefore, G is set when the egg is fertilized while M can be changed during gestation and at all stages of development. Therefore in this thesis, G->M is more likely the explanation for the majority of sites where G and M are associated. However, I acknowledge that M->G is possible but suggest that this must be rare during the life course. Our findings from comparison modeling in methylome and genotype data in neonate cohorts [4, 5] showed that the rate of G->M is about 25%. Others' studies on monozygotic (MZ) and dizygotic twins (DZ) [67, 148, 186, 202, 259-261] have concordantly shown 20-30% epigenetic changes were directly driven by genetic variants. Similar results can be also found in a recent result from van Dongen et. al [262]. The directionality of association between methylation and environmental influences is unclear and complicated by considerations such as tissue heterogeneity and I agree with the reviewer that although we can show that methylation states are associated with the environmental influences we can not definitely say whether E->M or M->E.

In this chapter, I provided evidence for the influence of genotype on inter-individual variability in neonate DNA methylation alternations. Using GEM_Gmodel, methQTLs were readily apparent in the methylome of neonates. The findings were consistent with a recent report [263] of methQTLs across the neonatal epigenome in umbilical cord tissues and cord bloods. They detected 14,189 ~ 32,318 methQTLs (3 ~7%) from

~450,000 CpGs by searching the SNPs (genotype 2.5 million) in 5 kb flanking regions either side of each CpG.

Cellular heterogeneity may also compromise the ability to account for variance across the methylome. The umbilical cords included a mixture of cell types. Recent studies in blood suggest that some of the inter-individual variation in methylation is accounted for by differences in cellular content [31, 101]. The cell types in the GUSTO umbilical cords were unable to be separated as the tissues were frozen at collection, which was a limitation of this study. However, a recently published study on adult human blood which did adjust for cell counts estimated a remarkably similar proportion of inter-individual variation in methylation, to be genotype dependent (19%)[262].

The one million SNPs studied on the OmniExpress and exome arrays is only a subset of the SNPs in the genome. Missing SNPs could have accounted for additional variance in the results. However the one million SNPs were distributed relatively evenly across the genome and represented the majority of haplotype blocks. Finally, the Infinium 450K arrays used tend to cover coding genes and so the study in this chapter will have had less opportunity to cover variable methylation in gene deserts or non-coding genes.

The analysis was intended to estimate the relative, potential influence of genotypic, environmental and gene x environment interactive effects on the neonatal methylome. The models should not be considered as fully determinative of specific outcomes. Indeed, many of the environmental factors considered here are inter-dependent. Parity and maternal age, for example, are obviously correlated. Moreover, other potential sources of environmental influence, such as socio-economic status, were not considered. Similarly, the associations shown in Figure 40 derive from models that did not control for potentially important covariates. In addition, due to the number of predictive factors tested it is difficult to ascribe significance to any individual association. Instead we report the model that best explains the variance in methylation values. Nevertheless our data serve to underscore the importance of GxE interactions and suggest that models of epigenetic variation should consider such interactive influences.

The work described in this chapter is the world's first attempt to quantify the relative influence of genotype and environment, and their interaction on the human epigenome. Modelling based GEM tools facilitate the research successfully. The results strongly

suggested that computational modelling empowers biomedical discovery especially when it is hypothesis-free in 'omics' data. GEM based modelling answered fundamental questions about how the product of human evolution (genetic factors) interact with the individual circumstances that human are born into, to shape the well-being.

6 Extending the story of *HIF3A* association with adiposity to before birth

6.1 Introduction

A recent large scale EWAS analysis with replication in independent cohorts, reported that methylation levels at three CpGs in the first intron of the Hypoxia Inducible Factor 3A (*HIF3A*) loci (cg22891070, cg27146050, and cg16672562) were positively associated with body mass index (BMI) as measured in adult Caucasian whole blood and adipose tissues [29]. The authors postulated that *HIF3A* has a role in acquired obesity, perhaps through regulating adipocyte differentiation.

The study also showed that two SNPs located within the first intron (rs3826795) and upstream (rs8102595) of the *HIF3A* locus have independent associations with methylation at cg22891070, one of the three BMI associated *HIF3A* CpGs [29]. However rs3826795 and rs8102595 genotypes at the SNPs did not predict BMI in any of the reported cohorts. Therefore, the authors suggest that increased DNA methylation at the *HIF3A* locus is a consequence of increasing BMI and not causal to BMI or a product of a confounding factor which independently drives both BMI and *HIF3A* methylation. As previous studies have shown the origins of obesity could begin in the womb [237], these findings motivated us to examine the relation between *HIF3A* CpGs and SNPs and adiposity among neonates in the GUSTO birth cohort [176]. The GUSTO birth cohort was designed, in part, to determine epigenetic marks which associate with early life environments and later phenotypes.

In this study, a series of GEM like models were developed to examine in 991 GUSTO individuals, whether umbilical cord methylation levels at the CpGs of *HIF3A* identified in the previous study [29] associate with neonate birth weight and adiposity. Specifically, the three *HIF3A* CpGs were investigated for their association with infant weight, height and adiposity across the first two years of life. In agreement with the previous finding [29], we find that *HIF3A* DNA methylation at birth associates with measures of infant size and adiposity, the association of *HIF3A* DNA methylation and adiposity must begin in gestation or previous generations. This may have significance for either mechanistic and/or biomarker studies.

6.2 Study design and data

6.2.1 GUSTO participants and methylation/genotype data processing

A total of 991 mother-offspring from the GUSTO cohort was used for this analysis. The GUSTO participants have been described in greater detail in previous chapters and published in [4], the methylation and genotype dataset are from 450K and Omniexpress arrays respectively. Data processing and quality control were presented in the previous chapter.

Three CpGs in the first intron of the *HIF3A* locus (cg22891070, cg27146050, and cg16672562) and two SNPs located within the first intron (rs3826795) and upstream (rs8102595) of the *HIF3A*, shown in Figure 42, were studied in 991 infant umbilical cord tissue samples.

Figure 42 Location of methylation probes associated and SNPs in the *HIF3A* locus reported and studied.

We have compared our study design and findings to those from [29] in Table 21.

Table 21 Comparison of our study design with those from Dick et al [29].

	Dick et al (2014) [29]	Our work [5]
Data cohort	479 adult European individuals in discovery cohort; 339 and 1789 European individuals in two replication cohorts respectively.	991 Asian neonates from GUSTO birth cohort, across three different ethnic groups (Chinese, Malay and Indian).
Tissue	Blood and adipose.	Umbilical cord.
Phenotype	BMI	Weight*, length*, BMI* (*measured at birth, 6, 12, 18 and 24months of age), subscapular skinfold, triceps skinfold and the ratio of these two (birth, 18 and 24 months).
CpGs and SNPs	Three CpGs (cg22891070, cg27146050 and cg16672562) and two SNPs (rs3826795 and rs8102595) within the HIF3A locus.	Same CpGs and SNPs.
Platform	Infinium450K	Same platform

To study the association between *HIF3A* and infant weight and adiposity, six anthropometry measures including weight, length, BMI, subscapular skinfold, triceps skinfold and the ratio of subscapular to triceps skinfold were investigated as offspring outcomes. We also investigated the association between maternal pregnancy health status and methylation and infant outcomes. These 25 variables included maternal micronutrient levels, life style and glucose levels (see Table 22). All analyses were adjusted for gestational age (continuous), child sex (binary), ethnicity (category), and cell types (continuous). Characteristics of participants are shown in Table 22 and Table 23 and analysis models are shown in Figure 43.

Table 22 Characteristics of participants in the GUSTO cohort studied in the analysis.

Characteristic	Time point	n (%)	Mean (SD)
Ethnicity:	Delivery		
Chinese		568 (57)	
Malay		251 (25)	
Indian		172 (17)	
Child sex:	Delivery		
Male		530 (53)	
Female		461 (47)	
Gestational age (weeks)	Delivery	961	38.9 (1.0)
Weight (g)	Delivery	961	3150 (376)
	6 months	872	7744 (919)
	12 months	848	9405 (1073)

	18 months	808	10770 (1304)
	24 months	820	11996 (1542)
Length (cm)	Delivery	916	49 (2)
	6 months	876	67 (3)
	12 months	850	75 (3)
	18 months	695	82 (3)
	24 months	720	88 (4)
BMI(g/cm^2)	Delivery	961	1.3 (0.1)
	6 months	872	1.7 (0.2)
	12 months	847	1.6 (0.1)
	18 months	693	1.6 (0.1)
	24 months	720	1.6 (0.1)
Subscapular skinfold (mm)	Delivery	962	5.0 (1.2)
	18 months	673	6.4 (1.4)
	24 months	760	6.4 (1.6)
Triceps skinfold (mm)	Delivery	963	5.5 (1.2)
	18 months	714	8.6 (1.7)
	24 months	734	8.8 (1.8)
Subscap:triceps	Delivery	962	0.93 (0.17)
	18 months	648	0.76 (0.14)
	24 months	723	0.73 (0.14)
cg27146050	Delivery	991	15% (4%)
cg16672562			19% (5%)
cg22891070			31 (6%)
rs8102595 minor allele frequency (G)			0.22
rs3826795 minor allele frequency (T)			0.41

Table 23 Maternal status and maternal conditions

	Prenatal environmental variable	time point
maternal antenatal lifestyle	1_job (yes/no)	during pregnancy
	2_time_spent_for_sleeping(hour)	during pregnancy
	3_gentle_exercise (yes/no)	during pregnancy
	4_gentle_exercise_period (hour)	during pregnancy
	5_consuming_special_diet (yes/no)	during pregnancy
	6_drinking_alcohol (yes/no)	before pregnancy
	7_drinking_alcohol (yes/no)	during pregnancy
	8_smoked_regularly (yes/no)	before pregnancy
	9_smoked_regularly (yes/no)	during pregnancy
	10_anyone_living_in_home_smoke_daily (yes/no)	during pregnancy
	11_job_with_smoking_exposure (yes/no)	during pregnancy
maternal anthropometry	12_maternal_height (cm)	26-28 weeks
	13_weight_gain (kg)	26-26 weeks self-reported weight at enrolment
	14_prepregBMI (kg/m ²)	Pre-pregnant weight self-reported at enrolment
maternal plasma micronutrients	15_VitD (nmol/L)	26-28 weeks
	16_VitB12 (pg/ml)	26-28 weeks
	16_VitB12_categorized (0: <300 pg/ml, 1: ≥300 pg/ml)	26-28 weeks
	17_Copper_(ug/L)	26-28 weeks
	18_Iron (ug/L)	26-28 weeks
	18_Iron_Categorized (0: <560 ug/L, 1: ≥560 ug/L)	26-28 weeks
	19_Zinc (ug/L)	26-28 weeks
	20_Folate (ng/ml)	26-28 weeks gestation
maternal glucose level	21_ogtt_26w_2hr (mmol/L)	26-28 weeks
	22_ogtt_26w_fast (mmol/L)	26-28 weeks
others	23_mother_age (year)	first clinic visit
	24_birthorder (1/other)	delivery
	25_mode_of_delivery (normal/LSCS)	delivery

6.2.2 Estimation of cell fractions from methylation data

DNA methylation levels were measured in frozen umbilical cords that are made up of multiple cell types. CpG site methylation levels vary by cell types, and differences in cellular content between samples affect methylation levels independent of inter-

individual variation within cell types [101, 264]. Using a methodology similar to that applied for blood samples [103], we calculated methylation signatures for the three main cell types found in umbilical cord using published data from umbilical cord isolated homogenous cell types [263]. We used these signatures to estimate cellular content in the 991 GUSTO umbilical cords. The estimated cellular proportions (only two of the three variables were used to prevent collinearity) were then adjusted as covariates in all regression models that included methylation as one of the variables.

6.3 Statistical analysis and results

The data and association studies were illustrated in Figure 43. In this study, the new factor Y as the adiposity measurement was introduced as the outcomes of methylation changes. GEM models were used to examine the associations of *HIF3A* methylation, genotype and prenatal environment factors. For example, function F_5 to explain *HIF3A* methylation changes by 25 prenatal factors is GEM_Emodel, function F_2 for exploring genotype and methylation relation is GEM_Gmodel and function F_4 is GEM_GxEmodel to find the interaction between genotype and birth adiposity outcomes. Besides, a series GEM like models were developed to predict birth outcomes by *HIF3A* methylation as the explanatory variables, like function F_1 and F_3 , in Table 24

Matrix Y presents six birth adiposity measures ($i=1$ to 6 for weight, length, BMI, subscapular skinfold, triceps skinfold and the ratio of subscapular to triceps skinfold), and methylation at *HIF3A* CpG sites is M_j ($j=1,2,3$ for the three CpGs). The covariates cvt_k ($k=1,2,3,4$, for four cofounders) and 25 environment factors E_l ($l = 1, \dots, 25$) are also listed in Table 24.

Figure 43 Study design for HIF3A: data and modelling.

Table 24 Models in HIF3A analysis

The infant anthropometric outcomes were represented as Y_i ($i = 1, 2, 3, 4, 5, 6$) for weight, length, BMI, subscapular skinfold, triceps skinfold and the ratio of subscapular to triceps skinfold. HIF3A methylation was M_j ($j=1,2,3$) for three CpGs and the genotyping for two SNPs as G_n ($n = 1, 2$) for two SNPs using additive minor allele effect as the value 0 for AA, 1 for AB and 2 for BB. Maternal status and conditions were denoted as E_l ($l = 1, \dots, 25$) and all models were adjusted by four covariates as sex, ethnicity, gestational age and cell types, as cvrt_k ($k = 1, 2, 3, 4$) for four confounders.		
Model	Objectives	Equation index
$\text{lm}(Y \sim M + \text{cvrt})$	To test if HIF3A methylation can predict the anthropometric outcomes	F_1
$\text{lm}(M \sim G + \text{cvrt})$	To test if HIF3A CpGs is a methQTL	F_2
$\text{lm}(Y \sim G + \text{cvrt})$	To test if birth outcomes driven by the HIF3A <i>cis</i> -genotype	F_3
$\text{lm}(M \sim Y \times G + \text{cvrt})$	To test if the interaction between adiposity and HIF3A genotyping can explain methylation variants.	F_4
$\text{lm}(M \sim E + \text{cvrt})$	To test the association between prenatal environment and HIF3A methylation	F_5
$\text{lm}(Y \sim E + \text{cvrt})$	To test the association between prenatal environment and offspring outcomes	F_6

Association between umbilical cord methylation at three sites in *HIF3A* and child weight measured at 0, 6, 12, 18 and 24 months were tested by general estimation. Regression coefficients (Est) and 95% confidence interval (CI) were reported as percentage change in child weight for every 10% increase in methylation level. P-values are two-sided. Using generalized estimating equations (implemented by *gee* function in R scripts) to account for repeated measures, log-transformed weight was regressed against methylation at each CpG site, adjusting for fixed effects of time, child sex, ethnicity, cell type proportions and interaction between ethnicity and proportions. Time was coded using a binary variable for each distinct time point and interaction terms of time with all variables (methylation, child sex, ethnicity, cell type proportions and interactions between ethnicity and proportions) were included.

6.3.1 *HIF3A* methylation in umbilical cords was positively associated with birth weight and adiposity

Higher umbilical cord *HIF3A* methylation level at all three CpGs was associated with greater infant birth weight, after adjusting for infant sex, ethnicity, estimated cellular proportions. Results generated from model F_1 are presented in Table 25. There were also significant associations with adiposity at birth measured by BMI and, for two of the three CpGs, by subscapular skinfold and the ratio of subscapular to triceps skinfold.

Table 25 Association between methylation at three *HIF3A* CpGs and neonatal anthropometric outcomes at birth and gestational age.

	cg27146050			cg16672562			cg22891070		
	Est	95% CI	p-value	Est	95% CI	p-value	Est	95% CI	p-value
Birth Weight (g)	3.61	(0.68, 6.63)	0.015	3.34	(1.4, 5.32)	0.00068	2.05	(0.32, 3.82)	0.02
Birth Length (cm)	0.6	(-0.29, 1.5)	0.19	0.46	(-0.13, 1.05)	0.13	0.35	(-0.18, 0.88)	0.2
BMI at birth (g/cm ²)	2.38	(0.23, 4.58)	0.03	2.4	(0.97, 3.84)	0.00096	1.35	(0.07, 2.64)	0.039
Subscapular skinfold at birth (mm)	-0.77	(-5.97, 4.72)	0.78	5.44	(1.76, -9.25)	0.0035	3.27	(0, 6.64)	0.05
Triceps skinfold at birth (mm)	-1.28	(-6.42, 4.14)	0.64	0.75	(-2.75, 4.38)	0.68	0.25	(-2.90, -3.51)	0.88
Ratio of Subscapular to triceps at birth	0.5	(-3.85, 5.04)	0.83	4.67	(1.66, 7.77)	0.0022	3.02	(0.34, 5.78)	0.027

For every 10% increase in methylation of cg27146050, cg16672562 and cg22891070, birth weight was 3.61% (95% CI: 0.68–6.63), 3.34% (1.4–5.32) and 2.05% (0.32–3.82) higher, respectively (Table 25). These effect sizes are slightly smaller than the corresponding effect estimates on adult BMI of 7.8% (5.1–10.4), 3.2% (2.0–4.4) and 3.6% (2.4–4.9) as previously reported in [29].

6.3.2 The association of *HIF3A* methylation and birth weight was independent of the contribution of *cis*-acting SNPs; but the association was stronger in one genotype group.

We genotyped the SNPs identified previously [29] as *cis*-influencing on *HIF3A* methylation in the GUSTO cohort. The effect allele of rs8102595 is G with allele frequency 0.22 and the effect allele of rs3826795 is T with allele frequency 0.41. The two SNPs were associated with the three CpG's methylation. The association of two SNPs at the *HIF3A* with methylation at three loci in *HIF3A* was done by linear regression model as F_2 , methylation at each CpG site is regressed against each SNP. Regression coefficients and standard errors reflect change in methylation beta values per copy of the effect allele, adjusting for ethnicity, gender and cell types, as in Table 26.

The tests on genotyping and the six outcomes were conducted by F_3 and none of the SNPs was associated with outcomes. Table 27 showed neither SNP was associated with birth weight. The associations of birth weight with methylation at all three sites were similar when adjusted for genotype at both SNPs (Table 5 in [5].)

Table 26 Association of two *cis*- SNPs (rs3826795, rs8102595) at *HIF3A* with methylation at three sites in *HIF3A*.

	cg27146050			cg16672562			cg22891070		
	Est	95% CI	p-value	Est	95% CI	p-value	Est	95% CI	p-value
CpG ~ rs8102595	1.2	(0.96, 1.5)	5.7E-19	2.9	(2.5, 3.3)	1.6E-45	3.2	(2.8, 3.7)	5.7E-47
CpG ~ rs3826795	0.64	(0.4, 0.87)	1.1E-07	1.6	(1.2, 1.9)	3.2E-18	1.5	(1.1, 1.9)	3.2E-14

Table 27 Association between SNPs at *HIF3A* locus and birth weight

	BW		
	Est	95% CI	p-value
rs8102595	0.39	(-0.89, -1.7)	0.55
rs3826795	0.059	(-1.0, 1.2)	0.92

The results from testing on *HIF3A* methylation with the interaction of genotype and birth weight, as in model F_4 showed the associations were only significant at p-value <0.05 for two of the CpGs and rs3826795 (Table 28). In the CC genotype group, birth weight was more strongly correlated with *HIF3A* methylation, as shown in Figure 44.

Table 28 Two sided p-values from testing for interaction term between log-transformed birth weight and SNPs, with methylation as outcomes, adjusting for cofounders.

model	cg27146050	cg16672562	cg22891070
	Pvalue	Pvalue	Pvalue
M ~ log(BW) × rs8102595	0.32	0.077	0.30
M ~ log(BW) × rs3826795	0.34	0.019	0.033

Figure 44 Association between methylation at cg16672562 and birth weight, stratified by rs3826795 genotypes. (A) Scatter plots of birth weights against methylation. Each panel displays data of each genotype group. (B) Box plot of rs3826795 genotype against methylation.

6.3.3 None of the prenatal environmental variables examined was consistently associated with HIF3A methylation and birth weight.

In order to examine whether *HIF3A* methylation and birth weight were related to the prenatal environment, we examined for significant associations with both *HIF3A* methylation and birth weight for a range of factors related to the prenatal environment, both as main effects and in interaction with *HIF3A* genotype. Factors investigated included antenatal lifestyle factors, such as activity level and smoking, maternal prepregnancy BMI and pregnancy weight gain, maternal glucose tolerance during pregnancy indicative of gestational diabetes, maternal micronutrients levels and maternal age. Together there were 25 prenatal environment status and measurements (Table 23). There were some suggestive associations, for example, maternal glucose tolerance was significantly associated with *HIF3A* methylation at one of the CpGs and birth weight. Furthermore, maternal weight gain during pregnancy was significantly associated both with infant birth weight and *HIF3A* methylation interaction with genotype. However, there was no prenatal environmental variable that showed consistently significant association with all three *HIF3A* sites and birth weight and none of them yielded substantial effect size.

6.3.4 *HIF3A* methylation levels at birth predicted infant weight in the first two years of life

The longitudinal analysis of the associations between anthropometric outcomes and *HIF3A* methylation was conducted generalized estimation equation. Interestingly, *HIF3A* methylation levels at birth were also associated with weight at later ages, albeit possibly with slightly lower effect size, see Figure 45. The trends were maintained when the model adjusted for birth weight.

6.4 Discussion

Based on a series designed models, we explored the reported *HIF3A* methylation CpGs, SNPs with prenatal environmental factors and infant birth weight and adiposity status for 991 GUSTO mother-offspring pairs. Comparable results (shown in Table 29) demonstrated that this study extended the previous finding that *HIF3A* DNA methylation in adults associated with BMI to show that the link between *HIF3A* DNA methylation with weight and adiposity can be detected at birth. We suggest that as the association between *HIF3A* methylation and adiposity is detectable so early in life, *HIF3A* may be a potential biomarker of metabolic trajectory.

Table 29 Comparison of our study findings with those from Dick et al [29].

Dick et al, published in The LANCET (March13, 2014) [29]	Our work [5]	Comparison
For every 0.1 unit (10%) methylation increase, adult BMI increased by 7.8% (5.1-10.4), 3.2% (2.0-4.4) and 3.6% (2.4-4.9) for cg27146050 and cg16672562, cg22891070 respectively.	For every 0.1 unit (10%) increase in methylation value, birth weight increased by 3.61% (95% CI: 0.68-6.63), 3.34% (1.4-5.32) and 2.05% (0.32-3.82) for cg27146050, cg16672562 and cg22891070 respectively.	Effect sizes in both studies are comparable.
Both cis SNPs were associated with three methylation sites, but not significantly associated with BMI.	Both cis SNPs were associated with three methylation sites, but not significantly associated with birth weight.	Comparable.

Future studies should track both *HIF3A* methylation and offspring outcomes longitudinally during gestation, at birth and at subsequent ages to determine whether the association of *HIF3A* methylation with adiposity is consistent over life course. As blood and adipose tissues are invasive to sample in infants, *HIF3A* methylation levels should be investigated in other sample types such as saliva, buccal swabs or circulating fetal cells obtained prenatally. Longitudinal studies can help determine, if *HIF3A* is a signature for obesity risk and assess the utility of using *HIF3A* methylation to evaluate the efficacy of intervention as lifestyle modification. Future studies should also examine the impact of *HIF3A* methylation changes on gene expression in relevant tissues.

7 Summary and future work

7.1 What we know now, we didn't know then

When this work was started in 2012, there was empirical evidence that

1. Methylation marks could be measured on a genome wide scale [15]. After Infinium 450K [81] array was generated in year 2011, it quickly replaced its forerunners and became a well-accepted platform for DNA methylation profiling of clinical samples and widely adopted in EWAS.
2. Early life environment influenced later life phenotype [119]. The “Barker hypothesis”, also known as the “fetal programming hypothesis” emphasized that early environment, especially fetal life was the “developmental origins of health and disease” (DOHaD) [57, 119] .
3. Some epigenetic marks were influenced by environment [129]. Studies searching for new associations were typically limited to hypothesis-driven candidate CpGs. For example, a study on 68 candidate CpGs screened on five neonates’ umbilical cords found hypermethylation of a *RXRA* CpG at chr9:136355885 was associated with lower maternal carbohydrate intake in early pregnancy. A subset of environment associated epigenetic marks could predict later phenotype, for example the same *RXRA* CpG was linked with higher neonatal adiposity [129]
4. Methylation marks are influenced by *in cis* genotype. In a study of 27,000 CpGs from human brain tissues with variation at 537,411 SNPs, Gibbs [144] found 4%~5% of the total CpGs were influenced by SNPs within 1MB region in the same chromosomal region (*cis*-). A few other studies had similar findings and some of the methQTLs were independently discovered across different studies [146, 258]

This thesis contributed (with many others) to the current state of our knowledge. We now know:

1. Genome wide methylation marks can be measured robustly and at high resolution on arrays [1] and via sequencing modalities that improve coverage [2]. Data generated from such is prone to technical artifact but these can be corrected through careful processing.
2. The computational barriers inherent in the analysis of large-scale multi-dimensional combined genetic, clinical and epigenetic data can be overcome. Data can be efficiently processed using standard computational infrastructure by GEM [3].

3. Methylation marks that are variable among individuals are reflective of both environmental influences and *in cis* genotype and most are the results of an interaction between the two [4]. Van Dongen [262] recently showed that along with genotype and environment, the interaction between sex and age can shape the inter-individual differences in DNA methylomes.
4. Methylation marks in the *HIF3A* and *VEGFB* genes may be on the causal pathway to obesity and metabolic syndrome [5, 9, 265]. They may also contribute to the inter-generational transmission of obesity [266].

Therefore, we have expanded the evidence base for epigenetic marks as promising biomarkers for developmental disease trajectories. We now have the computational wherewithal to conduct EWAS for a wide range of environments and diseases.

7.2 Summarization and contributions of this thesis

The chapters in this thesis discussed insights into data analysis methods of diverse technologies surrogating DNA methylation and comprehensive modellings of epigenetic processes in associating with genetic variants and environment. Two major contributions have been done in this work, which including computational methodologies and epigenetic discoveries. Data analysis in epigenetic measurements cross different technology platforms were improved by comparing measurements from clinical samples, and GEM package was developed to enable epigenetic studies in much larger scale with multiple dimensional data. Based on the developed methods, we explored neonate methylome, genetic variants and developmental environment, and we found DNA methylation levels of *HIF3A* as the potential biomarkers in adiposity development.

There are seven chapters cross the whole thesis started from introduction in **Chapter 1** and ended by summarization in **Chapter 7**. Five other chapters in between the introduction and summarization presented five projects that were done during the study, which included processing the data to achieve minimized bias and high quality, exploring the data relationship with comprehensive methodologies and employing the methodologies to answer questions with result interpretation.

Chapter 2 presented a novel pipeline in processing and normalizing DNA methylation measured in Infinium 450K technology. The pipeline was validated by concordance study for same seven clinical samples with RRBS technology [1]. The experiment by 450K for the clinical samples suggested that batch effects were prevalent and data

normalization was very challenging. We then studied the technology and proposed a pipeline with two major processes, color adjustment and typeI/typeII correction, and accomplished the aims to eliminate the impact of technical defects and reduce the batch effect. This study represented the largest dataset for evaluation of the 450K with RRBS, and the first pipeline proposed to improve 450K analysis, at the time when results were published [1] and it has been cited by more than eighteen other publications based PubMed record.

Next generation sequencing is an advanced technology in DNA methylation detection and Methyl Capture Seq (MC-Seq) is with good coverage and designed targets for CpGs. In **Chapter 3**, we explored the latest DNA methylation profiling, MC-Seq based on comparison between MC-Seq and 450K for same seven buccal epithelium samples. Results confirmed MC-Seq could be an attractive cost-effective alternative to whole-genome bisulfite sequencing and improves on the lower level of methylome coverage offered by the Infinium array. We were the first in evaluating the technical performance of MC-Seq in human buccal DNA samples. The comparison of 450K and MC-Seq was comprehensively across key functional genomic regions in the context of their coverage and concordance of methylation calls in clinical samples. Finally, as proof-of-principle, in a small sample, the utility of MC-Seq was examined in studying inter-individual methylome variation and distinguishing subjects from different ethnic groups.

After the technologies were well studied and the methodologies for data processing were ready, in **Chapter 4**, a fast and efficient tool suite named GEM [3] was created to explore the associations of Gene, Environment and Methylation. It provides three models denoted GEM_Emodel, GEM_Gmodel and GEM_GxEmodel to facilitate analyses in EWAS. The GEM_Emodel tests the association of methylome marks and environmental factors; the GEM_Gmodel creates a methQTL genome-wide map; finally, the GEM_GxE model tests the ability of gene and environmental interaction models to predict DNA methylation levels. GEM was evolved into a R package with efficient implementation which was based on matrix operation. GEM R package was then hosted in BioConductor <http://bioconductor.org/packages/GEM/> [3] We benchmarked the performance of the GEM operations by the methylation, genotyping and environmental factors from publically available data. Our results demonstrated that the GEM package can facilitate reliable EWAS analyses within minutes, in a standard computational setting (processor=2.2GHz, RAM=8G, system = window7 64bit).

Based on GEM models, we studied the relationships among environment, genotype and methylation in 237 GUSTO infants and their mothers in **Chapter 5**. This chapter focused on the examination of the influence of genotypic (G), environmental (E) and gene x environment (G x E) effects on 237 multi-ethnicity Asian neonatal methylomes [4] by statistical modelling. We found methylation quantitative trait loci (methQTLs) were readily detected in variable methylated regions. Further, genotype alone is the best explanation for about 25% of the methylation changes. About 75% of methylation changes were best explained by the interaction of genotype with different *in utero* environments, including maternal smoking, maternal depression, maternal BMI, infant birth weight, gestational age and birth order. Lastly, segregation studies the VMRs best explained by GxE demonstrated differential effects on the methylome in different genetic groups, which suggested the hypothesis that certain genetic groups are more sensitive to the effects of particular environment and methylate according. This was the first EWAS study to offer insights on the complex relationship between genetic inheritance and individual prenatal experience in interpreting neonatal epigenetic variation. In particular this work extended that genetic variation can moderate the relationship between environmental observations and epigenetic status.

Chapter 6 is based on our publication in [5]. Motivated by a DNA methylation biomarker, *HIF3A*, which was reported with association with adult BMI by Dick et al (2014) [29], a series GEM models were used to examine the *HIF3A* methylation CpGs and SNPs with prenatal environmental factors and infant birth weight and adiposity status for 991 GUSTO mother-offspring pairs. Our study in umbilical cord DNA replicated the existing findings [29], and extended the reported association of *HIF3A* DNA methylation and adiposity to before birth. Results suggested that as the association between *HIF3A* methylation and adiposity is detectable so early in life, *HIF3A* may be a potential biomarker of metabolic trajectory.

7.3 Limitations in this study

7.3.1 The issue of tissue heterogeneity

DNA methylation patterns are tissue specific. A recent study on DNA methylation signatures in cord found that methylation at 49.2% of CpG sites on the array differed by cell type [267]. The clinical samples conducted in this study are neonates from GUSTO birth cohort and DNA methylation profiling was majorly based on umbilical cord tissues. It has been well known that cell type can influence the epigenome [268].

The caveat is relevant to our discovery of association among genetic variants, methylome changes and environment factors [4]. Umbilical cords were frozen at collection so no fractionation by cell type was possible. The confounding contribution of cell type heterogeneity and *in silico* methods for cell type devolution were not widely accepted at the time of the study. However, as the influence of genotype was clearly shown, cellular heterogeneity did not substantially mask inter-individual methylation patterns. It may be that the environmental influence was under-estimated as environmentally-driven epigenetic patterns are more likely to differ between cell types than genetically driven ones.

With the availability of a reference data set consisting of DNA methylation measurements on isolated cell types, it is possible to impute cell proportions and adjust for them. By the time we conducted our follow-up study of *HIF3A* methylation association with adiposity [5], we practiced careful control for cellular heterogeneity. We applied three methods, one is a reference method [103], and two reference free methodologies [104, 269] for adjusting association of DNA methylation in tissues for cellular components. The estimated cell fractions were then adjusted as covariates in the regression models.

Reference-free algorithms can provide more flexibility when reference sets are incomplete or unavailable [270]. However, existing in-silico methods are restricted in dealing with the complexity of the nature of samples. Cell type separation on relevant sample will definitely provide golden parameters for association modelling and better interpretation of associations evident from DNA methylation.

7.3.2 Limitations in computational analysis

In this study we proposed data analysis pipeline for Infinium 450K array, applied linear regression in association studies and created a software package to operate epigenome wide association study in an efficient and fast way. All methodologies were well accepted in the research community, though there are quite a lot of imperfections.

(1) The pipeline in this study is dependent on GenomeStudio

Our published algorithm [1] requires signal intensities exported from Illumina's GenomeStudio before correction for colour bias and the different dynamic ranges of the two types of probes included within Infinium arrays. This is in contrast to other

comprehensive methylation data packages subsequently developed such as “Minfi” [193]. Minfi uses subset quantile within array normalization (SWAN [183]) to perform corrections for probe type, which in our hands performed as well or slightly worse than our correction. However, the dependence of our tools for GenomeStudio data exporting has been recently overcome by publication of the “illumio” tool [193].

(2) Covariates in linear regression is too complicated to be carefully addressed

Covariate adjustment in the linear regression is widely used in studying the association between methylation changes with maternal condition and infant birth outcomes (E), genotype variants (G) and the interaction between environmental factors and genotype (GxE) in our study. There are continuous covariates (like maternal age and gestational age) and categorical covariates (like infant gender, ethnicity) and the misspecification between covariates and outcomes can lead to reduced power and potentially incorrect conclusions [271]. In the study of HIF3A, we found the estimated cell types were also related to ethnicity of the samples. So instead of using two independent covariates in the linear model, we used the interaction between “ethnicity” and “celltypes” to surrogate the associations these two factors [5]. We suggested much comprehensive test among all known parameters before the design of the association models to avoid collinearity.

(3) Missing methylation values can not be included in matrix based operations and imputation could introduce error.

Missing values are commonly present in methylation profiles. One reason for missing methylation values is typed CpGs vary from different technologies. For instance, Table2 in Chapter 2 for describes the comparison of 450K and RRBS data for the same tissues and samples, and showed there were data on only 100,000CpGs from both technologies, out of 450,000 measured by 450K and 6,000,000 by RRBS. The number decreased to 65,000 after quality control. In Chapter 3, the jointly measured CpGs between Infinium450K and MC-Seq numbered 395,000 (Figure 19 in Chapter3). Quality control and removal of probes with problematic signals produces missing values as well. In the DNA methylation screening of 1000+ GUSTO umbilical cords on Infinium 450K, 255,000 CpGs remained after discarding CpGs with detection p-value less than 0.05 in any sample, invariable CpGs that shows less than 0.1 methylation change across samples, and those on the sex chromosomes. Among the

255,000 CpGs that were tested, 18% of them had at least one sample with a missing value. Missing values cannot be included in matrix multiplication. A common approach is to exclude CpGs with any missing values but this causes loss of a large proportion of the data. In GEM we used the mean value across all samples with data to impute missing values. This method produced data very similar to the regression analysis (which can handle missing data) in the benchmark study on 1000 samples [3] in Chapter 4. However, when sample size is small, mean value imputation introduces error.

Recent research has attempted to translate the techniques from comprehensive genotyping imputation [272] and gene expression prediction [273] to methylation data. Zhang *et al* [274] applied penalized functional regression models [275] to impute missing methylation data across Infinium27K and Infinium450K datasets. Ma *et al* [276] used support vector machines to predict missing methylation profiles across different tissues. Both were able to validate their results by simulation studies. However, Zhang *et al* [274] assumed that DNA methylation exhibits local correlation among neighboring probes and this assumption does not always hold, especially when methylation values at a particular locus varied across individuals. Similarly, Ma *et al* make the assumption that CpG methylation values correlate across tissues in [276] and this does not hold especially for sites of tissue-specific methylation, which are extensive. Data based imputation methods should always be collaborated with biological domain knowledge. Recently, gene ontology was used to measure the expression similarity for microRNA expression data imputation [277], yet the similar practice to impute methylation has not been seen. In my opinion, the method presented in Zhang *et al* [274] has utility to impute data missing from Infinium 450K arrays relative to Infinium 850K, for comparison of the same tissues, however further evaluation is warranted in using it to impute missing methylation values between microarray and NGS technologies.

7.4 Future work

The EWAS studies described here using the Infinium450K platform are limited by the fact that we are studying only 450K CpGs, which is less than 2% of total ~28 million CpGs present in the genome. In GUSTO, the newly developed Infinium850K [96] will be applied for longitudinal sampling in different tissues, which enlarges the accessibility of existing DNA methylation detection from regulatory elements to

enhancers, though it is still not able to disclose the full picture of human epigenome. At the same time, evolving NGS technologies continually shape the way of methylation detection in genomic coverage and research specificity, which also requests computational methods to be robust across data platforms. The comparisons of Infinium450K with RRBS and MC-Seq in clinical samples demonstrated the efforts are intensive and non-trivial [1, 2]. My future work will be continued in technology exploration and optimal pipeline development to suit the data analysis demands in EWAS.

Epigenetics is the putative mediator that links genetics and environments to downstream phenotypes, by influencing gene functions. Future work could expand our mechanistic understanding to the consequence of altered DNA methylation on gene expression and subsequently health outcomes, see

Figure 46. Correlation of epigenetic marks to transcriptome has been conducted by our collaborators on a locus by locus basis e.g. quantitative PCR was used to assess *CXCL14* transcript levels and correlate the epigenetic state of *CXCL14* and suboptimal birth outcomes [7]. However, these approaches are candidates driven and employed on a limited scale. RNA sequencing (RNA-seq) empowers an in-depth view of the whole transcriptome [124] and the evolving technologies in RNA-seq can detect dynamic range of expression from the vast diversity of clinically relevant RNA species [278]. Therefore, EWAS studies must be combined with RNA-seq datasets in order to construct a complete pathway to investigate how changes in the methylome influence phenotype. Tools to combine the genome, epigenome and transcriptome may have a similar structure to the GEM tools described in this thesis and be an extension of them. However they will need adaptation to reflect the fact that whilst the epigenome can influence the transcriptome, the transcriptome can also influence the epigenome. Extension of the GEM suite of tools to perform high-throughput causal analysis and perhaps mendelian randomization [279] is warranted.

Figure 46 Schematic diagram illustrating a putative causal pathway for how the effects of genome and environment on phenotype are transmitted through the epigenome and transcriptome.

Lack of statistical power is a continuing problem for EWAS, especially for developmental phenotypes and non-communicable diseases which include a small signal-to-noise ratio as epigenetic assay is noisy but the expected effect sizes are subtle (compared for example to the dramatic epigenetic changes seen in tumor tissues). The number of variables that can be measured in methylome analyses far exceed typical sample sizes available in birth cohorts like GUSTO. Increasing the sample size is not feasible and very costly, as epigenetic sampling measures not only molecular level profiling like DNA methylation and genotype, but also very detailed environment status for the individuals, like social economic status, living style and healthy conditions. Cross-omics analyses will help improve statistical power. In addition, my future work is designed to address these challenges by reducing noise and increasing specificity through statistical modeling with efficient computational operation.

Large international and interdisciplinary studies are beginning to provide a good framework to better understand gene expression regulation in humans. ENCODE has examined 147 cell lines to search for functional elements in the genome; the Epigenomics Roadmap examined 127 tissue biopsies from embryos and adults, both diseased and healthy, with a focus on identifying functional regulatory elements; the International Human Epigenome Consortium and the Genome Tissue Expression project (GTEx) explores the relationship between genetic variation and tissue-specific gene expression and alternate splicing by sequencing whole genomes and examining gene expression in over 20,000 tissue samples from 900 individuals. Recent work from a UK group for the accessible resource for integrated epigenomics studies [280] inspired me. My future work will include shared data services with comprehensive epigenome browser for methylation profiles from different tissues, population or phenotype equipped with CpG query, more visualization function and bioinformatics tools to facilitate for better research. There are 1079 GUSTO mothers, 732 fathers and 1073 infants genotyped by Illumina Omniexpress+exome genotyping arrays. DNA methylation

profiles were screened in 1000+ umbilical cord tissues, 350+ maternal bloods and 350+ cord bloods. We have also results from GWAS, EWAS and GxE with regards to 1000+ clinical data including maternal condition, birth outcomes and longitudinal data for development. With all these different dimensional data, a systemized way to manage data in a genome browser style (<https://genome.ucsc.edu>) will be very helpful for data sharing and enable in depth study.

Epigenetics is exciting. Understanding the mechanisms that regulate gene expression and function provides an opportunity for targeted therapeutic intervention. DNA methylation biomarkers have great potential to contribute to personalized medicine throughout life though there are still a lot hurdles. Multi-disciplinary teams including biomedical researchers and bioinformaticians are essential, especially when data is big and signal is complex by nature. A deeper understanding of the complex inter-relationships between environment, behavior, physiological response and epigenetics will impact human health and with increasing knowledge comes optimism.

Reference

1. Pan H, Chen L, Dogra S, Teh AL, Tan JH, Lim YI, Lim YC, Jin S, Lee YK, Ng PY *et al*: **Measuring the methylome in clinical samples: improved processing of the Infinium Human Methylation450 BeadChip Array.** *Epigenetics : official journal of the DNA Methylation Society* 2012, **7**(10):1173-1187.
2. Teh AL, Pan H, Lin X, Lim YI, Patro CP, Cheong CY, Gong M, MacIsaac JL, Kwoh CK, Meaney MJ *et al*: **Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples.** *Epigenetics : official journal of the DNA Methylation Society* 2016, **11**(1):36-48.
3. Pan H, Holbrook JD, Karnani N, Kwoh CK: **Gene, Environment and Methylation (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment.** *BMC Bioinformatics* 2016, **17**:299.
4. Teh AL, Pan H, Chen L, Ong ML, Dogra S, Wong J, MacIsaac JL, Mah SM, McEwen LM, Saw SM *et al*: **The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes.** *Genome research* 2014, **24**(7):1064-1074.
5. Pan H, Lin X, Wu Y, Chen L, Teh AL, Soh SE, Lee YS, Tint MT, MacIsaac JL, Morin AM *et al*: **HIF3A association with adiposity: the story begins before birth.** *Epigenomics* 2015, **7**(6):937-950.
6. Stunkel W, Pan H, Chew SB, Tng E, Tan JH, Chen L, Joseph R, Cheong CY, Ong ML, Lee YS *et al*: **Transcriptome changes affecting Hedgehog and cytokine signalling in the umbilical cord: implications for disease risk.** *PloS one* 2012, **7**(7):e39744.
7. Cheong CY, Chng K, Lim MK, Amrithraj AI, Joseph R, Sukarieh R, Chee Tan Y, Chan L, Tan JH, Chen L *et al*: **Alterations to DNA methylation and expression of CXCL14 are associated with suboptimal birth outcomes.** *Journal of human genetics* 2014, **59**(9):504-511.
8. Chen L, Pan H, Tuan TA, Teh AL, MacIsaac JL, Mah SM, McEwen LM, Li Y, Chen H, Broekman BF *et al*: **Brain-derived neurotrophic factor (BDNF) Val66Met polymorphism influences the association of the methylome with maternal anxiety and neonatal brain volumes.** *Development and psychopathology* 2015, **27**(1):137-150.
9. Huang RC, Garratt ES, Pan H, Wu Y, Davis EA, Barton SJ, Burdge GC, Godfrey KM, Holbrook JD, Lillycrop KA: **Genome-wide methylation analysis identifies differentially methylated CpG loci associated with severe obesity in childhood.** *Epigenetics : official journal of the DNA Methylation Society* 2015, **10**(11):995-1005.
10. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al*: **Whole-genome**

- random sequencing and assembly of Haemophilus influenzae Rd.** *Science* 1995, **269**(5223):496-512.
11. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome.** *Science* 2001, **291**(5507):1304-1351.
 12. Venter JC, Smith HO, Adams MD: **The Sequence of the Human Genome.** *Clinical chemistry* 2015, **61**(9):1207-1208.
 13. Siggens L, Ekwall K: **Epigenetics, chromatin and genome organization: recent advances from the ENCODE project.** *Journal of internal medicine* 2014, **276**(3):201-214.
 14. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ *et al*: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518**(7539):317-330.
 15. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL *et al*: **High density DNA methylation array with single CpG site resolution.** *Genomics* 2011, **98**(4):288-295.
 16. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A: **Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling.** *Nat Protoc* 2011, **6**(4):468-481.
 17. Hitchler MJ, Rice JC: **Genome-wide epigenetic analysis of human pluripotent stem cells by ChIP and ChIP-Seq.** *Methods in molecular biology* 2011, **767**:253-267.
 18. Li Q, Suzuki M, Wendt J, Patterson N, Eichten SR, Hermanson PJ, Green D, Jeddeloh J, Richmond T, Rosenbaum H *et al*: **Post-conversion targeted capture of modified cytosines in mammalian and plant genomes.** *Nucleic acids research* 2015, **43**(12):e81.
 19. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R: **Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.** *Nucleic acids research* 2005, **33**(18):5868-5877.
 20. Mohn F, Weber M, Schubeler D, Roloff TC: **Methylated DNA immunoprecipitation (MeDIP).** *Methods in molecular biology* 2009, **507**:55-64.
 21. Waddington CH: **An introduction to modern genetics;** 1939.
 22. Haig D: **The (dual) origin of epigenetics.** *Cold Spring Harb Symp Quant Biol* 2004, **69**:67-70.
 23. Russo VE, Martienssen RA, Riggs AD: **Epigenetic mechanisms of gene regulation:** Cold Spring Harbor Laboratory Press; 1996.

24. Bernstein BE, Meissner A, Lander ES: **The mammalian epigenome.** *Cell* 2007, **128**(4):669-681.
25. Sarda S, Hannenhalli S: **Next-generation sequencing and epigenomics research: a hammer in search of nails.** *Genomics Inform* 2014, **12**(1):2-11.
26. Heard E, Martienssen RA: **Transgenerational epigenetic inheritance: myths and mechanisms.** *Cell* 2014, **157**(1):95-109.
27. Bonasio R, Tu S, Reinberg D: **Molecular signals of epigenetic states.** *Science* 2010, **330**(6004):612-616.
28. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Grealley JM, Gut I, Houseman EA, Izzi B, Kelsey KT, Meissner A *et al*: **Recommendations for the design and analysis of epigenome-wide association studies.** *Nat Methods* 2013, **10**(10):949-955.
29. Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aissi D, Wahl S, Meduri E, Morange PE, Gagnon F, Grallert H *et al*: **DNA methylation and body-mass index: a genome-wide analysis.** *Lancet* 2014.
30. Hidalgo B, Irvin MR, Sha J, Zhi D, Aslibekyan S, Absher D, Tiwari HK, Kabagambe EK, Ordovas JM, Arnett DK: **Epigenome-wide association study of fasting measures of glucose, insulin, and HOMA-IR in the Genetics of Lipid Lowering Drugs and Diet Network study.** *Diabetes* 2014, **63**(2):801-807.
31. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M *et al*: **Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis.** *Nat Biotechnol* 2013, **31**(2):142-147.
32. Mill J, Heijmans BT: **From promises to practical strategies in epigenetic epidemiology.** *Nat Rev Genet* 2013, **14**(8):585-594.
33. Holliday R, Pugh JE: **DNA modification mechanisms and gene activity during development.** *Science* 1975, **187**(4173):226-232.
34. Kouzarides T: **Chromatin modifications and their function.** *Cell* 2007, **128**(4):693-705.
35. Berger SL: **The complex language of chromatin regulation during transcription.** *Nature* 2007, **447**(7143):407-412.
36. Li B, Carey M, Workman JL: **The role of chromatin during transcription.** *Cell* 2007, **128**(4):707-719.
37. Mattick JS, Makunin IV: **Non-coding RNA.** *Hum Mol Genet* 2006, **15 Spec No 1**:R17-29.
38. **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**(7414):57-74.

39. Heidari N, Phanstiel DH, He C, Grubert F, Jahanbani F, Kasowski M, Zhang MQ, Snyder MP: **Genome-wide map of regulatory interactions in the human genome.** *Genome research* 2014, **24**(12):1905-1917.
40. Li S: **Implication of posttranslational histone modifications in nucleotide excision repair.** *International journal of molecular sciences* 2012, **13**(10):12461-12486.
41. Kebede AF, Schneider R, Daujat S: **Novel types and sites of histone modifications emerge as players in the transcriptional regulation contest.** *The FEBS journal* 2015, **282**(9):1658-1674.
42. Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, Landry J, Kauer M, Tackett AJ, Chait BT, Badenhorst P *et al*: **A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling.** *Nature* 2006, **442**(7098):86-90.
43. Davidovich C, Cech TR: **The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2.** *RNA (New York, NY)* 2015, **21**(12):2007-2022.
44. Margueron R, Trojer P, Reinberg D: **The key to development: interpreting the histone code?** *Current opinion in genetics & development* 2005, **15**(2):163-176.
45. Struhl K: **Histone acetylation and transcriptional regulatory mechanisms.** *Genes Dev* 1998, **12**(5):599-606.
46. Merschbaecher K, Haettig J, Mueller U: **Acetylation-mediated suppression of transcription-independent memory: bidirectional modulation of memory by acetylation.** *PloS one* 2012, **7**(9):e45131.
47. Graff J, Tsai LH: **Histone acetylation: molecular mnemonics on the chromatin.** *Nature reviews Neuroscience* 2013, **14**(2):97-111.
48. Dirks RA, Stunnenberg HG, Marks H: **Genome-wide epigenomic profiling for biomarker discovery.** 2016, **8**:122.
49. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129**(4):823-837.
50. Furey TS: **ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions.** *Nat Rev Genet* 2012, **13**(12):840-852.
51. Sun W, Poschmann J, Cruz-Herrera Del Rosario R, Parikshak NN, Hajan HS, Kumar V, Ramasamy R, Belgard TG, Elanggovan B, Wong CC *et al*: **Histone Acetylome-wide Association Study of Autism Spectrum Disorder.** *Cell* 2016, **167**(5):1385-1397.e1311.
52. Chen LL, Yang L: **Regulation of circRNA biogenesis.** *RNA biology* 2015, **12**(4):381-388.

53. Saeidimehr S, Ebrahimi A, Saki N, Goodarzi P, Rahim F: **MicroRNA-Based Linkage between Aging and Cancer: from Epigenetics View Point.** *Cell journal* 2016, **18**(2):117-126.
54. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, Sabo PJ, Lu Y, Rohs R, Stamatoyannopoulos JA *et al*: **Probing DNA shape and methylation state on a genomic scale with DNase I.** *Proc Natl Acad Sci U S A* 2013, **110**(16):6376-6381.
55. Wermann H, Stoop H, Gillis AJ, Honecker F, van Gurp RJ, Ammerpohl O, Richter J, Oosterhuis JW, Bokemeyer C, Looijenga LH: **Global DNA methylation in fetal human germ cells and germ cell tumours: association with differentiation and cisplatin resistance.** *The Journal of pathology* 2010, **221**(4):433-442.
56. Rakyan VK, Down TA, Balding DJ, Beck S: **Epigenome-wide association studies for common human diseases.** *Nat Rev Genet* 2011, **12**(8):529-541.
57. Feinberg AP: **Epigenetics at the epicenter of modern medicine.** *JAMA* 2008, **299**(11):1345-1350.
58. Ptashne M: **Epigenetics: core misconception.** *Proc Natl Acad Sci U S A* 2013, **110**(18):7101-7103.
59. Bird A: **The essentials of DNA methylation.** *Cell* 1992, **70**(1):5-8.
60. Kim H, Jang WY, Kang MC, Jeong J, Choi M, Sung Y, Park S, Kwon W, Jang S, Kim MO *et al*: **TET1 contributes to neurogenesis onset time during fetal brain development in mice.** *Biochemical and biophysical research communications* 2016, **471**(4):437-443.
61. Rajabi H, Mohseni-Kouchesfehani H, Mohammadi-Sangcheshmeh A, Farifteh-Nobijari F, Salehi M: **Pronuclear epigenetic modification of protamine deficient human sperm following injection into mouse oocytes.** *Systems biology in reproductive medicine* 2016, **62**(2):125-132.
62. Guo JU, Su Y, Shin JH, Shin J, Li H, Xie B, Zhong C, Hu S, Le T, Fan G *et al*: **Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain.** *Nature neuroscience* 2014, **17**(2):215-222.
63. Holliday R: **Epigenetics: a historical overview.** *Epigenetics : official journal of the DNA Methylation Society* 2006, **1**(2):76-80.
64. Horvath LM, Li N, Carrel L: **Deletion of an X-inactivation boundary disrupts adjacent gene silencing.** *PLoS genetics* 2013, **9**(11):e1003952.
65. Zamudio N, Bourc'his D: **Transposable elements in the mammalian germline: a comfortable niche or a deadly trap?** *Heredity* 2010, **105**(1):92-104.
66. Bartolomei MS, Ferguson-Smith AC: **Mammalian genomic imprinting.** *Cold Spring Harbor perspectives in biology* 2011, **3**(7).

67. Ollikainen M, Smith KR, Joo EJ, Ng HK, Andronikos R, Novakovic B, Abdul Aziz NK, Carlin JB, Morley R, Saffery R *et al*: **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* 2010, **19**(21):4176-4188.
68. Jung M, Pfeifer GP: **Aging and DNA methylation.** *BMC biology* 2015, **13**:7.
69. Teschendorff AE, Liu X, Caren H, Pollard SM, Beck S, Widschwendter M, Chen L: **The dynamics of DNA methylation covariation patterns in carcinogenesis.** *PLoS computational biology* 2014, **10**(7):e1003709.
70. Holliday R: **Epigenetics comes of age in the twentyfirst century.** *Journal of genetics* 2002, **81**(1):1-4.
71. Jones PA: **Functions of DNA methylation: islands, start sites, gene bodies and beyond.** *Nat Rev Genet* 2012, **13**(7):484-492.
72. Jones PA, Takai D: **The role of DNA methylation in mammalian epigenetics.** *Science* 2001, **293**(5532):1068-1070.
73. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes.** *Journal of molecular biology* 1987, **196**(2):261-282.
74. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB *et al*: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454**(7205):766-770.
75. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D: **A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA.** *Cell* 1985, **40**(1):91-99.
76. Saxonov S, Berg P, Brutlag DL: **A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.** *Proc Natl Acad Sci U S A* 2006, **103**(5):1412-1417.
77. Deaton AM, Bird A: **CpG islands and the regulation of transcription.** *Genes Dev* 2011, **25**(10):1010-1022.
78. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP: **Orphan CpG islands identify numerous conserved promoters in the mammalian genome.** *PLoS genetics* 2010, **6**(9):1001134.
79. Bird A: **DNA methylation patterns and epigenetic memory.** *Genes Dev* 2002, **16**(1):6-21.
80. Smith AA, Huang YT, Eliot M, Houseman EA, Marsit CJ, Wiencke JK, Kelsey KT: **A novel approach to the discovery of survival biomarkers in glioblastoma using a joint analysis of DNA methylation and gene expression.** *Epigenetics : official journal of the DNA Methylation Society* 2014, **9**(6):873-883.

81. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F: **Evaluation of the Infinium Methylation 450K technology.** *Epigenomics* 2011, **3**(6):771-784.
82. Fouse SD, Nagarajan RO, Costello JF: **Genome-scale DNA methylation analysis.** *Epigenomics* 2010, **2**(1):105-117.
83. Lin X, Barton S, Holbrook JD: **How to make DNA methylome wide association studies more powerful.** *Epigenomics* 2016, **8**(8):1117-1129.
84. Branco MR, Ficz G, Reik W: **Uncovering the role of 5-hydroxymethylcytosine in the epigenome.** *Nat Rev Genet* 2012, **13**(1):7-13.
85. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y *et al*: **Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications.** *Nat Biotechnol* 2010, **28**(10):1097-1105.
86. Beck S: **Taking the measure of the methylome.** *Nat Biotechnol* 2010, **28**(10):1026-1028.
87. Laird PW: **Principles and challenges of genomewide DNA methylation analysis.** *Nat Rev Genet* 2010, **11**(3):191-203.
88. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, Klotzle B, Bibikova M, Fan JB, Gao Y *et al*: **Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates.** *Mol Cell* 2013, **49**(2):359-367.
89. Saffery R, Gordon L: **Time for a standardized system of reporting sites of genomic methylation.** *Genome biology* 2015, **16**:85.
90. Lee YK, Jin S, Duan S, Lim YC, Ng DP, Lin XM, Yeo G, Ding C: **Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples.** *Biological procedures online* 2014, **16**(1):1.
91. Lienhard M, Grimm C, Morkel M, Herwig R, Chavez L: **MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments.** *Bioinformatics* 2014, **30**(2):284-286.
92. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M: **Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome.** *Epigenetics : official journal of the DNA Methylation Society* 2011, **6**(6):692-702.
93. Engel SM, Joubert BR, Wu MC, Olshan AF, Haberg SE, Ueland PM, Nystad W, Nilsen RM, Vollset SE, Peddada SD *et al*: **Neonatal genome-wide methylation patterns in relation to birth weight in the Norwegian Mother and Child Cohort.** *American journal of epidemiology* 2014, **179**(7):834-842.
94. **International network of cancer genome projects.** *Nature* 2010, **464**(7291):993-998.

95. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nat Genet* 2013, **45**(10):1113-1120.
96. Moran S, Arribas C, Esteller M: **Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences.** *Epigenomics* 2016, **8**(3):389-399.
97. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, Doucet D, Thomas NJ, Wang Y, Vollmer E *et al*: **High-throughput DNA methylation profiling using universal bead arrays.** *Genome research* 2006, **16**(3):383-393.
98. Yu M, Hon GC, Szulwach KE, Song CX, Jin P, Ren B, He C: **Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine.** *Nat Protoc* 2012, **7**(12):2159-2170.
99. Touleimat N, Tost J: **Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation.** *Epigenomics* 2012, **4**(3):325-341.
100. Liang L, Cookson WO: **Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies.** *Hum Mol Genet* 2014, **23**(R1):R83-88.
101. Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, Kobor MS: **Factors underlying variable DNA methylation in a human community cohort.** *Proc Natl Acad Sci U S A* 2012, **109** Suppl 2:17253-17260.
102. Jaffe AE, Irizarry RA: **Accounting for cellular heterogeneity is critical in epigenome-wide association studies.** *Genome biology* 2014, **15**(2):2014-2015.
103. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT: **DNA methylation arrays as surrogate measures of cell mixture distribution.** *BMC Bioinformatics* 2012, **13**:86.
104. Houseman EA, Molitor J, Marsit CJ: **Reference-free cell mixture adjustments in analysis of DNA methylation data.** *Bioinformatics* 2014, **30**(10):1431-1439.
105. Feinberg AP: **Epigenomics reveals a functional genome anatomy and a new approach to common disease.** *Nat Biotechnol* 2010, **28**(10):1049-1052.
106. Wessely F, Emes RD: **Identification of DNA methylation biomarkers from Infinium arrays.** *Frontiers in genetics* 2012, **3**:161.
107. Jones PA, Baylin SB: **The fundamental role of epigenetic events in cancer.** *Nat Rev Genet* 2002, **3**(6):415-428.

108. Feinberg AP, Tycko B: **The history of cancer epigenetics.** *Nature reviews Cancer* 2004, **4**(2):143-153.
109. Sandoval J, Esteller M: **Cancer epigenomics: beyond genomics.** *Current opinion in genetics & development* 2012, **22**(1):50-55.
110. Kitchen MO, Bryan RT, Haworth KE, Emes RD, Luscombe C, Gommersall L, Cheng KK, Zeegers MP, James ND, Devall AJ *et al*: **Methylation of HOXA9 and ISL1 Predicts Patient Outcome in High-Grade Non-Invasive Bladder Cancer.** *PloS one* 2015, **10**(9):e0137003.
111. Jayaraman S: **Epigenetics of autoimmune diabetes.** *Epigenomics* 2011, **3**(5):639-648.
112. Maier S, Olek A: **Diabetes: a candidate disease for efficient DNA methylation profiling.** *The Journal of nutrition* 2002, **132**(8 Suppl):2440S-2443S.
113. Kucukali CI, Kurtuncu M, Coban A, Cebi M, Tuzun E: **Epigenetics of Multiple Sclerosis: An Updated Review.** *Neuromolecular medicine* 2014.
114. Akbarian S: **Epigenetics of schizophrenia.** *Current topics in behavioral neurosciences* 2010, **4**:611-628.
115. Rando TA: **Epigenetics and aging.** *Experimental gerontology* 2010, **45**(4):253-254.
116. Feinberg AP: **The epigenetic basis of common human disease.** *Trans Am Clin Climatol Assoc* 2013, **124**:84-93.
117. Bird A: **Perceptions of epigenetics.** *Nature* 2007, **447**(7143):396-398.
118. Hanson M, Godfrey KM, Lillycrop KA, Burdge GC, Gluckman PD: **Developmental plasticity and developmental origins of non-communicable disease: Theoretical considerations and epigenetic mechanisms.** *Prog Biophys Mol Biol* 2011.
119. Barker DJ, Eriksson JG, Forsen T, Osmond C: **Fetal origins of adult disease: strength of effects and biological basis.** *Int J Epidemiol* 2002, **31**(6):1235-1239.
120. Gluckman PD, Hanson MA: **Maternal constraint of fetal growth and its consequences.** *Semin Fetal Neonatal Med* 2004, **9**(5):419-425.
121. Meaney MJ, Ferguson-Smith AC: **Epigenetic regulation of the neural transcriptome: the meaning of the marks.** *Nature neuroscience* 2010, **13**(11):1313-1318.
122. Seckl JR, Meaney MJ: **Early life events and later development of ischaemic heart disease.** *Lancet* 1993, **342**(8881):1236.
123. Waterland RA, Michels KB: **Epigenetic epidemiology of the developmental origins hypothesis.** *Annual review of nutrition* 2007, **27**:363-388.

124. Boney CM, Verma A, Tucker R, Vohr BR: **Metabolic syndrome in childhood: association with birth weight, maternal obesity, and gestational diabetes mellitus.** *Pediatrics* 2005, **115**(3):e290-296.
125. Parets SE, Conneely KN, Kilaru V, Fortunato SJ, Syed TA, Saade G, Smith AK, Menon R: **Fetal DNA Methylation Associates with Early Spontaneous Preterm Birth and Gestational Age.** *PloS one* 2013, **8**(6):e67489.
126. Pilsner JR, Hu H, Ettinger A, Sanchez BN, Wright RO, Cantonwine D, Lazarus A, Lamadrid-Figueroa H, Mercado-Garcia A, Tellez-Rojo MM *et al*: **Influence of prenatal lead exposure on genomic methylation of cord blood DNA.** *Environ Health Perspect* 2009, **117**(9):1466-1471.
127. Roseboom TJ, van der Meulen JH, Ravelli AC, Osmond C, Barker DJ, Bleker OP: **Effects of prenatal exposure to the Dutch famine on adult disease in later life: an overview.** *Mol Cell Endocrinol* 2001, **185**(1-2):93-98.
128. Skilton MR, Viikari JS, Juonala M, Laitinen T, Lehtimäki T, Taittonen L, Kahonen M, Celermajer DS, Raitakari OT: **Fetal growth and preterm birth influence cardiovascular risk factors and arterial health in young adults: the Cardiovascular Risk in Young Finns Study.** *Arterioscler Thromb Vasc Biol* 2011, **31**(12):2975-2981.
129. Godfrey KM, Sheppard A, Gluckman PD, Lillycrop KA, Burdge GC, McLean C, Rodford J, Slater-Jefferies JL, Garratt E, Crozier SR *et al*: **Epigenetic gene promoter methylation at birth is associated with child's later adiposity.** *Diabetes* 2011, **60**(5):1528-1534.
130. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE, Lumey LH: **Persistent epigenetic differences associated with prenatal exposure to famine in humans.** *Proc Natl Acad Sci U S A* 2008, **105**(44):17046-17049.
131. McGowan PO, Sasaki A, D'Alessio AC, Dymov S, Labonte B, Szyf M, Turecki G, Meaney MJ: **Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse.** *Nature neuroscience* 2009, **12**(3):342-348.
132. Weaver IC, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, Seckl JR, Dymov S, Szyf M, Meaney MJ: **Epigenetic programming by maternal behavior.** *Nature neuroscience* 2004, **7**(8):847-854.
133. Haworth KE, Farrell WE, Emes RD, Ismail KM, Carroll WD, Hubball E, Rooney A, Yates AM, Mein C, Fryer AA: **Methylation of the FGFR2 gene is associated with high birth weight centile in humans.** *Epigenomics* 2014, **6**(5):477-491.
134. Monick MM, Beach SR, Plume J, Sears R, Gerrard M, Brody GH, Philibert RA: **Coordinated changes in AHRH methylation in lymphoblasts and pulmonary macrophages from smokers.** *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 2012, **159B**(2):141-151.

135. Philibert RA, Beach SR, Brody GH: **Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers.** *Epigenetics : official journal of the DNA Methylation Society* 2012, **7**(11):1331-1338.
136. Breitling LP, Salzmann K, Rothenbacher D, Burwinkel B, Brenner H: **Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease.** *European heart journal* 2012, **33**(22):2841-2848.
137. Zhang Y, Yang R, Burwinkel B, Breitling LP, Brenner H: **F2RL3 methylation as a biomarker of current and lifetime smoking exposures.** *Environ Health Perspect* 2014, **122**(2):131-137.
138. Fryer AA, Nafee TM, Ismail KM, Carroll WD, Emes RD, Farrell WE: **LINE-1 DNA methylation is inversely correlated with cord plasma homocysteine in man: a preliminary study.** *Epigenetics : official journal of the DNA Methylation Society* 2009, **4**(6):394-398.
139. Fryer AA, Emes RD, Ismail KM, Haworth KE, Mein C, Carroll WD, Farrell WE: **Quantitative, high-resolution epigenetic profiling of CpG loci identifies associations with cord blood plasma homocysteine and birth weight in humans.** *Epigenetics : official journal of the DNA Methylation Society* 2011, **6**(1):86-94.
140. Emes RD, Clifford H, Haworth KE, Farrell WE, Fryer AA, Carroll WD, Ismail KM: **Antiepileptic drugs and the fetal epigenome.** *Epilepsia* 2013, **54**(1):e16-19.
141. Wu P, Farrell WE, Haworth KE, Emes RD, Kitchen MO, Glossop JR, Hanna FW, Fryer AA: **Maternal genome-wide DNA methylation profiling in gestational diabetes shows distinctive disease-associated changes relative to matched healthy pregnancies.** *Epigenetics : official journal of the DNA Methylation Society* 2016:0.
142. Gonzalez-Rodriguez PJ, Xiong F, Li Y, Zhou J, Zhang L: **Fetal hypoxia increases vulnerability of hypoxic-ischemic brain injury in neonatal rats: role of glucocorticoid receptors.** *Neurobiology of disease* 2014, **65**:172-179.
143. Kruglyak L: **Prospects for whole-genome linkage disequilibrium mapping of common disease genes.** *Nat Genet* 1999, **22**(2):139-144.
144. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J *et al*: **Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain.** *PLoS genetics* 2010, **6**(5):e1000952.
145. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, Craig DW, Redman M, Gershon ES, Liu C: **Genetic control of individual differences in gene-specific methylation in human brain.** *Am J Hum Genet* 2010, **86**(3):411-419.
146. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK: **DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.** *Genome biology* 2011, **12**(1):R10.

147. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA *et al*: **DNA methylation profiling of human chromosomes 6, 20 and 22.** *Nat Genet* 2006, **38**(12):1378-1385.
148. Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, Buil A, Busche S, Yuan W, Nisbet J, Sekowska M *et al*: **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* 2013, **93**(5):876-890.
149. Zhang FF, Cardarelli R, Carroll J, Fulda KG, Kaur M, Gonzalez K, Vishwanatha JK, Santella RM, Morabia A: **Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood.** *Epigenetics : official journal of the DNA Methylation Society* 2011, **6**(5):623-629.
150. Fraser HB, Lam LL, Neumann SM, Kobor MS: **Population-specificity of human DNA methylation.** *Genome biology* 2012, **13**(2):R8.
151. Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, Myers J, Godley LA, Dolan ME, Zhang W: **Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits.** *Genetics* 2013, **194**(4):987-996.
152. Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, Monk D, Hata K, Marques-Bonet T, Wang L *et al*: **DNA methylation contributes to natural human variation.** *Genome research* 2013, **23**(9):1363-1372.
153. Yousefi M, Karmaus W, Zhang H, Ewart S, Arshad H, Holloway JW: **The methylation of the LEPR/LEPROT genotype at the promoter and body regions influence concentrations of leptin in girls and BMI at age 18 years if their mother smoked during pregnancy.** *International journal of molecular epidemiology and genetics* 2013, **4**(2):86-100.
154. Soto-Ramirez N, Arshad SH, Holloway JW, Zhang H, Schauburger E, Ewart S, Patil V, Karmaus W: **The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years.** *Clinical epigenetics* 2013, **5**(1):1.
155. Klengel T, Binder EB: **Allele-specific epigenetic modification: a molecular mechanism for gene-environment interactions in stress-related psychiatric disorders?** *Epigenomics* 2013, **5**(2):109-112.
156. Keller MC: **Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution.** *Biological psychiatry* 2014, **75**(1):18-24.
157. Bjornsson HT, Fallin MD, Feinberg AP: **An integrated epigenetic and genetic approach to common human disease.** *Trends in genetics : TIG* 2004, **20**(8):350-358.
158. Holbrook JD: **An epigenetic escape route.** *Trends in genetics : TIG* 2015, **31**(1):2-4.

159. Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, Daunay A, Busato F, Mein CA, Manfras B *et al*: **Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis.** *PLoS genetics* 2011, **7**(9):e1002300.
160. Murphy TM, Mill J: **Epigenetics in health and disease: heralding the EWAS era.** *Lancet* 2014, **383**(9933):1952-1954.
161. Ng JW, Barrett LM, Wong A, Kuh D, Smith GD, Relton CL: **The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities.** *Genome biology* 2012, **13**(6):246.
162. Ong ML, Holbrook JD: **Novel region discovery method for Infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways.** *Aging cell* 2014, **13**(1):142-155.
163. Ladd-Acosta C, Shu C, Lee BK, Gidaya N, Singer A, Schieve LA, Schendel DE, Jones N, Daniels JL, Windham GC *et al*: **Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood.** *Environmental research* 2016, **144**(Pt A):139-148.
164. Joubert BR, den Dekker HT, Felix JF, Bohlin J, Ligthart S, Beckett E, Tiemeier H: **Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns.** 2016, **7**:10577.
165. Haberg SE, London SJ, Nafstad P, Nilsen RM, Ueland PM, Vollset SE, Nystad W: **Maternal folate levels in pregnancy and asthma in children at age 3 years.** *The Journal of allergy and clinical immunology* 2011, **127**(1):262-264, 264.e261.
166. Kruithof CJ, Kooijman MN, van Duijn CM, Franco OH, de Jongste JC, Klaver CC, Mackenbach JP, Moll HA, Raat H, Rings EH *et al*: **The Generation R Study: Biobank update 2015.** *European journal of epidemiology* 2014, **29**(12):911-927.
167. Sun YV: **The Influences of Genetic and Environmental Factors on Methylome-wide Association Studies for Human Diseases.** *Current genetic medicine reports* 2014, **2**(4):261-270.
168. Lemire M, Zaidi SH, Ban M, Ge B, Aissi D, Germain M, Kassam I, Wang M, Zanke BW, Gagnon F *et al*: **Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci.** *Nature communications* 2015, **6**:6326.
169. Soh SE, Tint MT, Gluckman PD, Godfrey KM, Rifkin-Graboi A, Chan YH, Stunkel W, Holbrook JD, Kwek K, Chong YS *et al*: **Cohort profile: Growing Up in Singapore Towards healthy Outcomes (GUSTO) birth cohort study.** *Int J Epidemiol* 2014, **43**(5):1401-1409.
170. Barker DJ: **The developmental origins of insulin resistance.** *Horm Res* 2005, **3**:2-7.

171. **R: A language and environment for statistical computing** [<http://www.R-project.org/>]
172. Huber W, Carey VJ, Gentleman R, Anders S: **Orchestrating high-throughput genomic analysis with Bioconductor**. 2015, **12**(2):115-121.
173. Gu H, Bock C, Mikkelsen TS, Jager N, Smith ZD, Tomazou E, Gnirke A, Lander ES, Meissner A: **Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution**. *Nat Methods* 2010, **7**(2):133-136.
174. Hayatsu H, Wataya Y, Kai K, Iida S: **Reaction of sodium bisulfite with uracil, cytosine, and their derivatives**. *Biochemistry* 1970, **9**(14):2858-2865.
175. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM: **Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis**. *BMC Bioinformatics* 2010, **11**:587.
176. Soh SE, Tint MT, Gluckman PD, Godfrey KM, Rifkin-Graboi A, Chan YH, Stunkel W, Holbrook JD, Kwek K, Chong YS *et al*: **Cohort Profile: Growing Up in Singapore Towards healthy Outcomes (GUSTO) birth cohort study**. *Int J Epidemiol* 2013.
177. Bock C, Tomazou EM, Brinkman AB, Muller F, Simmer F, Gu H, Jager N, Gnirke A, Stunnenberg HG, Meissner A: **Quantitative comparison of genome-wide DNA methylation mapping technologies**. *Nat Biotechnol* 2010, **28**(10):1106-1114.
178. Roessler J, Ammerpohl O, Gutwein J, Hasemeier B, Anwar SL, Kreipe HH, Lehmann U: **Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc**. *BMC Res Notes* 2012, **5**(1):210.
179. Bland JM, Altman DG: **Statistical methods for assessing agreement between two methods of clinical measurement**. *Lancet* 1986, **1**(8476):307-310.
180. Rodgers JS, Nicewander WA: **Thirteen ways to look at the correlation coefficients**. *The American Statistician* 1988, **42**(1):59-66.
181. Zar JH: **Spearman rank correlation**. In: *Encyclopedia for Biostatistics*. Edited by Armitage P, Colton T: Wiley; 2005.
182. Feinberg AP: **Genome-scale approaches to the epigenetics of common human disease**. *Virchows Arch* 2010, **456**(1):13-21.
183. Maksimovic J, Gordon L, Oshlack A: **SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips**. *Genome biology* 2012, **13**(6):2012-2013.
184. Sun Z, Chai HS, Wu Y, White WM, Donkena KV, Klein CJ, Garovic VD, Therneau TM, Kocher JP: **Batch effect correction for genome-wide**

- methylation data with Illumina Infinium platform.** *BMC Med Genomics* 2011, **4**:84.
185. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8**(1):118-127.
 186. Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, Sommer IE, Kahn RS, Ophoff RA: **The relationship of DNA methylation with age, gender and genotype in twins and healthy controls.** *PloS one* 2009, **4**(8):e6767.
 187. El-Maarri O, Becker T, Junen J, Manzoor SS, Diaz-Lacava A, Schwaab R, Wienker T, Oldenburg J: **Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males.** *Hum Genet* 2007, **122**(5):505-514.
 188. Schroeder JW, Conneely KN, Cubells JC, Kilaru V, Newport DJ, Knight BT, Stowe ZN, Brennan PA, Krushkal J, Tylavsky FA *et al*: **Neonatal DNA methylation patterns associate with gestational age.** *Epigenetics : official journal of the DNA Methylation Society* 2011, **6**(12):1498-1504.
 189. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA: **Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.** *Int J Epidemiol* 2012, **41**(1):200-209.
 190. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F: **A comprehensive overview of Infinium HumanMethylation450 data processing.** *Briefings in bioinformatics* 2013, **29**:29.
 191. Du P, Kibbe WA, Lin SM: **lumi: a pipeline for processing Illumina microarray.** *Bioinformatics* 2008, **24**(13):1547-1548.
 192. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S: **A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data.** *Bioinformatics* 2013, **29**(2):189-196.
 193. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA: **Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.** *Bioinformatics* 2014, **30**(10):1363-1369.
 194. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S: **ChAMP: 450k Chip Analysis Methylation Pipeline.** *Bioinformatics* 2014, **30**(3):428-430.
 195. Morris TJ, Beck S: **Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data.** *Methods (San Diego, Calif)* 2015, **72**:3-8.

196. Harris RA, Nagy-Szakal D, Kellermayer R: **Human metastable epiallele candidates link to common disorders.** *Epigenetics : official journal of the DNA Methylation Society* 2013, **8**(2):157-163.
197. Warden CD, Lee H, Tompkins JD, Li X, Wang C, Riggs AD, Yu H, Jove R, Yuan YC: **COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis.** *Nucleic acids research* 2013, **41**(11):e117.
198. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, Kelsey KT, Marsit CJ, Houseman EA, Brown R: **Review of processing and analysis methods for DNA methylation array data.** *British journal of cancer* 2013, **109**(6):1394-1402.
199. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F: **A comprehensive overview of Infinium HumanMethylation450 data processing.** *Briefings in bioinformatics* 2014, **15**(6):929-941.
200. Glossop JR, Emes RD, Nixon NB, Haworth KE, Packham JC, Dawes PT, Fryer AA, Matthey DL, Farrell WE: **Genome-wide DNA methylation profiling in rheumatoid arthritis identifies disease-associated methylation changes that are distinct to individual T- and B-lymphocyte populations.** *Epigenetics : official journal of the DNA Methylation Society* 2014, **9**(9):1228-1237.
201. Ong ML, Tan PY, MacIsaac JL, Mah SM, Buschdorf JP, Cheong CY, Stunkel W, Chan L, Gluckman PD, Chng K *et al*: **Infinium monkeys: Infinium 450K array for the Cynomolgus macaque (*Macaca fascicularis*).** *G3 (Bethesda, Md)* 2014, **4**(7):1227-1234.
202. van Dongen J, Ehli EA, Sliker RC, Bartels M, Weber ZM, Davies GE, Slagboom PE, Heijmans BT, Boomsma DI: **Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells.** *Genes* 2014, **5**(2):347-365.
203. Choi M, Lee J, Le MT, Nguyen DT, Park S, Soundrarajan N, Schachtschneider KM, Kim J, Park JK, Kim JH *et al*: **Genome-wide analysis of DNA methylation in pigs using reduced representation bisulfite sequencing.** *DNA research : an international journal for rapid publication of reports on genes and genomes* 2015, **22**(5):343-355.
204. De Meyer T, Bady P, Trooskens G, Kurscheid S, Bloch J, Kros JM, Hainfellner JA, Stupp R, Delorenzi M, Hegi ME *et al*: **Genome-wide DNA methylation detection by MethylCap-seq and Infinium HumanMethylation450 BeadChips: an independent large-scale comparison.** *Scientific reports* 2015, **5**:15375.
205. Field SF, Beraldi D, Bachman M, Stewart SK, Beck S, Balasubramanian S: **Accurate measurement of 5-methylcytosine and 5-hydroxymethylcytosine in human cerebellum DNA by oxidative bisulfite on an array (OxBS-array).** *PloS one* 2015, **10**(2):e0118202.

206. Kato N, Loh M, Takeuchi F: **Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation.** 2015, **47**(11):1282-1293.
207. Warton K, Samimi G: **Methylation of cell-free circulating DNA in the diagnosis of cancer.** *Frontiers in molecular biosciences* 2015, **2**:13.
208. Zaghlool SB, Al-Shafai M, Al Muftah WA, Kumar P, Falchi M, Suhre K: **Association of DNA methylation with age, gender, and smoking in an Arab population.** *Clinical epigenetics* 2015, **7**(1):6.
209. Zheleznyakova GY, Nilsson EK, Kiselev AV, Maretina MA, Tishchenko LI, Fredriksson R, Baranov VS, Schiöth HB: **Methylation levels of SLC23A2 and NCOR2 genes correlate with spinal muscular atrophy severity.** *PloS one* 2015, **10**(3):e0121964.
210. Wang WY, Barratt BJ, Clayton DG, Todd JA: **Genome-wide association studies: theoretical and practical concerns.** *Nat Rev Genet* 2005, **6**(2):109-118.
211. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: **Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.** *Nature* 2008, **452**(7184):215-219.
212. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM *et al*: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462**:315-322.
213. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis.** *Cell* 2008, **133**(3):523-536.
214. Lee YK, Jin S, Duan S, Lim YC, Ng DP, Lin XM, Yeo GS, Ding C: **Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples.** *Biol Proced Online* 2014, **16**:1.
215. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R: **Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.** *Nucleic Acids Res* 2005, **33**(18):5868-5877.
216. Magdalena J, Goval JJ: **Methyl DNA immunoprecipitation.** *Methods in molecular biology* 2009, **567**:237-247.
217. Jacinto FV, Ballestar E, Esteller M: **Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome.** *BioTechniques* 2008, **44**(1):35, 37, 39 passim.
218. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM *et al*: **A Bayesian deconvolution strategy for**

- immunoprecipitation-based DNA methylome analysis.** *Nat Biotechnol* 2008, **26**(7):779-785.
219. Hing B, Ramos E, Braun P, McKane M, Jancic D, Tamashiro KL, Lee RS, Michaelson JJ, Druley TE, Potash JB: **Adaptation of the targeted capture Methyl-Seq platform for the mouse genome identifies novel tissue-specific DNA methylation patterns of genes involved in neurodevelopment.** *Epigenetics : official journal of the DNA Methylation Society* 2015, **10**(7):581-596.
 220. Allum F, Shao X, Guenard F, Simon MM, Busche S, Caron M, Lambourne J, Lessard J, Tandre K, Hedman AK *et al*: **Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants.** *Nature communications* 2015, **6**:7211.
 221. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.** *Bioinformatics* 2011, **27**(11):1571-1572.
 222. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**(4):357-359.
 223. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Mol Cell* 2010, **38**(4):576-589.
 224. Price ME, Cotton AM, Lam LL, Farre P, Emberly E, Brown CJ, Robinson WP, Kobor MS: **Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array.** *Epigenetics Chromatin* 2013, **6**(1):4.
 225. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R: **Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray.** *Epigenetics : official journal of the DNA Methylation Society* 2013, **8**(2):203-209.
 226. Clifford H, Wessely F, Pendurthi S, Emes RD: **Comparison of clustering methods for investigation of genome-wide methylation array data.** *Frontiers in genetics* 2011, **2**:88.
 227. Akaike: **Information theory and an extension of the maximum likelihood principle.** *Proceedings of the Second International Symposium on Information Theory, (ed BN Petrov, F Caski)* 1973:267-281.
 228. Schwarz G: **Estimating the Dimension of a Model.** 1978:461-464.
 229. Aho K, Derryberry D, Peterson T: **Model selection for ecologists: the worldviews of AIC and BIC.** *Ecology* 2014, **95**(3):631-636.
 230. Vrieze SI: **Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the**

- Bayesian information criterion (BIC).** *Psychological methods* 2012, **17**(2):228-243.
231. Shabalin AA: **Matrix eQTL: ultra fast eQTL analysis via large matrix operations.** *Bioinformatics* 2012, **28**(10):1353-1358.
 232. Hoyo C, Murtha AP, Schildkraut JM, Jirtle RL, Demark-Wahnefried W, Forman MR, Iversen ES, Kurtzberg J, Overcash F, Huang Z *et al*: **Methylation variation at IGF2 differentially methylated regions and maternal folic acid use before and during pregnancy.** *Epigenetics : official journal of the DNA Methylation Society* 2011, **6**(7).
 233. Tobi EW, Heijmans BT, Kremer D, Putter H, Delemarre-van de Waal HA, Finken MJ, Wit JM, Slagboom PE: **DNA methylation of IGF2, GNASAS, INSIGF and LEP and being born small for gestational age.** *Epigenetics : official journal of the DNA Methylation Society* 2011, **6**(2):171-176.
 234. Relton CL, Groom A, St Pourcain B, Sayers AE, Swan DC, Embleton ND, Pearce MS, Ring SM, Northstone K, Tobias JH *et al*: **DNA methylation patterns in cord blood DNA and body size in childhood.** *PloS one* 2012, **7**(3):e31821.
 235. Mardia KV, J. T. Kent and J. M. Bibby: **Multivariate Analysis.** London: Academic Press; 1979.
 236. Barker DJ, Winter PD, Osmond C, Margetts B, Simmonds SJ: **Weight in infancy and death from ischaemic heart disease.** *Lancet* 1989, **2**(8663):577-580.
 237. Gillman MW, Rifas-Shiman S, Berkey CS, Field AE, Colditz GA: **Maternal gestational diabetes, birth weight, and adolescent obesity.** *Pediatrics* 2003, **111**(3):e221-226.
 238. Hofman PL, Regan F, Jackson WE, Jefferies C, Knight DB, Robinson EM, Cutfield WS: **Premature birth and later insulin resistance.** *N Engl J Med* 2004, **351**(21):2179-2186.
 239. Hillier TA, Pedula KL, Schmidt MM, Mullen JA, Charles MA, Pettitt DJ: **Childhood obesity and metabolic imprinting: the ongoing effects of maternal hyperglycemia.** *Diabetes Care* 2007, **30**(9):2287-2292.
 240. Bouhours-Nouet N, Dufresne S, de Casson FB, Mathieu E, Douay O, Gatelais F, Rouleau S, Coutant R: **High birth weight and early postnatal weight gain protect obese children and adolescents from truncal adiposity and insulin resistance: metabolically healthy but obese subjects?** *Diabetes Care* 2008, **31**(5):1031-1036.
 241. Painter RC, Osmond C, Gluckman P, Hanson M, Phillips DI, Roseboom TJ: **Transgenerational effects of prenatal exposure to the Dutch famine on neonatal adiposity and health in later life.** *BJOG* 2008, **115**(10):1243-1249.

242. Broekman BF, Chan YH, Chong YS, Quek SC, Fung D, Low YL, Ooi YP, Gluckman PD, Meaney MJ, Wong TY *et al*: **The influence of birth size on intelligence in healthy children.** *Pediatrics* 2009, **123**(6):e1011-1016.
243. Alisi A, Cianfarani S, Manco M, Agostoni C, Nobili V: **Non-alcoholic fatty liver disease and metabolic syndrome in adolescents: Pathogenetic role of genetic background and intrauterine environment.** *Ann Med* 2011.
244. Sohi G, Revesz A, Hardy DB: **Permanent implications of intrauterine growth restriction on cholesterol homeostasis.** *Semin Reprod Med* 2011, **29**(3):246-256.
245. Dancause KN, Laplante DP, Fraser S, Brunet A, Ciampi A, Schmitz N, King S: **Prenatal exposure to a natural disaster increases risk for obesity in 5(1/2)-year-old children.** *Pediatr Res* 2012, **71**(1):126-131.
246. Schwarze CE, Mobascher A, Pallasch B, Hoppe G, Kurz M, Hellhammer DH, Lieb K: **Prenatal adversity: a risk factor in borderline personality disorder?** *Psychol Med* 2012:1-13.
247. Alisi A, Cianfarani S, Manco M, Agostoni C, Nobili V: **Non-alcoholic fatty liver disease and metabolic syndrome in adolescents: pathogenetic role of genetic background and intrauterine environment.** *Ann Med* 2012, **44**(1):29-40.
248. Alwan NA, Greenwood DC, Simpson NA, McArdle HJ, Godfrey KM, Cade JE: **Dietary iron intake during early pregnancy and birth outcomes in a cohort of British women.** *Hum Reprod* 2011, **26**(4):911-919.
249. Belsky J, Jonassaint C, Pluess M, Stanton M, Brummett B, Williams R: **Vulnerability genes or plasticity genes?** *Molecular psychiatry* 2009, **14**(8):746-754.
250. Simons RL, Lei MK, Beach SR, Brody GH, Philibert RA, Gibbons FX: **Social Environmental Variation, Plasticity Genes, and Aggression: Evidence for the Differential Susceptibility Hypothesis.** *American sociological review* 2011, **76**(6):833-912.
251. Pujadas E, Feinberg AP: **Regulated noise in the epigenetic landscape of development and disease.** *Cell* 2012, **148**(6):1123-1131.
252. Patel CJ, Chen R, Kodama K, Ioannidis JP, Butte AJ: **Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus.** *Hum Genet* 2013, **132**(5):495-508.
253. Dal GM, Erguner B, Sagiroglu MS, Yuksel B, Onat OE, Alkan C, Ozcelik T: **Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair.** *Journal of medical genetics* 2014, **51**(7):455-459.
254. Li R, Montpetit A, Rousseau M, Wu SY, Greenwood CM, Spector TD, Pollak M, Polychronakos C, Richards JB: **Somatic point mutations occurring early**

in development: a monozygotic twin study. *Journal of medical genetics* 2014, **51**(1):28-34.

255. Zhang R, Thiele H, Bartmann P, Hilger AC, Berg C, Herberg U, Klingmuller D, Nurnberg P, Ludwig M, Reutter H: **Whole-Exome Sequencing in Nine Monozygotic Discordant Twins.** *Twin research and human genetics : the official journal of the International Society for Twin Studies* 2016, **19**(1):60-65.
256. Abdellaoui A, Ehli EA, Hottenga JJ, Weber Z, Mbarek H, Willemsen G, van Beijsterveldt T, Brooks A, Hudziak JJ, Sullivan PF *et al*: **CNV Concordance in 1,097 MZ Twin Pairs.** *Twin research and human genetics : the official journal of the International Society for Twin Studies* 2015, **18**(1):1-12.
257. Morgan HD, Santos F, Green K, Dean W, Reik W: **Epigenetic reprogramming in mammals.** *Hum Mol Genet* 2005, **14** Spec No 1:R47-58.
258. Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schubeler D: **Identification of genetic elements that autonomously determine DNA methylation states.** *Nat Genet* 2011, **43**(11):1091-1097.
259. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suner D, Cigudosa JC, Urioste M, Benitez J *et al*: **Epigenetic differences arise during the lifetime of monozygotic twins.** *Proc Natl Acad Sci U S A* 2005, **102**(30):10604-10609.
260. Gordon L, Joo JE, Powell JE, Ollikainen M, Novakovic B, Li X, Andronikos R, Cruickshank MN, Conneely KN, Smith AK *et al*: **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 research* 2012.
261. Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C *et al*: **DNA methylation profiles in monozygotic and dizygotic twins.** *Nat Genet* 2009, **41**(2):240-245.
262. van Dongen J, Nivard MG, Willemsen G, Hottenga JJ, Helmer Q, Dolan CV, Ehli EA, Davies GE, van IJterson M, Breeze CE *et al*: **Genetic and environmental influences interact with age and sex in shaping the human methylome.** *Nature communications* 2016, **7**:11115.
263. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A *et al*: **Passive and active DNA methylation and the interplay with genetic variation in gene regulation.** *Elife* 2013, **2**:e00523.
264. Jaffe AE, Irizarry RA: **Accounting for cellular heterogeneity is critical in epigenome-wide association studies.** *Genome Biol* 2014, **15**(2):R31.
265. Holbrook JD: **Does diet interact with genotype to cause epigenetic modification of angiogenesis genes, on the trajectory to obesity?** *Epigenomics* 2016.

266. Richmond RC, Sharp GC, Ward ME, Fraser A, Lyttleton O, McArdle WL, Ring SM, Gaunt TR, Lawlor DA, Davey Smith G *et al*: **DNA Methylation and BMI: Investigating Identified Methylation Sites at HIF3A in a Causal Framework.** *Diabetes* 2016, **65**(5):1231-1244.
267. Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, McKenney S, Witter F, Walston J, Feinberg AP, Fallin MD: **DNA methylation of cord blood cell types: Applications for mixed cell birth studies.** *Epigenetics : official journal of the DNA Methylation Society* 2016:0.
268. Kim K, Zhao R, Doi A, Ng K, Unternaehrer J, Cahan P, Huo H, Loh YH, Aryee MJ, Lensch MW *et al*: **Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells.** *Nat Biotechnol* 2011, **29**(12):1117-1119.
269. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments.** *Bioinformatics* 2012, **28**(6):882-883.
270. Houseman EA, Kelsey KT, Wiencke JK, Marsit CJ: **Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective.** *BMC Bioinformatics* 2015, **16**:95.
271. Kahan BC, Rushton H, Morris TP, Daniel RM: **A comparison of methods to adjust for continuous covariates in the analysis of randomised trials.** *BMC medical research methodology* 2016, **16**(1):42.
272. Li Y, Willer C, Sanna S, Abecasis G: **Genotype imputation.** *Annual review of genomics and human genetics* 2009, **10**:387-406.
273. de Souto MC, Jaskowiak PA, Costa IG: **Impact of missing data imputation methods on gene expression clustering and classification.** *BMC Bioinformatics* 2015, **16**:64.
274. Zhang G, Huang KC, Xu Z, Tzeng JY, Conneely KN, Guan W, Kang J, Li Y: **Across-Platform Imputation of DNA Methylation Levels Incorporating Nonlocal Information Using Penalized Functional Regression.** *Genetic epidemiology* 2016, **40**(4):333-340.
275. Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D: **Penalized Functional Regression.** *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 2011, **20**(4):830-851.
276. Ma B, Wilker EH, Willis-Owen SA, Byun HM, Wong KC, Motta V, Baccarelli AA, Schwartz J, Cookson WO, Khabbaz K *et al*: **Predicting DNA methylation level across human tissues.** *Nucleic acids research* 2014, **42**(6):3515-3528.
277. Yang Y, Xu Z, Song D: **Missing value imputation for microRNA expression data by using a GO-based similarity measure.** *BMC Bioinformatics* 2016, **17** Suppl 1:10.

- 278. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW: **Translating RNA sequencing into clinical diagnostics: opportunities and challenges.** *Nat Rev Genet* 2016.
- 279. Davey Smith G, Hemani G: **Mendelian randomization: genetic anchors for causal inference in epidemiological studies.** *Hum Mol Genet* 2014, **23**(R1):R89-98.
- 280. Relton CL, Gaunt T, McArdle W, Ho K, Duggirala A, Shihab H, Woodward G, Lyttleton O, Evans DM, Reik W *et al*: **Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES).** *Int J Epidemiol* 2015, **44**(4):1181-1190.