Epigenetic variation associated with genetic and environmental factors in the aetiology of Type 2 diabetes

A thesis submitted to Queen Mary University of London for the degree of PhD

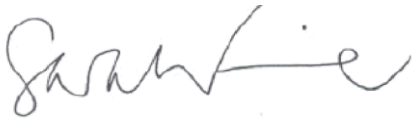
Dr Sarah Finer

Barts and The London School of Medicine and Dentistry,

Queen Mary University of London

Declaration

I declare that the contents of this thesis are my own work. Where other individuals have contributed to the work presented, their role and level of involvement has been stated.



Sarah Finer

1st February 2013

Acknowledgements

This work presented in this thesis has been supervised by Professor Graham Hitman and Dr Vardhman Rakyan, and their support has been invaluable and greatly appreciated. Graham Hitman has nurtured my transition from clinical to research training, allowed me to develop my own interests and ambitions, and has become an important role model; for this I am extremely grateful. Within the Centre for Diabetes, Blizard Institute, Chris Mathews has provided expert bioinformatic input into much of the work presented in this thesis, with additional contributions from Robert Lowe and Guillermo Carbajosa. Melissa Smart, Michelle Holland and Carolina Gemma have helped and supported me in learning laboratory techniques and assisting with wet lab work. Irene Smith and Susanne Bell have provided administrative support. The North-East London Diabetes Research Network, notably Alison Fiddler and Gill Hood, have supported the recruitment to the study presented in Chapter 7, and it was a privilege to recruit and work with the participants themselves who offered their spare time to help with this study. The Blizard Institute has been an exciting and friendly research environment to work in and Mike Curtis, its director, has been very supportive of my work. I have worked with some fantastic external collaborators who have enabled this work to expand its original horizons, and they include:

Type 2 Diabetes project:

University of Cambridge – Thomas Down

UCL – Stephan Beck, Christopher Bell

Oxford – Mark McCarthy, Cecilia Lindgren

Kings College London – Jonathan Mill, Ruth Pidsley

Pune Maternal Nutrition Study

King Edward Memorial Hospital – Chittaranjan Yajnik

Centre for Cellular and Molecular Biology, Hyderabad – Giriraj Chandak, Aparna Duggirala

Matlab Famine Study

ICDDR, Bangladesh – Dewan Alam, Shamim Iqbal

Abstract

Type 2 diabetes, as a complex disease, has a range of genetic and environmental factors that underpin its aetiology. It is hoped that the emerging study of epigenetic processes will provide the necessary mechanistic insight into the genetic and environmental interactions that, to date, are poorly understood. This thesis considers the role of DNA methylation, an epigenetic modification, in the aetiology of type 2 diabetes. A range of different genome-wide and whole genome techniques are applied to a study of established type 2 diabetes and experimental models (human and animal) of fetal programming.

Samples from a recent genome-wide association study of type 2 diabetes were used to identify DNA methylation patterns at areas of genetic variation associated with disease risk. Analysis of data from methylated DNA immunoprecipitation and microarray identified a genetic-epigenetic interaction in the *FTO* gene. At this locus, the presence or absence of a SNP created or abrogated a CpG site capable of methylation and further analyses highlighted possible functional relevance via enhancer activity.

Models of fetal programming were then used to identify whether variation in DNA methylation may underlie the 'programmed' phenotype of diabetes and related cardiometabolic disease. Pre-existing human models of programming via maternal vitamin B12 deficiency and maternal famine exposure have been used to generate exploratory evidence of such mechanisms. Whole genome-based techniques (Medip-seq and Illumina 450k methylation array) were used to profile DNA methylation in whole blood samples from the offspring born to each of these studies. Custom bioinformatic analysis was performed to identify differences in methylation between offspring exposed versus unexposed to the in utero environmental insult. Technical replication and validation studies are ongoing to confirm or refute the presence of regions of differential methylation.

Finally, this thesis considers whether a state of 'over nutrition' gestational diabetes, may play a role in fetal programming. This condition is of increasing prevalence across the world and is characterised by maternal hyperglycaemia and insulin resistance, often resulting in fetal overgrowth. A mouse model using an inbred strain (Lepr) of mice induced a programmed phenotype of glucose intolerance and obesity in aged offspring born to mothers with gestational diabetes. Medip-seq performed on the livers of late gestation mouse embryos identified differential methylation in cases vs. controls, located at genomic regions with potential functional relevance. A human cohort of women with gestational diabetes was collected to develop further hypothesis around the multiple environmental factors that could interact in pregnancy. Prevalent nutritional deficiencies of vitamin D, iron and one-carbon

metabolites were found in women with and without gestational diabetes recruited from a local antenatal clinic.

This thesis presents preliminary findings that variation in DNA methylation may be involved in the genetic and environmental risk of type 2 diabetes. The work presented highlights how future studies must incorporate integrated genetic, epigenetic and functional analysis with sufficient sample size if their results are to be translatable to diverse populations at risk of diabetes.

List of Headings

2.2 Common Experimental methods

2.2.1 DNA extraction and purification

2.2.1.1 Phenol:chloroform extraction

2.2.1.2 DNA extraction and purification kits

2.2.2 Checking DNA concentration and quality

2.2.2.1 Agarose gels

2.2.1.2 Checking DNA concentration

2.2.3 Sonication

2.3 Enrichment-based DNA methylation experiments

2.3.1 Medip

2.3.1.1 qPCR to determine Medip efficiency

2.3.2 Medip-chip

2.3.2.1 Amplification of Medip and input DNA

2.3.2.2 Custom-designed targeted sequencing arrays for Medip-chip

2.3.3 Medip-seq

2.3.3.1 Medip-seq library preparation

2.3.3.2 Multiplexing Medip-seq samples

2.3.3.3 Size selection of finished libraries

2.3.3.4 qPCR to determine Medip library efficiency

2.3.3.5 Illumina GAIIx sequencing

2.4 Methylation array-based experiments

2.4.1 Illumina HumanMethylation 450 BeadChip array

2.4.1.1 Bisulphite conversion

2.4.1.1.1 Testing bisulphite conversion efficiency

2.4.1.2 Array hybridisation

2.4.1.3 Data output

2.5 Data analysis of epigenomic datasets

2.5.1 Important considerations in the analysis of epigenomic datasets

2.5.1.1 Epigenomic profiling – what is normal?

2.5.1.2 Sample size and power calculations

2.5.1.3 Description of methylation values

2.5.1.4 Quality control

2.5.1.5 Normalisation strategies

2.5.1.6 Problems of multiple testing

2.5.1.7 Identification of methylation variation

2.5.1.8 Integration of epigenetic and genetic data

2.5.1.9 Validation and replication

2.5.2 Analysis of Medip-chip

2.5.3 Analysis of Medip-seq

2.5.3.1 Quality control

2.5.3.2 DMR calling strategies

2.5.3.2.1 Batman algorithm

2.5.3.2.2 Thomas Down DMR caller

2.5.3.2.3 USeq

2.5.3.2.4 Combined Thomas Down + USeq DMR callers

2.5.3.3 Gene ontology analysis

2.5.4 Analysis of 450k methylation array

2.5.4.1 Quality control checks

2.5.4.2 Normalisation

2.5.4.3 Data filtering

2.5.4.4 Methylation values

2.5.4.5 MVP calls

2.5.4.6 Sensitivity analysis

2.5.4.7 Identification of SNP-associated methylation variation

List of Tables

Chapter 1. General Introduction

Chapter 2. Materials and Methods

Chapter 3. Type 2 Diabetes Study

Chapter 4. Pune Maternal Nutrition Study

Chapter 5. Matlab Famine Study

5a. Summary of clinical data from participants in the Matlab famine study
5b. Clinical data from Matlab famine study birth cohort
5c. Details of used in Matlab famine study epigenomic experiments.
5d. Sample numbers used for analysis of batch 1
5e. Sample numbers used for analysis of batch 2
5f. MVPs identified in pairwise calls in Batch 1 and Batch 2, after LOOCV, with p value ranges
5g. GO analysis of C vs. A validated MVPs.
5h. Overlap between differentially methylated regions in Gambian and Dutch Winter Hunger Studies with Matlab

Chapter 6. Mouse Model of Gestational Diabetes

6a. Number, sex and genotypes of mice used in mouse model
6b. Maternal and litter characteristics
6c. Results from glucose tolerance tests of female mice at 17.5dpc
6d. Phenotypic characteristics of all male offspring at 6 months' of age
6e. Phenotypic characteristics of all female offspring at 6 months of age
6f. Medip-seq sequencing read counts through before, during and after read processing
6g. DMR counts using different bioinformatic algorithms
6h. Genomic locations and characteristics of DMR tophits
6i. Genomic details of DMR tophits

Chapter 7. Human Model of Gestational Diabetes

7a. Summary of blood tests performed on maternal samples
7b. Maternal metabolic and micronutrient parameters
7c. Prevalence of micronutrient deficiencies in all women
7d. Multiple regression modelling to identify independent variables of (log)homocysteine
7e. Summary of treatment methods and dosing GDM mothers
7f. Data collected at delivery of offspring
7g. Multiple regression modelling to identify independent variables of birth weigh

List of Figures

Chapter 1. General Introduction

Chapter 2. Materials and Methods

Chapter 3. Type 2 Diabetes Study

Chapter 4. Pune Maternal Nutrition Study

4j. Medip-sequence read counts per sample and per flowcell that passed each stage of pre-processing.
4k. Medip enrichment sample matching
4l. Example UCSC upload of Medip-seq data used for sanity checking of DMRs.
4m(i). UCSC screenshots of Medip-seq DMRs at PPARD
4m(ii). UCSC screenshots of Medip-seq DMRs at LUZP2
4n. Infinium probe I bisulphite conversion QC plot
4o. Infinium probe II bisulphite conversion QC plot
4p. Density plot of all array samples (cases = 1, controls =2) showing the density of varying beta values across all probes in a typical bimodal distribution
4q. MDS plot of beta values detected from 430975 probes across all samples
4r. DMRs represented by USeq and multiple overlapping 450k array probes

Chapter 5. Matlab Famine Study

5a. Agarose gels of DNA samples extracted from Matlab study participants
5b. Beta values from Y chromosome probes in a selection of samples from batch 1
5c. MDS plot of 90 Matlab samples from batch 1 (after normalisation).
5d. MDS plot of batch 1 samples (green) and batch 2 samples (orange).
5e. MDS plot of batch 1 samples (green) and batch 2 samples (orange) using normalised data
5f. XY plot examining the differences between batch 1 (X axis) and batch 2 (Y axis) using the sample duplicates run across both batches
5g. Validation of Groups A vs. B MVPs
5h. Validation of Groups A vs. C MVPs
5i. Validation of Groups B vs. C MVPs
5j. Venn diagram showing validated MVPs from pairwise analyses.

Chapter 6. Mouse Model of Gestational Diabetes

6a. Outline of mouse breeding model
6b. Diagram of tetraARMS PCR designed to detect Lepr genotype
6c. Pre-pregnancy body weights of females
6d. Maternal body weights at 17.5dpc
6e. Maternal liver weights at 17.5dpc
6f. Glucose tolerance test results of 17.5dpc mice
6g. Glucose tolerance test results of aged male mice
6h. Glucose tolerance test results of aged female mice
6i. AUC analysis of glucose tolerance tests
6j. Mean body weight of aged mice
6k. Mean gonadal fat pad weights
6l. Mean retroperitoneal fat pad weights
6m. Mean liver weights
6n. Calibration plot of one case-control pair to assess Medip-seq enrichment efficiency

Chapter 7. Human Model of Gestational Diabetes

7a. The one-carbon cycle
7b. Recruitment and sample numbers collected.

7c. Serum glucose measurements in control (non-GDM) and GDM women at their diagnostic glucose tolerance test

7d. Multi-dimensional scaling plot of normalised 450k data from cord blood (green) and placenta (orange) samples.

List of Abbreviations

BATMAN – Bayesian Tool for Methylation Analysis
BMI – Body Mass Index
BS-conversion – Bisulphite conversion
BS-seq – Bisulphite sequencing
CGI – CpG iland
ChIP – Chromatin immunoprecipitation
Chip-seq – Chromatin sequencing
CpG – Cytosine-guanine dinucleotide
DMR – Differentially methylated region
DNA – Deoxyribunucleic Acid
DNMT – DNA methyltransferase
DPC – Days post coitus
EWAS – Epigenome-wide association study
FDR – False discovery rate
FTO – Fat mass and obesity-associated gene
GDM – Gestational diabetes
GWAS – Genome-wide association study
HCNE – Highly conserved non-coding element
Holo-TC – holo-transcobalamin
HOMA-IR – Homeostatic model assessment of insulin resistance
IFG – Impaired fasting glycaemia
IGT – Impaired glucose tolerance
IPGTT – Intraperitoneal glucose tolerance test
LOOCV – Leave-one-out cross-validation
MDS plot – Multi-dimensional scaling plot
Medip – Methylated DNA immunoprecipitation
Medip-seq – Medip-sequencing
MMA – Methylmalonic acid
MVP – Methylation variable position
PCR – Polymerase Chain Reaction
PMNS – Pune Maternal Nutrition Study
OGTT – Oral glucose tolerance test
RNA – Ribonucleic acid
RNA-seq – RNA-sequencing
RRBS-seq – Reduced representation bisulphite sequencing
SNP – Single nucleotide polymorphism
STR – Short tandem repeat
T2D – Type 2 diabetes
WT – Wildtype
5MTHF – 5-methyltetrahydrofolate
27k array – Illumina HumanMethylation 27k BeadChip
450k array – Illumina HumanMethylation 450k BeadChip

Chapter 1. General introduction

The notion that both genetic and environmental factors are important factors in the susceptibility to type 2 diabetes is long known, yet poorly understood. Over recent years, the advent of molecular techniques to study epigenetic modifications, the chemical changes around DNA that affect gene activity, have provided an exciting insight into the processes that may underlie these gene-environment interactions. Epigenetic signatures set down in early developmental life and modified by a range of stochastic and programmed processes throughout the lifecourse are thought to have an important role in an individual's susceptibility to complex disease.

This thesis attempts to use epigenetic discovery to yield insights into the aetiology of type 2 diabetes, and related 'cardiometabolic' conditions such as obesity. The application of epigenetic studies to the understanding of complex disease is a recent molecular approach that is being used to elucidate mechanisms by which genetic and environmental factors interact and in how genetic susceptibility variants may exert a functional effect. The experiments set out in this thesis will use epigenomic techniques to assay and profile epigenome-wide DNA methylation in human and animal models of type 2 diabetes, to understand (i) genetic-epigenetic interactions in established type 2 diabetes and (ii) environment-epigenetic interactions in fetal programming.

1.1 Type 2 diabetes

1.1.1 Epidemiology

The increasing global prevalence of Type 2 diabetes and obesity is well-characterised, and is known to be increasing across the world, with age-standardised diabetes prevalence of 9.8% (men) and 9.2% (women) in 2008, compared to 8.3 and 7.5% in 1980 (1). The rise in diabetes prevalence is particularly apparent in certain global contexts, including the high-income populations of Australasia, North America and Western Europe, but also countries in South Asia under rapid population transition (see Figure 1a). The vascular complications associated with type 2 diabetes are the predominant reason why the disease burden is associated with excess morbidity and mortality. The recent Global Burden of Disease report (2) estimated that the 'years of life lost' due to ischaemic heart disease and stroke have increased by 17-28%

across the world between 1990 and 2010.

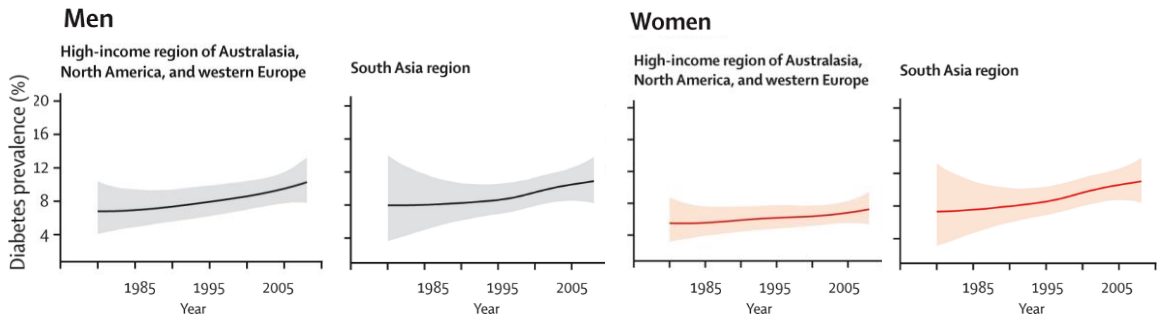


Figure 1a. Global patterns of diabetes prevalence. The figure shows diabetes prevalence (%) over the last 3 decades, with 95% confidence intervals (shaded area). (adapted from (1))

1.1.2 Complex disease aetiology

Complex diseases, such as type 2 diabetes, are thought to have a combination of genetic and environmental risk factors underlying their onset. The genetic basis of these conditions is known to be polygenic, rather than monogenic, and the advent of genome-wide association studies (GWAS) has provided significant insight into understanding these multiple influences though gene discovery. Despite these advances in the detection of the multiple disease-associated genetic variants in type 2 diabetes, little is known about their role in aetiology and how they interact with other factors that play a role in disease susceptibility.

Over 40 genetic variants have now been associated with Type 2 diabetes but, even when taken in combination, account for only a small excess in disease risk on an individual level. Furthermore, the majority of these genetic variants identified relate the pancreatic beta-cell dysfunction in type 2 diabetes pathogenesis, leaving the mechanisms of insulin resistance underrepresented. The small effect size of the genetic polymorphisms associated with type 2 diabetes have led researchers to conceptualise a 'missing heritability' of Type 2 diabetes and to seek alternative mechanisms to explain it, e.g. via the identification of rare or structural variants or interactions between genetic variants (3)(4). Despite this focus on genetic factors in disease causation, there are numerous other factors that may predispose an individual to developing type 2 diabetes. Environmental determinants are well characterised in epidemiological literature, but little is known about their mechanistic roles in disease causation (5) (6). The potential for genetic and environmental influences to interact is significant and indeed could account for the supposed 'missing heritability' concept. 'Heritability' of disease may be affected by non-genetic factors, such as shared environmental susceptibility within families or populations.

For many years, the existence of 'gene-environment interactions' has been supposed through study of genetically-susceptible populations in increasingly 'diabetogenic' environments, however, only recently have plausible 'epigenetic' mechanisms underlying these interactions been described (7). Epigenetic modifications, including DNA methylation, post-translational modification of histones, or other chromatin modifying factors, are highly variable and plastic features of the genome. These modifications may influence phenotype through downstream or bi-directional influences over gene expression (8). The epigenetic 'landscape' of an organism is reflected by the functional diversity of the genome despite its predetermined genetic state. A range of processes, both normal and aberrant, may alter the epigenetic state of an organism, and it is expected that an increased understanding of them may shed light on aetiology of complex phenotype and disease. These epigenetic processes will be discussed further in section 1.3.

1.1.3 Pathogenesis of type 2 diabetes

Type 2 diabetes is a disorder of glucose homeostasis, characterised by insulin resistance in peripheral tissues including liver, muscle and adipocytes; disturbances in insulin secretion from the pancreatic beta cell; and abnormalities in uptake and metabolism of glucose in the gut, via incretin hormones and the splanchnic circulation (9).

Insulin resistance occurs in peripheral tissues such as liver, muscle and adipocytes and is the harbinger of overt type 2 diabetes. Insulin resistance is usually asymptomatic and is manifest by impaired glucose tolerance (IGT) or impaired fasting glycaemia (IFG) during oral glucose tolerance testing or fasting glucose measurement. IGT and IFG are more common in individuals with central adiposity, but usually have no sensitive or specific clinical features associated with them, except sometimes the development of acanthosis nigricans, a skin manifestation of insulin resistance. Peripheral insulin resistance is a complex and poorly understood phenomenon, contributed to by a failure of the suppression of hepatic glucose production, and reduced uptake of glucose by peripheral tissues (mostly muscle). Recent insights into defective insulin intracellular signalling pathways have been gained, identifying defects at phosphorylation of the insulin receptor subunit and downstream phosphorylation cascades including the phosphatidylinositol 3-kinase pathway and its downstream targets including AKT and the insulin-dependent glucose transporter, GLUT4 (10).

Beta cell failure occurs progressively in response to the increasing demands on the pancreas to produce sufficient insulin to overcome peripheral insulin resistance. A range of factors are involved in beta cell failure, including the effects of glucolipotoxicity, where elevated glucose

and free fatty acids are thought to lead to an acquired defect in insulin secretion. In addition, a blunting of the incretin response, whereby gut peptides such as gastric inhibitory polypeptide (GIP) and glucagon-like peptide 1 (GLP-1) increase the insulin response to glucose ingestion, is thought to impair insulin secretion. Specific genetic polymorphisms, identified through GWAS and candidate gene studies (11), and include genes such as TCF7L2, a transcription factor with a role in glucose-stimulated insulin secretion in the pancreatic beta cell (12).

1.1.4 Related cardiometabolic diseases and complications

Vascular complications are the predominant cause of morbidity and mortality in individuals with type 2 diabetes. The seminal work performed by the UKPDS study groups identified the risks of increasing glycaemia on macrovascular (e.g. ischaemic heart disease and stroke) and microvascular (e.g. retinopathy and nephropathy) complications of diabetes (13). Recently, this data has been challenged in long-term outcome studies of individuals with type 2 diabetes suggesting an increased risk of death from macrovascular disease in those who have been intensively treated with regard to glycaemia in the ACCORD trial and others (reviewed in (14)). However, the risks of glycaemia in respect to microvascular outcomes remain, and the concept of an early 'metabolic memory' has emerged suggesting that the reduction in macrovascular complications is restricted to those individuals who have achieved tight glycaemic control in the early stages of disease (15). Hyperglycaemia is not the sole mediator of long-term cardiometabolic complications in individuals with type 2 diabetes; the effects of advanced glycation end products and systemic inflammation contribute to tissue damage, as do the commonly coexisting conditions of dyslipidaemia and hypertension (16).

1.1.5 Treatment

Treatment of type 2 diabetes follows a stepwise approach, starting with lifestyle advice and then incorporating drugs targeted to treat insulin resistance (e.g. metformin) or overcome reduced insulin secretion from the beta cell (sulphonylureas, insulin). Newer therapies (e.g. liraglutide) have been developed to augment the defective incretin response in people with diabetes, thus aiding the insulin response to oral glucose.

A personalised approach is taken to treatment of type 2 diabetes, but factors affecting individual responsiveness to treatment are not known. Pharmacogenomics has identified a genetic variant in the ATM gene that affects the glycaemic response to metformin, but as yet,

there is no direct translational benefit of this discovery. A greater understanding of functionally relevant epigenetic processes in diabetes pathogenesis may enable tailored drug design and personalised therapy to be developed in the future.

1.1.6 Prevention

Prevention strategies use a broad-brush approach to reduce diabetes risk and are based on population-based studies. Whilst such studies have shown the effectiveness of lifestyle-based prevention strategies, e.g. the Finnish Diabetes Prevention Study, which reduced relative risk of progression from impaired glucose tolerance to type 2 diabetes by 43% over 7 years through intensive diet and exercise (17), they are considered to have had minimal impact on wider populations. The 'one size fits all' approach to lifestyle and health promotion advice, based around achieving weight loss and increased physical activity is often considered to be ineffective, particularly in populations where societal and personal barriers to these changes exist.

1.2 Fetal programming

The concept of fetal programming, or the developmental origins of disease (DoHAD), has become increasingly described and studied over the last two decades. Fetal programming refers to the hypothesis that adverse environmental factors in utero can have lasting effects on the physiological development of the fetus that can influence, and increase, its propensity to develop certain diseases in later life, such as type 2 diabetes. This theory is gaining increasing recognition from a multidisciplinary perspective, combining basic scientists with clinicians, epidemiologists and social scientists. Hales and Barker (18), were the first to describe the phenomenon of fetal programming in routinely collected clinical data from Hertfordshire, UK. These researchers used birth records collected by midwives in the 1920s and 30s, and were able to follow up 463 of men for whom these records existed in late adulthood. In doing so, Barker and Hales noted that there was an inverse relationship between birth weight and risk of adult diseases, including type 2 diabetes and cardiovascular disease (18) (19).

The contrasting thrifty phenotype and fetal insulin hypotheses are two concepts that seek to understand fetal programming in greater detail and will now be discussed. In addition, the range of different environmental insults that can give rise to fetal programming, notably gestational diabetes and nutritional insults, will be considered.

1.2.1 Thrifty phenotype hypothesis

Barker and Hales described the 'thrifty phenotype hypothesis' to try and conceptualise their observations (18). They explained that a fetus that develops in an adverse pregnancy environment may be born small and in later life be 'programmed' to function in a similarly restricted environment. These adaptations allow the human to exist through continued 'thrift', perhaps with limited growth potential, but in a state of health. For those individuals whose environment changes into one of nutritional excess, a mismatch can occur between their thrifty design and the challenges that their body is put under. Type 2 diabetes, obesity, and related disorders can ensue from this mismatch and the inability for the human to cope with the extra demands on it due to reduced plasticity that was programmed in its early development.

Proponents of this hypothesis see the potential for this phenomenon to underlie the changing global patterns of diabetes and obesity and put fetal programming at the heart of the putative gene-environment interactions underlying their aetiology. In particular, it has been noted that the 'thrifty phenotype' hypothesis has the potential to explain the increasing prevalence of these diseases in Asia, as many people moved from rural, undernourished environments, to urban areas with calorific excess and reduced physical activity.

1.2.2 Fetal insulin hypothesis

Other researchers have proposed a contrasting theory to explain the fetal origins of type 2 diabetes and obesity: the fetal insulin hypothesis (20). This hypothesis explains that interacting genetic influences between a mother and her developing child can regulate both birth weight, and future risk of diabetes. These important genetic determinants of birth weight are located around insulin control genes, e.g. glucokinase, which have a combined effect on birth weight, through insulin's actions as a growth factor, and their role in future type 2 diabetes risk, e.g. ADCY5 (21). However, these genetic variants contribute only a small amount to birth weight variation, e.g. for the 9% of Europeans estimated to carry 4 birth weight-lowering alleles, there is 113g reduction in birth weight, compared to those carrying zero or one allele. This suggests the potential for multiple additional regulators of birth weight, including both genetic and environmental factors that may also play a role in the long-term risk of cardiometabolic diseases.

1.2.4 Gestational diabetes

1.2.4.1 Epidemiology

Gestational diabetes mellitus (GDM) is a common metabolic disorder of pregnancy, manifest by maternal impaired glucose tolerance in mothers in the late 2nd trimester of pregnancy and onwards. The prevalence of GDM is not well described, and as it is usually asymptomatic in its early stages, diagnosis relies on biochemical screening policies. The National Institute of Clinical Excellence guidance on diabetes in pregnancy in 2008 (www.**nice**.org.uk/**nice**media/pdf/CG063Guidance.pdf) has suggested that the prevalence of GDM in England and Wales is 3.5%, but the American Diabetes Association report an average of 7% prevalence across studies. Estimates from a local audit study (N. Tomkins, unpublished) at the Royal London Hospital suggests that GDM affected 11.9% of the 4612 receiving antenatal care during 2011, highlighting the excessive prevalence of the disorder in at-risk groups, such as women of South Asian origin (predominantly Bangladeshi), which comprise 49% of the pregnant population at the Royal London Hospital. Other risk factors for gestational diabetes have been described by the American Diabetes Association include obesity, previous gestational diabetes, a first degree family history of diabetes, and South Asian, black Caribbean, or Middle-Eastern ethnic origin, and identification of these triggers oral glucose tolerance tests (OGTT) screening at 24-28 weeks gestation (22). One of the reasons that GDM prevalence estimates vary is due to differences in diagnostic criteria for the condition defined by various researchers and advisory committees, including the International Association of Diabetes and Pregnancy Study Groups (23) and the UK-based NICE guidance (described above) (see table 1a). The IADPSG criteria have been based on more recent data from the HAPO Study (to be discussed later) suggesting a linear association between maternal glycaemia and adverse pregnancy outcomes. However, many have questioned whether the criteria for fasting and 1 hour post-challenge glucose were too stringent and lead to over-diagnosis and whether the 2 hour post-glucose challenge cutoff is too lax, especially for women of Asian origin who are known to exhibit greater post-prandial hyperglycaemia (24).

	Oral glucose challenge	Fasting glucose	1 hour glucose post-challenge	2 hour glucose post- challenge
Carpenter and Coustan (25)	100g	5.3	8.7	7.8
WHO criteria (26)	75g	7.0	Not used	7.8
NICE guidelines (2008)	75g	7.0	Not used	7.8
IADPSG (23)	75g	5.1	10.0	8.5
Royal London Hospital (current practice)	75g	5.8	Not used	7.8

Table 1a. Diagnostic criteria for gestational diabetes based on different clinical guidelines. Glucose cutoffs are all greater than or equal to the value presented and units are mmol/l

1.2.4.2 Aetiology

Gestational diabetes is thought to have a similar aetiological basis to that of type 2 diabetes. Risk factors for GDM include ethnicity (women of Asian, Middle-Eastern and African origin are particularly at risk), maternal obesity, increasing age and parity, and a family history of GDM and type 2 diabetes. The genetic basis of gestational diabetes is not well characterised, but it is thought to overlap with that of type 2 diabetes. Only one genome-wide association study has been performed GDM (in Korean women) and identifies CDKAL1 and MTNR1B as genetic variants associated with the condition, at genome-wide significance (27). The CDKAL1 and MTNR1B variants have also been identified as risk variants in type 2 diabetes; it is believed that they play a role in beta cell functioning and control of fasting glucose levels. Other studies have tested for commonality of the known type 2 diabetes GWAS hits with GDM and have found considerable overlap, but this approach does not fully address the question of whether they are associated with GDM or later T2DM risk, or both, in the individuals tested.

1.2.4.3 Pathogenesis

During normal pregnancy, alterations to glucose and insulin metabolism occur to provide extra fuel for developing fetus. During the first trimester, a pregnant woman becomes leptin resistant, enabling the accumulation of fat stores as fuel deposits for later pregnancy and lactation. Increasing maternal insulin resistance and gluconeogenesis during gestation results in increased glucose transit to the fetus. Maternal pancreatic beta cell expansion occurs to cope with increased demands for insulin due to this change in maternal glucose metabolism, but as maternal insulin does not cross the placenta, the maternal hyperinsulinaemia does not prevent the transit of sufficient glucose across to the fetus. Placental glucose transfer is

mediated by specific members of the glucose transporter (GLUT) family, e.g. GLUT1 and GLUT3. Recent evidence suggests that the GLUT genes may be dynamically regulated by epigenetic processes (28). Fetal glucose metabolism is regulated by its own insulin production, in addition to other hormones, e.g. thyroid hormone, IGF1, cortisol, with a role in glucose homeostasis. Transfer of lipids across the placenta also requires complex metabolic pathways to supply the correct lipid substances (predominantly free fatty acids) to the developing fetus for development of white fat stores towards the end of gestation. Amino acid transport across the placenta enables the increasing demands of the fetus for a positive protein and nitrogen balance to support its growth. Amino acid transfer is dependent on the physiological increase in placental perfusion to assist active exchange across trophoblast membranes. The complexities of these normal physiological processes of placento-fetal nutrient transfer are great and, furthermore, their disruption in pathological conditions of pregnancy, such as gestational diabetes, are poorly understood and can have significant consequences on fetal growth.

Gestational diabetes is manifest by glucose intolerance in the late 2nd trimester due to a failure of the required beta-cell expansion to cope with increased requirements for insulin production as well as increased peripheral insulin resistance. The resulting maternal glucose intolerance is manifest as maternal hyperglycaemia and results in increased glucose transit to the developing fetus. Maternal hyperinsulinaemia ensues in response to hyperglycaemia but as maternal insulin does not cross the placental interface, the fetal environment is hyperglycaemic and results in the fetus producing its own insulin in excess to try and maintain normoglycaemia. Fetal hyperinsulinism results in accelerated fetal growth, leading to macrosomia and organomegaly, as well as the risk of neonatal hypoglycaemia when the neonate has to adapt rapidly to its own normoglycaemic environment post-partum after a prolonged state of hyperinsulinism during gestation.

The triggers to these abnormal homeostatic and metabolic processes in GDM are poorly understood. The hormones prolactin and human placental lactogen are thought to play a role in the beta cell expansion required in pregnancy and a recent study has shown that serotonin has an important role downstream of these which, if disrupted, may lead to failure of beta cell expansion and ensuing glucose intolerance (29). Leptin resistance is known to occur in normal pregnancy and is thought to be mediated centrally through hypothalamic pathways. Increased leptin resistance has been found in women with gestational diabetes and may play a role in reducing insulin sensitivity in pregnant women and inhibiting insulin secretion (reviewed in (30)). Yamashita et al (31) have used these insights to understand the role of leptin in gestational diabetes and adiposity using a leptin receptor mutant inbred mouse strain (see

Chapter 6) and suggest that leptin administration in pregnancy may improve maternal insulin resistance via reducing adiposity. Finally, specific cytokine members of inflammatory pathways, e.g. adiponectin and TNFa, may also play a role in the metabolic phenotypes of pregnancy and tendency towards gestational diabetes (32).

1.2.4.4 Clinical features

There are few specific clinical features of gestational diabetes in mothers, other than those associated with risk of developing the condition, including obesity, ethnicity and increasing age. The maternal risks associated with gestational diabetes include pregnancy-associated complications e.g. preterm birth, increased rate of Caesarean section and instrumental delivery, and are predominantly related to fetal macrosomia (33). In the longer-term, women with gestational diabetes are known to have an accelerated progression to type 2 diabetes (T2D), compared to those with and equivalent degree of impaired glucose tolerance unrelated to pregnancy. This observation, and the understanding that the fasting glucose level post-GDM pregnancy is predictive of T2D onset, suggests either a common genetic susceptibility to both GDM and T2D mediated by beta cell genes, or that having a GDM pregnancy in some way 'exhausts' beta cell function and therefore accelerates the progression to type 2 diabetes. Long-term follow up of the ALSPAC cohort has also shown an increase in 10 year cardiovascular disease risk in women who have had a pregnancy complicated by gestational diabetes (34). Further studies to address the exact mechanisms behind these observations are important and should yield insights into the aetiology and pathogenesis of GDM and T2DM.

The short-term risks of gestational diabetes to the developing fetus relate mostly to the complications of excessive fetal growth to specific organs, e.g. cardiac malformations, as well as overall macrosomia, which is known to leading to delivery-related complications such as shoulder dystocia. After delivery, fetal hypoglycaemia from beta cell hypertrophy in response to maternal glucose levels can occur and is associated with neonatal morbidity and mortality.

1.2.4.5 Role of GDM in fetal programming

The long-term risks to offspring exposed to gestational diabetes are increasingly recognised and considered to be a mode of fetal programming of adult disease. Multiple cross-sectional studies have identified an association between maternal gestational diabetes and a programmed phenotype of diabetes and obesity in offspring. Studies in mothers with gestational diabetes in the Pima Indian population of the USA provided early insights into this

model, showing that offspring born to these women had fasting hyperglycaemia, higher rates of abnormal glucose tolerance and obesity, compared to controls (35). Early criticisms of this data suggested that this observation could be explained by the inheritance from mother to child of genetic risk towards all of these phenotypes. A rat model performed by Gauguier (36) in which continuous intravenous infusion of glucose to dams in the third trimester of pregnancy rendered them hyperglycaemic and induced programmed glucose intolerance and impaired insulin secretion in F1 offspring. When female offspring were mated with unexposed rats, they exhibited further glucose intolerance in pregnancy and the programmed effects were transmitted to the F2 generation who were macrosomic and displayed abnormal glucose and insulin metabolism from birth and into adulthood. Since this seminal study, further animal evidence of the programming effects of maternal hyperglycaemia have been published (37) and more detailed human studies have been performed. However, a series of studies refines this hypothesis, identifying that exposure to varied maternal hyperglycaemia in utero does indeed 'programme' offspring to developing type 2 diabetes and obesity, independent of genetic influences. Another study of the Pima Indian population used siblings born to mothers in subsequent pregnancies when one had been exposed to diabetes and the other had not (38). The offspring exposed to maternal diabetes showed a 3.7 times greater risk of having diabetes and at follow-up than their unexposed (age-matched) sibling, with similar differences in obesity, thereby identifying a risk in excess of that explainable by genetic influence. Additional evidence for the non-genetic basis of this programming model comes from Lindsay et al (39) who have shown an increased propensity to overweight and obesity in offspring born to mothers with type 1 diabetes, and they hypothesise that this may be regulated by higher cord leptin.

The mechanisms underlying these programming events are not clear, but it seems likely that they have the potential to occur late in gestation, during the period of maximal fetal growth. Further studies are required to understand the precise mechanisms by which this programming can occur and whether it is induced through one environmental factor or many. Additional understanding of the contributory genetic factors (from mother, father and child) in modulating this programming is important, as highlighted by the observation that paternal insulin resistance can affect fetal growth in the Parthenon studies in Mysore, India (40).

1.2.5 Nutritional deficiency in pregnancy

1.2.5.1 Micronutrient deficiency – vitamin B12 and folate

Over recent years, human and animal studies have shown the potential for maternal deficiency of one-carbon cycle nutrients to programme a phenotype of diabetes and obesity in offspring, and have suggested that this programming occurs via epigenetic variation. Vitamin B12 and folate are two key components of the one carbon metabolic cycle: folate (from dietary tetrahydrofolate) exists as a substrate and B12 as a cofactor to methionine synthase for the efficient cycling of methionine to homocysteine (see figure 7a). Crucially, when this one-carbon cycle is working effectively, it produces methyl groups from the conversion of the substrate, S-adenosyl-methionine (SAM), to its product, S-adenosyl-homocysteine (SAH), using methyltransferase enzymes. Researchers have described the ratio of SAM:SAH as marker of 'cellular methylation capacity' and several studies have suggested that it has a direct impact on DNA methylation (41). The studies described use a range of different tissue types, including lymphocytes and solid organ tissue as well as cell lines to highlight how methyl donor deficiency results in reduced SAM:SAH and consequent disruption in DNA methylation. However, the techniques that have been used are now out-dated and rely on quantitative global estimates of DNA methylation (e.g. (42)) that do not provide genomic detail.

The role of B12 and folate in human fetal programming has been described elegantly in the Pune Maternal Nutrition Study (43). In this study, Yajnik and colleagues have performed a prospective study of young women and their children from pre-conception, through pregnancy and beyond. Thus far, the study has followed these children until their early teens with published studies on the childhood outcomes of different in utero influences. The most notable finding from this study is that women who were B12 deficient in early pregnancy (a common problem in India due to the predominant lacto-vegetarian diet) have children who are more insulin resistant and obese than children born to mothers who were B12 replete in pregnancy. These differences, measured using detailed anthropometry measurements and biochemical indices of glucose and insulin homeostasis, were observed in several hundred 6 year-old children born to the study. The possible role of altered DNA methylation in this example of fetal programming will be explored in Chapter 4.

A model of periconceptual methyl group deprivation in ewes has also been described and provides further support to the role of one-carbon metabolites in a programmed 'cardiometabolic' offspring phenotype (44). In this model, ewes were deprived of a range of methyl donors in their diet, including vitamin B12 and folate as well as methionine, pre-conceptually and at the time of conception The ewes receiving the deprived diet showed

biochemical deficiency of B12, folate and methionine with an expected rise in homocysteine. Blastocysts from these donor ewes were then transferred to normally fed surrogate ewes for the remainder of pregnancy, as well as weaning. Granulosa cell lysates were collected at the time of embryo transfer and these showed a reduced SAM:SAH ratio. The phenotype of offspring exposed to periconceptual maternal methyl deficiency was consistent with programming towards cardiometabolic disease. All methyl-deficient exposed offspring were heavier at birth and exhibited higher growth rates into their adulthood, and male offspring exhibited higher fat mass (judged by CT scan) at 12 months of age compared to controls. Male offspring also showed insulin resistance, independent of differences in adiposity, with similar, but smaller, differences seen in female offspring. By 23 months of age, male offspring showed elevated body fat-adjusted blood pressure in the exposed group vs. controls. These convincing phenotypic changes were associated with reported variation in DNA methylation. The researchers used methylation-sensitive restriction enzymes to characterise DNA methylation at 1400 CpG sites, mostly at promoter sequences of genes in the fetal livers of 18 control and 16 methyl deficient animals. In their analysis, they found that 57 loci (4% of the CpG sites studied) exhibited altered methylation in two or more animals in methyl-deficient offspring and suggest that this is more than would be expected by chance. Furthermore, the authors report that hypomethylation was more common than hypermethylation, as could be expected from methyl deficiency, and that the differences were more common in males than females, consistent with the phenotypic outcome. Although this method of studying DNA methylation provides only a limited perspective of the potential complex epigenomic phenomena, and has now been superseded by genome-wide technologies, this paper has given an early, and much-needed support of the potential mechanisms underlying this fetal programming hypothesis.

Other researchers have shown small methylation differences in target genes in association with maternal and infant one-carbon status and folic acid supplementation in healthy populations in the UK and USA (45) (46). These studies used bisulphite-pyrosequencing assays and detected small methylation differences (of up to 8%) and weak correlations in association with parameters of one-carbon status. More importantly, McKay and colleagues have provided evidence in their paper that epigenetic-genetic interactions may be important in determining methylation status. In their study, analysis of common genetic variants that regulate folate metabolism, such as the methyltetrahydrofolate (MTHFR) gene, identified an association between maternal and infant genotype with infant methylation. Although only preliminary findings, these highlight the complexity of interacting influences of gene and environment that regulate methylation patterns across the genome.

1.2.5.2 Exposure to famine

Several large studies based on periods of severe and widespread famine have refined the evidence that a restricted in utero environment can predispose to diabetes in later life. The Dutch Winter Hunger was a period of extreme famine in the 1940s in Holland. A combination of the after-effects of the Second World War and harsh weather conditions leading to crop failure meant that a large proportion of the Dutch population were exposed to extreme nutritional deprivation for a 5-month period. Epidemiological studies (47) have compared those offspring exposed to famine around the time of conception (periconceptual) or in fetal development, with their younger siblings who were not exposed to famine. Those offspring who were exposed to famine in utero were more likely to develop type 2 diabetes, obesity and other disorders such as schizophrenia, in later life, compared to their unexposed siblings. The individuals studied included those born before famine (n = 202), those exposed in early gestation (n = 63), mid-gestation (n = 100), late gestation (n = 116), and born after (n = 221). At follow-up, the mean age of participants was 50 years, mean BMI was 27kg/m^2 and there was a background risk of type 2 diabetes of 16% (in the conceived after famine group). This study did not identify significant differences in those with and without T2DM between groups, but 2 hours post-challenge glucose and insulin were higher in those exposed to famine, with an increase in 2 hour plasma glucose of 2.4% per kg/m^2 of BMI (95% CI 1.9-2.9). When adjusted for BMI and sex, was greatest in those exposed in late gestation, followed by mid-gestation exposure, with early exposure conferring the smallest excess risk of post-challenge hyperglycaemia. The glucose intolerance of offspring was inversely associated with maternal weight and birth weight, and those offspring who had the lowest birth weight but the highest adult BMI had the highest 120 minute glucose values. The authors have tried to elucidate whether the programmed offspring display insulin resistance and/or a beta cell secretory defect by measuring pro-insulin, split pro-insulin and insulin levels in the dynamic glucose studies performed. The fasting pro-insulin levels of offspring exposed to famine were elevated, however there was no other evidence of a beta-cell secretory defect from these fasting measures nor their incremental response to glucose, therefore the results are more consistent with an insulin resistant pathogenesis.

Subsequent studies on a smaller group of offspring born to the Dutch Winter Hunger model used more complex dynamic studies to assess the mechanism of glucose intolerance, calculating a disposition index from an intravenous glucose tolerance test as a measure of beta-cell secretion, adjusted for insulin resistance. In this study, the researchers studied 97 individuals with normal glucose tolerance, excluding anyone with impaired glucose tolerance or type 2 diabetes. Similar to the original study, there was no difference in the fasting glucose

and insulin parameters of famine-exposed and unexposed, however the glucose intolerance of the former group was reproduced in this study. The disposition index was lower in the group exposed to famine exposure in mid-gestation to a statistically significant level, however this was not seen in the groups exposed early or late in gestation. The authors of this study conclude that maternal famine exposure in mid-gestation induces glucose intolerance via insulin secretory defect (48). This is consistent with the theory of fuel-mediated teratogenesis (49), however the exclusion of individuals with established dysglycaemia is a significant omission in understanding the mechanisms across this group as a whole. The finding of beta-cell dysfunction is also in conflict with animal studies that suggest insulin resistance is the more common result of prenatal famine exposure (50).

Similar data comes from studies of offspring born during the severe Chinese famine in 1958-1961 (51). This famine occurred at the time of the 'Great Leap Forward', when widespread political and economic change across China is said to have caused a severe famine and millions of deaths. As with the Dutch Winter Hunger Study, population-based follow up studies have been performed on the offspring born before and after this famine, with n=1005 exposed and 1974 unexposed offspring studied at age 38-42 years. In this study, the severely famine-exposed (n=503) had a four-fold increased risk of hyperglycaemia (>6.1mmol/l fasting or >7.8mmol/l 2 hours post glucose load) at age compared to those in rural areas unaffected by famine (OR=3.92, 95% CI 1.64 – 9.39, p=0.002). Importantly, for those exposed to severe famine who followed a Western/affluent diet in later life, the increased risk of diabetes was even greater (OR = 7.63, 95% CI 2.41-24.1, p=0.0005), with a similar increase in risk seen in individuals who had a higher socioeconomic status in later life. Interestingly, mean BMI across both exposed and unexposed offspring was no different and within the normal weight range at 23kg/m^2 in both groups. This study has not examined trimester-specific effects on later diabetes risk.

A study (unpublished) in Bangladesh looking at individuals exposed to famine during the 1970s shows a higher rate of impaired glucose tolerance in young adults exposed to famine in utero, compared to those born either before, or after. This study was performed in a rural population at high genetic risk of type 2 diabetes, but where under-nutrition and low birth weight are still common and will be discussed in Chapter 4.

In contrast to these two studies, the Siege of Leningrad study, which followed up individuals following their exposure to famine, found no real differences in their risk of diabetes and obesity in later life compared to a matched group of individuals that were not exposed to the famine (52). This may be due to the continued poverty of this group of individuals that made the 'famine' expose less defined and more chronic, lasting well into postnatal and later life.

Criticisms of these studies have been levelled at the epidemiological assumption that an association between famine exposure and diabetes implies causation. A range of other factors may influence the increased risk of diabetes in these studies, including differences in the external environment and behaviours of these offspring from their postnatal life into adulthood. In such population-based cohorts and with retrospective methods of data collection, it is difficult to define these postnatal influences and to characterise the exact exposure in terms of nutritional deficits, accompanying diseases, social deprivation or stress.

Famine may involve exposure to multiple micronutrient and macronutrient deficiencies and its precise nutritional manifestation has not been characterised in detail in the studies above. Data from animal models have supported the role of specific micronutrient deficiencies (44), protein deficiency and vascular compromise in programming (reviewed in (53)).

The molecular basis of these models of fetal programming is poorly understood. Recent advances in the field of epigenetics suggests that the environmental susceptibility of the developing fetal epigenome, and predisposing genetic factors, are an important means of elucidating these molecular mechanisms and determining the aetiological and pathogenic influences on complex disease phenotype. This putative epigenetic process will be discussed in the next section.

1.3 Epigenetic variation

In recent years, numerous collaborative GWAS studies using large cohorts have yielded some insight into the genetic control of phenotypic diversity (3) (54). Similar studies of the epiallelic contribution to complex traits have so far been limited by technological and economic factors as well as the analytical complexity required. However, insight has been gained from the characterisation of epiallelically-induced phenotypic variation in monozygotic twins (55) and the identification of pleiotropic quantitative loci which account for phenotypic variance in mice (56). Large-scale 'epigenomic' initiatives, such as the NIH International Human Epigenomic Consortium, are being created to study the range of epigenetic modifications in normal and diseased states, and a variety of tissues. Open access data from such studies is increasingly displayed on standard genome browsers will foster further developments in this area. DNA methylation, as a stable epigenetic modification, lends itself to detailed study through a range of experimental platforms. Epigenomic patterns of methylation are now well characterised, and show a relationship between CpG density and genomic location, described Figure 1b.

Much focus has been placed on CpG islands (CGIs), regions where the observed density of CpG dinucleotides is higher than that expected from the background genome (defined by the Ensembl gene browser as >400 bp long and with a $CpG_{o/e}$ density of >0.6). Typically, CGIs overlie gene promoters and are relatively under-methylated compared to regions of lower CpG density. A detailed human epigenomic profile comes from Lister (57) who used a technique that couples bisulphite conversion with Next Generation Sequencing (called BS-seq) to study DNA methylation in a human cell line and embryonic stem cells. Using this approach, 1.16-1.18 billion reads were generated, providing coverage of over 86% of both strands of the 3.08Gb human reference sequence. This extensive profiling confirmed previous epigenomic observations, but also highlights the existence of non-CpG site methylation in embryonic stem cells. The role of non-CpG methylation is unclear, but it seems likely to play a role in cellular differentiation.

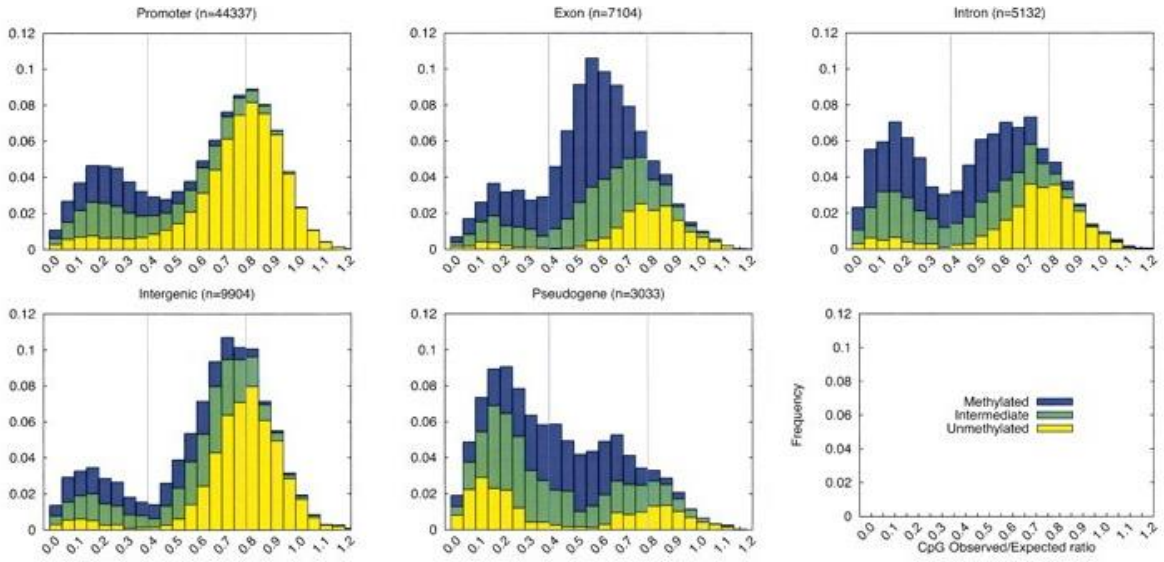


Figure 1b. Analysis of somatic DNA methylation profiles, using MeDIP-chip. Data were categorised into unmethylated (yellow: <40%), intermediate (green: 40-60%) and methylated (blue: >60%) regions, and included non-promoter CpG islands. X axis = CpG observed/expected ratio and Y axis = frequency Reproduced from Rakyan et al (58).

Techniques for studying protein-DNA binding, such as ChIP-seq, have been used to profile the range of histone modifications that contribute to the variable epigenetic landscape in humans, and their interaction with DNA methylation. High-resolution maps for up to 20 chromatin modifications have now been produced (59) (60), and these have also provided an insight into the possible role of epigenetic mechanisms in transcriptional regulation. A range of 'active' and 'repressive' histone modifications have now been well characterised, and these seem to

have predictable relationships with DNA methylation patterns, nucleosome positioning and gene expression (59) (61) (62) (Figure 1c).

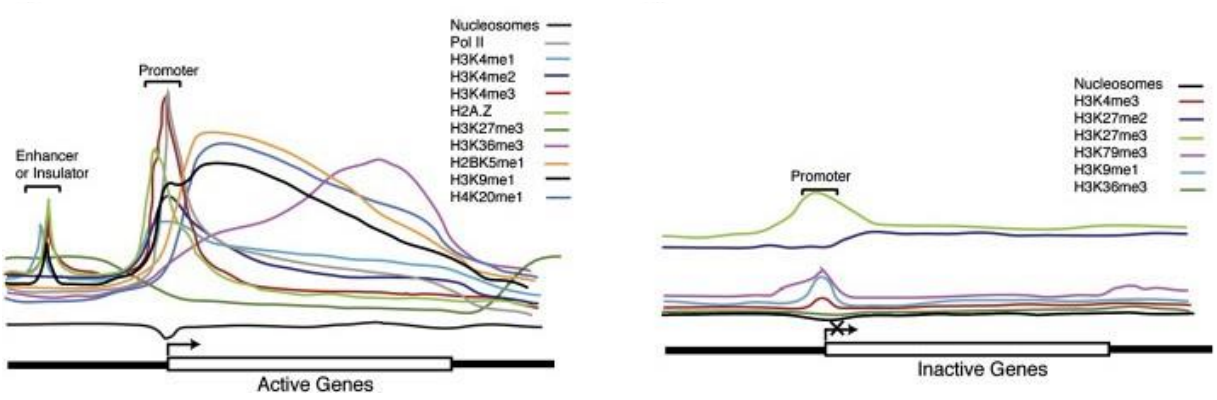


Figure 1c. Characteristic histone methylation profiles of active and silent genes. Y axis represents enrichment and X axis the genomic location. From Barski et al (59).

The characterisation of epigenetic modifications on their own gives a limited insight into their interactions with other factors, e.g. genetic variants, their acquisition and maintenance, and their functional role. Like genetic variants, or alleles, epigenetic variants are increasingly seen as stable and penetrant entities and have thus been defined as 'epialleles' by some (63). In defining an epiallele, they can be conceptualised first as regions where genetic, environmental and stochastic events converge to form clear 'signatures', and second, as having a defined outcome such as variable gene expression or phenotype. Richards has defined these epialleles as obligatory, facilitated and pure (summarised in Figure 1d), a classification that is helpful in understanding how to detect and understand their putative role in complex disease aetiology, especially with respect to their interaction with genetic and environmental factors.

1.3.1 Epialleles

1.3.1.1 Obligatory epialleles.

These are generated as a direct consequence of genetic polymorphism and are independent of environmental or stochastic influences (Figure 1d). Epialleles of this type may be associated with a specific haplotype, in which case they are said to occur in *cis*. Single-locus studies have identified obligatory epialleles for DNA methylation, where a single nucleotide polymorphism (SNP) is responsible for the creation, or deletion of a CpG site (64) (65). Genome-wide

approaches have identified further examples of SNP associated, allele-specific methylation that is more widespread (66) (67). Another recent study has mapped differentially methylated regions (DMRs) in the F1 generation from crosses of unique, inbred mouse strains, and reports that the majority of differentially methylated regions (DMRs) coincided with the underlying genotype (68). Similarly, it has been reported that local histone modifications may also segregate in a Mendelian manner, although the statistical power of this finding was weak (69). Allele-specific chromatin marks have recently been identified through the study of CTCF-binding sites across the genome; these commonly occurring trans-acting transcription and chromatin regulatory factors act as insulators, blocking enhancer activity. McDaniell and colleagues (70) have identified allele-specificity of CTCF-binding sites on a genomic level in humans, and a potential complex regulatory role in X inactivation.

When an epiallele results from a polymorphism at a distant genomic location, it is termed *trans*-obligatory (Figure 1d). Epialleles of this nature are less easy to identify, however, the mutagenesis screen utilised by Whitelaw and colleagues (71) (72) identified several chromatin modifiers using this principle. Mutations identified in their screen included core regulators of epigenetic phenomena, e.g. DNA methyl-transferase 1 (DNMT1), and unsurprisingly, were associated with severe phenotypes, many being homozygous embryonically lethal (72).

Although obligatory epialleles influence phenotype through introducing epigenetic variation, they exist purely as a direct result of genetic variation. Whether alleles of this nature should be considered as epialleles is subject to ongoing debate.

1.3.1.2 Facilitated epialleles

In other cases, genetic polymorphism may predispose to the formation of an epiallele without determining the epigenotype. Epialleles of this nature are termed 'facilitated', and their occurrence has been further defined as a severance in the direct link between genetic and epigenetic variation (63). Facilitated epialleles, although mitotically stable, do not conform to Mendelian segregation. Some of the best examples of facilitated epialleles include the *agouti viable yellow* (A^{vy}) (73) and *axin-fused* (74) alleles, both of which are associated with the insertion of an intracisternal-A particle (IAP) retrotransposon into a non-translated region of the endogenous gene. Promoters associated with these retrotransposons are capable of directing production of an aberrant transcript and are epigenetically regulated, potentially existing in a transcriptionally silent or active state. By contrast, in the absence of the IAP insertion, the endogenous gene does not exist as an epiallele, and thus the genetic event is

seen to be a necessary substrate for its creation. These epialleles have been studied in isogenic mice, eliminating confounding effects of *trans*-modifiers and revealing that the epigenetic state of the introduced IAP promoter is stochastically established early in development.

In humans, a possible example of a facilitated epiallele is observed in fragile-X syndrome, a condition associated with expansion of a repetitive CGG sequence in the 5' untranslated region of the X-linked *FMR1* gene. This expansion predisposes to aberrant methylation and reduced *FMR1* expression, producing the Fragile X phenotype. Furthermore, in a family of 5 brothers all carrying the same predisposing X allele, methylation and FMR1 expression are mosaic and variably expressed amongst the brothers, producing phenotypes ranging from silent to fully penetrant (75).

1.3.1.3 Pure epialleles

These epialleles arise independently of genetic variation. In such cases it is thought that environmental or 'stochastic' factors act to induce epigenetic variance, subsequently producing phenotypic plasticity (55) (76) (Figure 1e). A range of specific environmental factors have been implicated, including diet and drug/toxin exposure, whilst in many cases, the inducer remains ill defined. The concept that environment can induce epiallelic variance, with mitotic and/or meiotic stability, seems particularly relevant to elucidating the mechanisms of fetal programming. Animal models in which maternal low protein programmes a cardiometabolic phenotype in offspring have been used to identify epigenetic differences associated with this programming. One such study has identified hypomethylation at the glucocorticoid receptor (GR) and peroxisome proliferator-activated receptor alpha (PPARα) gene promoters in offspring exposed to the programmable insult and, if stable, may be examples of pure epialleles resulting from the *in utero* environmental insult (77). Further evidence for the epigenetic role in programming comes from the finding that supplementation with the methyl donor, folate, prevents the hypomethylation in response to low protein (78). The phenotypic consequence of exposure to low protein diet is an insulin resistant phenotype in offspring (79), replicating the observation from human cohorts of maternal famine exposure leading to excess risk of type 2 diabetes in offspring (47). These comparisons raise a tempting possibility that human epigenomic studies may also enable the identification of pure epialleles, and indeed a recent study of fetal programming has identified stable epigenetic variants associated with in utero nutritional supply (80). This notable study will be discussed in further detail in section 1.5.2 and Chapter 5.

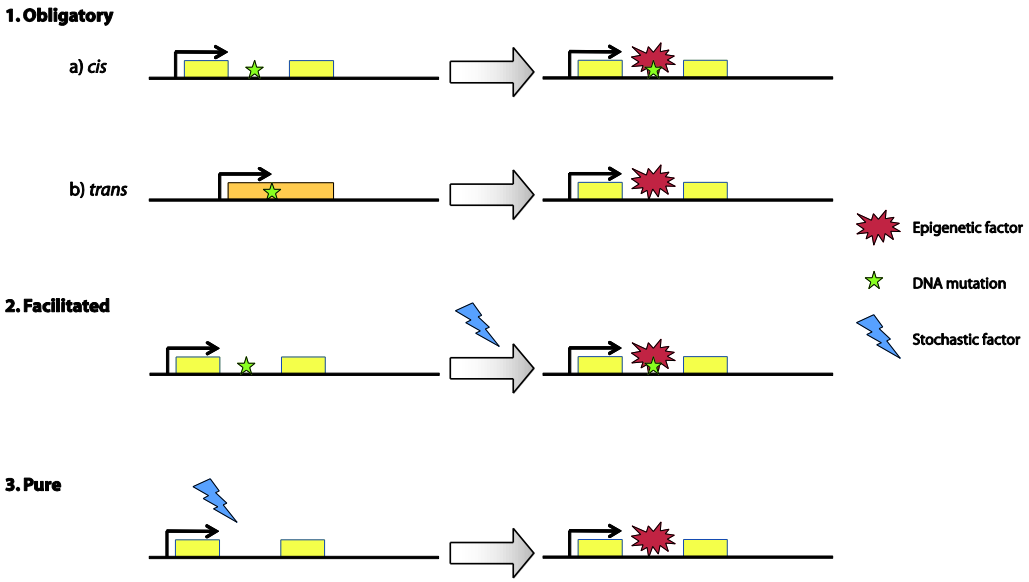


Figure 1d. Types of epialleles. 1. Obligatory epialleles. A genetic mutation (green star) either occurring at the site of the epiallele (red) (in *cis*), or at a distinct genomic location (in *trans*) is directly causative of epiallele formation. 2. Facilitated epialleles. A genetic mutation (green star) is a necessary substrate for epiallele (red) formation, but additional stochastic (blue bolt) factors are also required for epiallele formation. 3. Pure epialleles. Exist in the absence of any underlying genetic change and are derived as a result of stochastic factors (blue bolt) alone. Reproduced from Finer et al, (81).

1.4 Epigenetic changes during the mammalian life course

During the mammalian life course, the epigenetic profile of an organism undergoes a series of developmental changes and processes by which it maintains itself. Waddington (1951) originally proposed the concept that multiple stochastic influences during the lifecourse of an organism determines its fate (Figures 1e and 1f).

With further understanding of both the normal and aberrant processes that influence the epigenetic state of an organism, it is possible to understand the relative importance of different developmental time periods in determining its phenotypic fate. This knowledge is likely to be of particular importance in understanding the aetiology of type 2 diabetes and determining mechanisms underlying environmental and genetic susceptibility, rather than just epidemiological association.

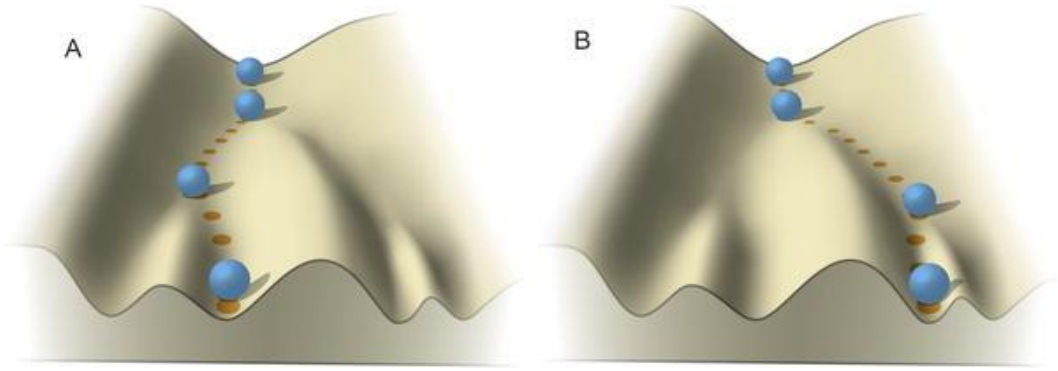


Figure 1e. The varied fate of an organism, determined by its genotype and its phenotypic development through its exposure to multiple stochastic influences during its lifecourse, represented by a ball (genotype) travelling through a varied environment (landscape) reproduced from Mitchell (82). This indirect relationship of genotype to phenotype is mediated by epigenetic influences. The diagrams represent two "runs" of the developmental process in two individuals (A and B) with the same starting genotype (as in monozygotic twins, for example). These two individuals therefore inherit the same probability of developing a certain phenotype but may have different actual phenotypic end points, determined by stochastic factors and environmental effects, especially at critical points.

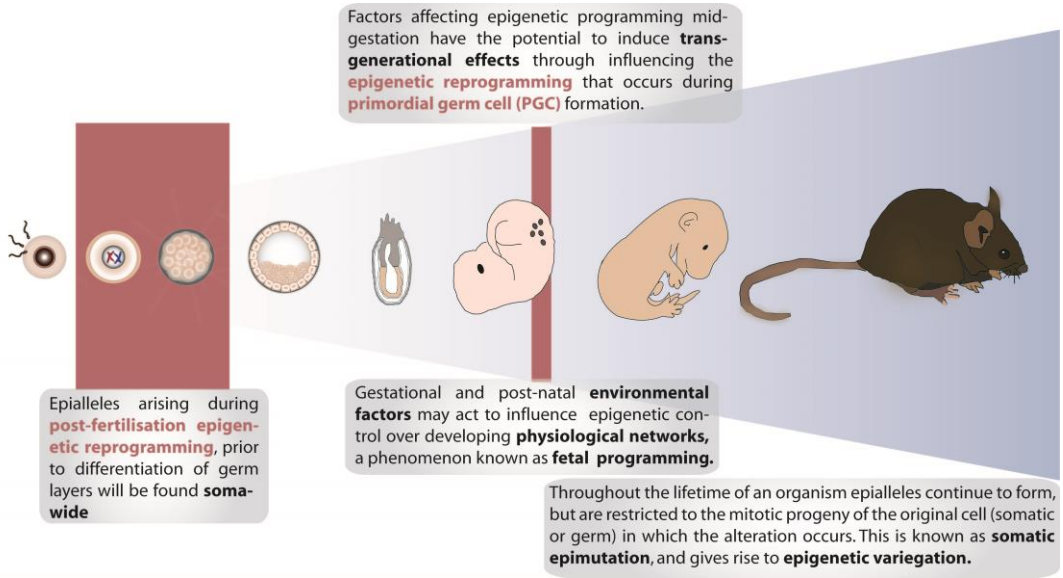


Figure 1f. The origins of epialleles. The developmental period at which an epiallele originates will determine how it might influence phenotype, as well as the potential for it to be meiotically heritable. Reproduced from Finer S et al (81).

1.4.1 Gametogenesis, embryogenesis and fetal development

In mammalian development, two periods of genome-wide epigenetic reprogramming occur, during which time development-specific epigenetic marks are established (figure 1g). The first event is timed during the formation of germ cells (gametogenesis) and is required for the establishment of a unique germ-cell specific gene expression signature, including the erasure and re-establishment of parental imprints and reactivation of the inactive X chromosome in preparation for meiosis (83). The second event occurs post-fertilisation, when maternally and paternally contributed genomes are processed differently; the male genome sees rapid replacement of its protamines with histones, and DNA methylation patterns are erased across male and female genomes by active and passive mechanisms, respectively (84).

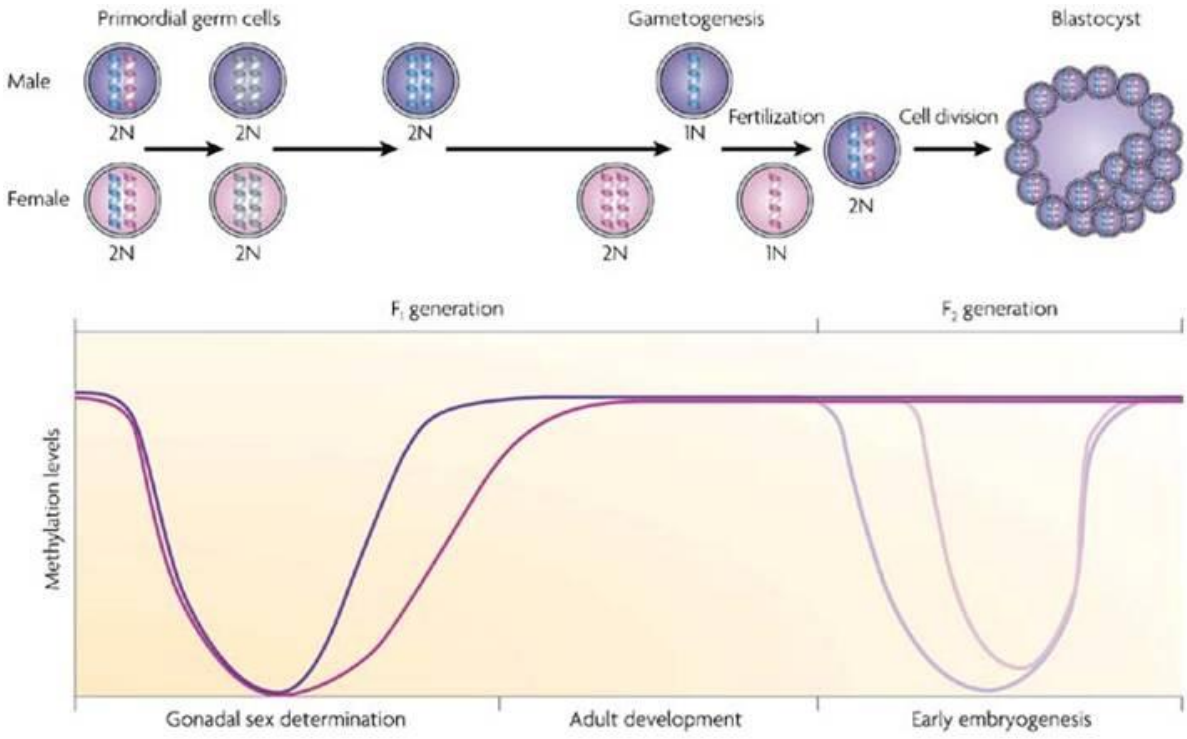


Figure 1g. Changes in methylation status during gametogenesis and early embryogenesis, reproduced from Jirtle (7)

Incomplete erasure or variation in the re-establishment of epigenetic marks in either of these two rounds of reprogramming has the potential to give rise to epialleles. The creation of these epialleles during gametogenesis or early embryogenesis results in their penetrance through all somatic lineages, exemplified by the human soma-wide epigenetic silencing of the mismatch repair gene, *MLH1* (85). The potential for disruption of these complex processes by external influences is great. Adverse environmental factors, such as gestational diet and toxin exposure, have been found to induce epialleles during these susceptible developmental windows. For example, when pregnant dams carrying A^{vy}/a offspring are fed a methyl donor supplemented diet mid-gestation, not only the F1 show a shift in phenotype towards increased silencing of the A^{vy} epiallele, but also the F2, indicating that exposure of the developing germ cells in F1 to methyl donors influences the A^{vy} epigenotype in a manner that survives post-fertilisation reprogramming (86).

The key role of the placenta in fetal development is well understood at a physiological level. Recently, patterns of placenta-specific methylation have been identified, for example in the regulation of vitamin D-responsive genes (87). The functional consequences of these unique mechanisms of epigenetic regulation at the materno-placental-fetal interface are not clear, but evidence is emerging for their role in fetal growth (88).

Furthermore, the application of epigenetic studies to models of fetal programming has yielded substantial insights into their possible role in the developmental origins of adult disease. DNA methylation changes have been identified at candidate loci in offspring exposed to the Dutch Hunger Winter *in utero* (89). These methylation changes are small and, to date, no association with gene expression or phenotypic outcome has been described. However, more direct evidence for the involvement of epigenetics in the establishment of lifelong gene expression profiles in response to environmental factors has been provided by animal studies of both nutritional and behavioural models (90) (91).

1.4.2 Adult life and ageing

Studies have suggested that perturbation of the epigenome can occur from stochastic environmental influences during the lifecourse, characterised most notably in the epigenome of monozygotic twins that become more dissimilar as twins age (76). In eliminating genetic differences, twin study design is able to identify potential stochastically-induced epigenetic variation, and recent studies have identified age-associated differences at bivalent chromatin domains, regions with important epigenetic regulatory potential (92). However, to date, no

such studies have been able to associate conclusively a link between methylation variation with phenotype. Larger-scale genome-wide studies of twins are likely to be able to uncover some of these associations, as well as the possible role of other genomic differences such as structural variation.

1.5 Role of epigenetic variation in aetiology and pathogenesis of Type 2 diabetes

Whilst epigenetic variation provides a tantalising means of studying type 2 diabetes aetiology, it is crucial to consider the potential interacting genetic, environmental influences on complex disease phenotypes and these should be taken into account when designing studies to determine its molecular basis. Complementary to this is to consider how and where these interactions occur and whether there are particular susceptible periods in the life course. Consideration must also be given to whether these interactions affect the organism as a whole, or induce tissue-specific effects with localised effects on gene expression and function.

1.5.1 Genetic-epigenetic interactions

The role of aberrant epigenetic marks is well established at single gene loci with rare imprinted disorders (93) and as well in cancer genomes with global hypomethylation and locus hypermethylation (94). Recent investigations in other complex diseases, such as type 1 diabetes (95) have been successful at identifying some preliminary insights into genetic-epigenetic interactions on a genomic scale. However, the relationship between epigenetic modifications and genetic variation across the genome is poorly understood. In complex diseases such as type 2 diabetes, genetic risk is influenced by polygenic influences and could therefore also be modulated by a myriad of interacting epigenetic modifications. It is hypothesised that such epigenetic variation could underlie the missing heritability of complex disease, but to date, there is only limited evidence to support this notion.

The first suggested evidence for direct genetic-epigenetic interactions in type 2 diabetes came from the identification of parent-of-origin specific associations of genetic susceptibility variants, putatively due to the occurrence of imprinting (65), whereby allelic expression occurs in a parent-of-origin specific manner. Since this observation, only one study identifies a

possible functional role of an imprinted locus in the aetiology of type 2 diabetes; Small et al (96) show that variants in KLF14, a maternally-expressed transcription factor can modulate adipose tissue gene expression via *trans* effects in relevant metabolic pathways. Currently, there is no other evidence for imprinting playing a role in the aetiology of type 2 diabetes, although related phenotypes can arise from rare single-gene imprinting disorders such as Silver-Russell and Beckwith-Wiedemann syndromes (97), and transient neonatal diabetes (98). Interestingly, despite their common role in developmental processes, imprinted genes are not thought to be susceptible to environmentally induced fetal programming (99).

More recently, many studies have identified epigenetic variation at genes with a known role in diabetes pathogenesis, e.g. increased DNA methylation associated with decreased PDX1 expression in human pancreatic islets from individuals with type 2 diabetes (100). However, such studies do not define whether these epigenetic variants are in fact stable epialleles, and whether they are induced stochastically, through environmental variation (e.g. pure epialleles) or in association with genetic variation (obligatory and facilitated epialleles). The association between genetic and epigenetic variation is an important consideration given their potential dual role in the development of complex disease phenotypes. Studies that integrate assays of genetic variants, e.g. single nucleotide polymorphisms, copy number variants etc, and epigenetic variants are required to understand these interactions in more detail.

1.5.2 Environment-epigenetic interactions

The notion of a pure epiallele, described above, suggests that environmental and/or stochastic factors can alone induce metastable epigenetic variants with secondary effects on gene function. To date, evidence for pure environmental-epigenetic interactions that occur independently of genetic variation and have a phenotypic consequence in humans is limited. In contrast, a range of studies using inbred mouse and rat strains have identified stable epigenetic variants induced through stimuli in the maternal environment, and the examples of these have been discussed in section 1.3.1.3.

Whilst many studies attempt to show evidence that pure epialleles exist in humans, most do not answer the difficult questions that are crucial to their definition: are these truly metastable changes that could be inherited through subsequent generations, do they predate disease onset and therefore have an aetiological role, and/or do they arise with complete independence from genetic risk variants? Secondly, most studies identifying methylation variation induced through environmental stimuli have not addressed whether these variants have an effect on gene function.

A study of epigenetic variation in type 1 diabetes has provided some evidence that epigenetic variants can predate the onset of type 1 diabetes, using longitudinal sampling from individuals before and after the onset of disease (101). This study also used disease-discordant monozygotic twins as the primary study group and can therefore conclude that the variation in DNA methylation they identify is occurring independently of genetic variation. No single environmental precipitant is identified as the cause of these epigenetic changes, nor are they proved to be metastable, however this is an important study in understanding the nature of environment-epigenetic interactions and their role in the onset of complex diseases.

Waterland and colleagues have performed a study that goes further towards identifying environmentally induced metastable epialleles in humans (80). The authors designed their study using an important principle of murine metastable epialleles, that they show inter-individual variation and are present across germ layers. A methylation-specific microarray was used to identify regions of the genome exhibiting inter-individual variation in DNA methylation that were present across mesodermal (peripheral leucocytes) and ectodermal (hair follicles) cell lineages. Of the 107 loci fulfilling these criteria, 34 were associated with SNPs, a further 35 were associated with probable copy number variation in sub-telomeric regions, and of the remaining loci, only 8 were technically validated by bisulphite-pyrosequencing. Further support of these candidate regions being stochastic or environmentally mediated, rather than genetically induced was provided by the observation that the majority of them showed inter-twin variation across 23 monozygotic twin pairs. Finally, the methylation state of these candidate regions was assayed in Gambian offspring where season of birth is associated different birth phenotypes and developmental programming. The authors found a statistically significant association ($p<0.05$) between season of birth and methylation at 5 of their selected candidate regions, with an increase in methylation of around 5% associated with birth during the nutritionally replete rainy season. The authors conclude that their findings provide evidence of metastable epialleles that have occurred independently of genetic variation and are induced through the environment at specific regions of methylation variation.

Beyond these issues, what remains is how to interpret the studies showing the direct influence of environmental perturbations on epigenetic marks where the study design does not seek to identify them as stable epialleles. Whilst the identification of a stable epiallele is key to understanding potential trans-generational inheritance of epigenetic marks, it is not necessarily crucial to understanding the role of the environment in inducing epigenetic variation and complex disease phenotype in a single generation. There is a wealth of studies that identify epigenetic variants in association with complex disease phenotypes; for example, small-scale targeted studies have identified an association between insulin resistance in elderly

subjects and promoter methylation at NDUFB6 (102) and methylation differences at PPARGC1A in human islets from twins with and without type 2 diabetes (103). These findings, at small numbers of CpG sites in gene candidates, do not provide sufficient evidence of their existence as stable epialleles but do provide insights into their association with environmental factors rather than genetic variation. The key question to ask of the observed environmental-epigenetic interactions and their association with disease phenotype is whether this is a primary and causative association or a secondary consequence of disease pathology. Away from the difficult question of whether or not an epigenetic variant is an epiallele, the differentiation between primary and secondary epigenetic variation is crucial if trying to understand their role in disease risk. Other studies have already shown in detail how epigenetic variants can be part of disease pathology, rather than being causative, in diabetes. El-Osta and colleagues (104) (105) (106) have identified histone modifications induced through transient hyperglycaemia at the promoter of the NF-kB subunit p65 in vascular cells with consequences on gene expression. These studies do not look at DNA methylation changes but provide a fascinating insight into how specific cellular environments associated with diabetes can induce long-standing changes to gene function that may underlie metabolic and vascular complications of diabetes and molecular memory of early glycaemic control. Other studies have shown variable DNA methylation in the leptin gene of women with gestational diabetes (107) and the existence of genome-wide DNA methylation variation in diabetic nephropathy (108). These studies contribute to a wider body of animal experiments, but all identify epigenetic variation that is likely to be a secondary phenomenon to the disease itself.

1.5.3 Gene and environment interacting through the epigenome

The difficulty in understanding genetic-epigenetic and environmental-epigenetic interactions as separate entities may be due to the fact that the majority of epigenetic variation occurs in tandem with genetic variation and is modulated through the environment. In addition, these dichotomous interactions are difficult to elucidate experimentally; the identification of pure epialleles necessitates complex study design that includes monozygotic twins and/or inbred animal strains. Identification of functional genetic-epigenetic variants requires large sample sizes, allele-specific functional studies and complex computational methods to identify *cis* and *trans* effects. The majority of published research on the role of epigenetic variation in complex disease aetiology is not designed to address the exact nature of these interactions at this level of detail and the conclusions of most published epigenomic studies at this time suggest putative roles of both environment and genetic variation. Furthermore, it could be argued

that the lack of evidence of pure epialleles in the many epigenomic studies performed to date may in fact reflect the fact that they are a rare occurrence.

As discussed earlier, the aetiology of type 2 diabetes and related disorders is thought to incorporate complex interactions between gene and environment and that early developmental life may be particularly susceptible to disruption of either, e.g. via polymorphisms or an aberrant environmental factor. Studies of fetal programming provide crucial insight into these convergent factors, and the impact of inherited genetic risk and environmental perturbations in utero can provide exciting insights into the nature of these interactions and their role in complex disease aetiology.

The body of evidence supporting the notion that an adverse maternal environment can predispose the developing fetus to adult disease in later life assumes a combined role of genetic and environmental influence. The potential for epigenetic regulatory mechanisms to lie at the interface between parental genotype, in utero environment and the developing offspring is strong. Identification of defined environmental insults in pregnancy has provided a useful starting point with which to explore the molecular mechanisms underlying fetal programming and there is now a wealth of preliminary data that is starting to elucidate these environmental-genetic-epigenetic interactions. The maternal low protein model, described above, is particularly well studied and multiple studies have shown that exposed offspring have a predictable phenotype of insulin resistance. Complementary molecular studies in this model have identified small methylation differences in relevant genes that have a role in insulin resistance (77), highlighting the convergence of genetic and environmental risk through epigenetic modifications. The low protein mouse model have lent support to human cohorts in which growth-restricted offspring born during periods of famine were found to have small methylation differences in IGF2, GNASAS and IL-10 and other genes, several of which are already implicated in the aetiology of diabetes. These changes occur in a time- and sex-specific manner, implicating gametogenesis as the period most susceptible to the environmental exposure (89) (109). However, these studies do not elucidate the functional consequences of the small methylation differences observed, nor do they address the important concern that the end phenotype associated with the programmed event may in fact be the cause of the epigenetic differences. Few studies can identify a true aetiological role for these interactions and sub-clinical phenotypes in human studies may further obscure the difference between primary and secondary changes. Indeed a robust study that detected stable epigenetic variants in Gambian offspring according to season of birth (as a surrogate for maternal nutrition) did not find any differences in DNA methylation at IGF2, GNASAS or IL-10 (80). Serial sampling from human cohorts could help to overcome these difficult issues with

the ability to identify the temporal relationship of epigenetic changes with phenotypic outcome. Recent work using Guthrie cards as a means with which to generate epigenomic insights of DNA methylation may provide a feasible and practical platform to address these important questions in the future (101).

One study that does identify the functional effects of epigenetic variants has been performed by the researchers who identified promoter methylation differences at NDUFB6 and PPARGCIA, described earlier (110). Muscle biopsies were taken from men in their 20s who had been born either at an extreme of low birth weight (LBW), or normal birth weight. Those who had had a low birth weight had convincing signs of early insulin resistance and obesity and differences in methylation at 2 CpG sites in the PPARGC1A promoter compared to those who had been born at normal birth weight. After high-fat feeding, the normal birth weight group showed increased methylation at the same CpG sites in association with increased expression of PPARG1CA, whereas those with LBW did not. Mouse studies of intrauterine growth restriction also support a true causative role through epigenetic silencing of a gene involved in pancreatic development prior to the onset of any programmed phenotype (111). A tissue-specific view of epigenetic variation in diabetes is also important in detecting its functional role, and this has been highlighted by Bhandare et al (112) in their genome-wide profiling of DNA methylation and histone modifications of human islets. They find, surprisingly, the absence of 'active' histone marks (H3K4me3) associated with transcriptionally active genes, such as those encoding the hormones insulin and glucagon, suggesting the need for detailed functional models to understand how epigenetic modifications, as well as their interaction with genetic and environmental factors, relate to disease aetiology and pathology.

In trying to characterise the role of genetic and environmental factors in epigenetic mechanisms of fetal programming, care must also be taken in the study design to address variables that may confound environmental and/or genetic risk factors and, in turn, influence outcome. The maternal low protein experiments offer a useful model of programming where the exposure is itself independent of genetic risk. In contrast, other models, such as that of programming through maternal hyperglycaemia may be prone to confounding by common genetic risk of diabetes inherited from the mother (with gestational diabetes) to the child, and studies also show a high prevalence of vitamin D deficiency in mothers with gestational diabetes (113). The 'adverse environments' described in human models of fetal programming may also include confounding variables and need careful thought to determine the nature of the complex nutritional and/or metabolic insults so that these can be elucidated. For example, the ability for famine exposure to induce developmental programming may in fact be due to

single nutrient deficiencies such as protein deficiency, or multiple factors related to the dietary insult itself or other factors associated with famine exposure such as recurrent infections or reduced physical activity. These complex factors are important to consider when comparing similar human studies and also when using animal models to simulate the human exposure.

1.6 Application of epigenomic studies to understanding the aetiology of type 2 diabetes

Much of the current research being performed to understand the role of epigenetics in the fetal origins of and gene-environment interactions in complex disease aetiology employs a gene-specific or targeted approach. Targeted studies offer a useful insight into specific molecular mechanisms or regions of interest within the epigenome, chosen using *a priori* hypotheses. Whilst this approach may be useful to complement existing candidate-based studies, it cannot be used to generate the unbiased insights into gene-epigenetic-environment interactions that are required in these early stages of understanding complex disease aetiology on this level. The approach of the work presented in this thesis is to use genome-scale and genome-wide approaches to generate novel insights into the aetiology of type 2 diabetes, and to generate insights for downstream hypothesis-driven studies.

1.6.1 Genome-scale genetic-epigenetic studies

Genome-scale studies implement genomic technologies to identify molecular events at multiple pre-defined areas of the genome and commonly use custom-designable arrays. These technologies have mostly been superseded by array-based and whole-genome-based approaches of recent years, but have been used in the study of type 2 diabetes presented in Chapter 3.

1.6.2 Genome-wide and whole genome studies of gene-environment-epigenome

Just as the unbiased genomic techniques have yielded an in depth understanding of candidate gene disease associations in recent years, whole 'epigenome' study is now being used to characterise the epigenetic mechanisms underlying complex phenotype and disease. Characterising and profiling the epigenomic state (including its variety of epigenetic marks,

their plasticity and interaction) brings significant technical challenges on top of those encountered in pure genomic investigation. DNA methylation is the best-studied epigenetic profile due to its relatively predictable interaction with the underlying DNA sequence and effects on gene expression, and several affordable and robust techniques now exist with which to do this. An optimal technical platform to study DNA methylation on an epigenomic scale is one in which methylation status can be identified quantitatively and at high resolution across whole genome, and one that is combined with simultaneous identification of the underlying genetic sequence.

MeDIP-seq has been used as an experimental platform that has been applied to the study of DNA methylation across mammalian genomes (114). The technique involves methylated DNA immunoprecipitation of short DNA fragments using a monoclonal antibody specific to 5-methylcytosine - enriched methylated DNA fragments are isolated for use in the sequencing stage of this process. Next Generation Sequencing (NGS) platforms are used to sequence methylated DNA fragments, having undergone the appropriate library preparation, and can produce 10s of millions of sequence reads in a few days that can be aligned to a reference genome and provide extensive coverage of methylated regions. Detection of the efficiency of MeDIP enrichment across overlapping genomic fragments allows the quantitative estimation of methylation across the genome and identification of 'differentially methylated regions' (DMRs); this will be discussed in more detail in Chapter 2.

The potential applications of MeDIP-seq are wide-ranging, allowing unbiased detailed study of the dynamic processes of DNA methylation, its role in developmental programming and the characterisation of disease-associated aberrant marks. Methylation profiles generated by MeDIP-seq are freely accessible via the Ensembl genome browser display. However, limitations of the MeDIP-seq approach exist, notably that it does not offer the single base pair resolution of bisulphite conversion coupled with Next Generation Sequencing (BS-seq). Bisulphite treatment of single-stranded DNA causes the selective conversion of unmethylated cytosines to uracil, leaving methylated cytosines unchanged, and thereby allowing the identification of methylation status per nucleotide residue. This gold-standard technique has now been successfully applied to animal (115) and human-based studies (116), but at the current time the sequencing requirements (i.e. cost and time) are prohibitive for many researchers. Reduced representation bisulphite sequencing (RRBS) provides a more pragmatic approach to BS-seq, using a MspI restriction digest specific to CpG sites and therefore enriches for regions of DNA methylation (117). Bisulphite conversion is then applied and after PCR, DNA libraries are applied to Next Generation Sequencing (NGS). This approach provides single nucleotide resolution, however by limiting itself to CpG dense regions, it has less extensive

coverage MeDIP-seq, and may miss DMRs in CpG-poor regions such as gene bodies, the functional relevance of which is not fully determined at this time. Another potential bias for RRBS- and BS-seq is the possibility of incomplete bisulphite conversion which may occur when double-stranded DNA is not fully denatured or has re-annealed, and is therefore not subject to the conversion that single-stranded DNA undergoes. If this occurs, unmethylated cytosines may represent experimental artefact and introduce bias into the analysis of data generated from this technique.

The three epigenomic approaches described above are reliant on detailed and expert bioinformatic analysis that can be time-consuming and short in supply given the rapid expansion in data since the recent technical advances in molecular biology. However, recent advances in the computational analysis of these techniques offer the exciting possibility of integration with related studies of histone modification (ChIP-seq) and functional investigation (RNA-seq) to build integrative maps (59).

Array-based techniques, in combination with MeDIP or restriction digests and hybridisation of methylated against unmethylated DNA, may be used to identify DMRs. These offer a cheap method of assessing DNA methylation at multiple genomic sites in large numbers of samples and generate data quickly often using relatively simple bioinformatic analytic techniques. Numerous examples of these platforms exist, including the Illumina HumanMethylation 450 BeadChip, a relatively cheap and easy platform to use in studies where large numbers of samples are being analysed and prohibit the use of BS-seq, RRBS-seq or Medip-seq.

Identification of regions of epigenetic variation, e.g. DMRs, can be validated using a targeted approach in a larger sample size using standard genetic sequencing of bisulphite-treated DNA and can be processed in a cost- and time-efficient manner.

1.7 Determining the functional role of epigenetic variants

To date, epiallelic variation is functionally characterised in terms of its influence over gene transcription. However, it is important to establish a deterministic link between an epiallele and altered gene expression, which could be influenced by unlinked genetic modifiers. Inbred mouse models provide a useful resource in which large numbers of genetically similar individuals can be simultaneously assessed for both the presence of the epiallele and its putative functional effect. An integrative approach, combining several genome-scale data sets, offers the most powerful strategy to detect epialleles that, in addition to *cis* effects, may exist mono-allelically, or act by influencing distant enhancers. The power of combining epigenomic and transcriptomic mapping to detect epialleles with functional and phenotypic outcome is huge, and is likely to yield considerable data resources that could be made accessible publicly though browsers such as Ensembl. However, at present, cost and analytical barriers limit the potential to scale up these experimental approaches to the sample sizes that have given insight into complex disease pathogenesis from the detection of genetic variants.

As previously discussed, the acquisition of an epiallele may be determined by the interaction of genetic and/or environmental factors. When designing an experimental approach to not only identify the presence of an epiallele, but also determine the factors that have led to its existence, appropriate study design is vital. Facilitated epialleles must be identified in the context of both genetic and epigenetic variation. Obligatory epialleles, which are induced solely through the existence of a genetic variant, may be detected by identifying the genetic variant or its functional consequence in the first instance and the epigenetic event second (118). To study pure epialleles, an isogenic environment offers a means with which to increase the power to detect non-genetic events, and this has been successfully performed in inbred mouse strains (119) and studies of disease-discordant monozygotic twins (95). Identification of environmentally-induced epialleles in human cohorts should also take careful note of potential confounding factors when considering their aetiological role (120).

Fetal programming of adult disease via environmentally determined epigenetic variants, with or without the association of genetic susceptibility, has been discussed earlier. To date, studies are limited and focus on individuals exposed to extreme environmental insults or animal models. Larger-scale human studies encompassing milder environmental insults or behavioural patterns will increase our understanding of the complex interaction of epigenetic marks, genetic risk and environment, and how they predispose to disease during an individual's lifetime. This knowledge is particularly pertinent given the increasing global

prevalence of complex diseases and furthermore, may provide insight into how environmental factors may be modified as a means of disease prevention. The identification of disease-associated epialleles either associated with disease induction, or pathogenesis, has the potential to translate into screening, diagnosis and identification of biomarkers (121). An understanding of the epiallelic contribution to disease pathogenesis will also provide a platform for the development of targeted drug treatments.

In all studies identifying epigenetic variants in association with disease, it is imperative to determine a causal role if trying to understand the aetiology of disease. This may be determined with functional experiments, however the ability to identify changes in gene expression due to epigenetic variation is complex and requires a careful experimental approach that is likely to incorporate the complexity of allele-specific expression and tissue-specificity (122) (123).

1.8 Objectives

This introduction has highlighted how the study of epigenetic variation may yield insights into the aetiology and pathogenesis of type 2 diabetes and related cardiometabolic conditions such as obesity. The application of discovery-based studies to understand how multiple interacting genetic and environmental factors influence type 2 diabetes risk through the epigenome has been introduced and will contribute to the main objectives of this thesis. In addition, the importance of the developing epigenome in fetal programming of cardiometabolic diseases, and its apparent susceptibility to environmental perturbations has been highlighted and will form the later objectives of this thesis. The broad objectives of this thesis are as follows:

- To identify epigenetic factors involved in susceptibility to Type 2 diabetes as a route to understand disease aetiology and pathogenesis.
- To apply animal and human models to elucidate the role of epigenetic mechanisms in fetal programming of Type 2 diabetes and related cardiometabolic conditions.
- To use 'discovery-based' techniques to determine interactions between genetic, epigenetic and environmental influences, on an epigenomic scale.

More specifically, these objectives will be applied to the following questions and contexts:

- **Chapter 3: Type 2 diabetes study**. This study will identify epigenetic variation at specific regions of genetic susceptibility to Type 2 diabetes to characterise putative regions of genetic-epigenetic interactions. A Medip-chip approach will be used to profile DNA methylation on a genome-wide scale and methylation profiles will be built across susceptibility haplotypes.
- **Chapter 4: Pune Maternal Nutrition Study**: Epigenomic techniques will be used to identify differentially methylated regions in whole blood from offspring from this cohort in which maternal one-carbon cycle defects have been shown to induce a programmed phenotype of insulin resistance and obesity in offspring.
- **Chapter 5: Matlab Famine Study**: Genome-wide techniques will be used to identify differentially methylated regions in offspring exposed to famine in utero or early childhood or unexposed.
- **Chapter 6: Mouse model of gestational diabetes**: A mouse model, analogous to human gestational diabetes, will be used to simulate fetal programming from exposure to hyperglycaemia in utero. The primary aim of this model will be to study epigenomic

variation in offspring to identify the consequences of exposure to hyperglycaemia, and hence elucidate mechanisms of fetal programming. The secondary aims of this model will be to investigate the role of tissue-, age- and sex-specific epigenetic factors and functional outcomes.

- **Chapter 7: Human model of gestational diabetes**: A human cohort of women with and without gestational diabetes will be developed with which to characterise the combined metabolic and nutritional influences of pregnant women of Asian origin living in London. This cohort will then be used to identify epigenetic variation associated with these pregnancy environments in cord blood and placenta. Results from this study and Chapter 6 will be used to complement each other.

Chapter 2: Methods

2.1 Materials

2.1.1 Chemicals and reagents

Agarose for routine use, Sigma-Aldrich

Ethidium bromide, Sigma-Aldrich

Triton X-100, Sigma-Aldrich

Tris-HCL, BioChemika Fluka

Ethylenediaminetetraacetic acid disodium dehydrate (EDTA), Sigma-Aldrich

Phenol Solution, Sigma-Aldrich

Phenol:Chloroform:Isoamyl alcohol 25:24:1, Sigma-Aldrich

Chloroform 99% minimum, Sigma-Aldrich

Orange G, Sigma-Aldrich

2.1.2 Buffers and Solution

All buffers and solutions were made with milli-Q water unless otherwise stated

TE Buffer: TricHCl 10mM, EDTA 1mM, pH 8.0

Proteinase K Buffer: TrisHCl , EDTA 10mM, SDS 0.5%, pH 8.0

2x IP buffer: 20mM sodium phosphate buffer pH7, 280mM sodium chloride, 0.1% Triton X-100

1x TBE: tris-borate, EDTA

DNA loading dye: 1x Orange G

D-glucose, diluted to 100mg/ml in sterile water

2.1.3 Enzymes, antibodies and kits

5-methyl cytidine antibody 500µg/50µl, Diagenode

Dynabeads – sheep anti-mouse IgG, Invitrogen

GenomePlex® Whole Genome Amplification Kit, Sigma-Aldrich

Klenow Fragment (3'-5' exo-), NEB

T4 Polynuclotide kinase, NEB

Quick Ligation kit, NEB

Proteinase K, PCR grade, Roche

Ribonuclease A solution from bovine pancreas, Sigma-Aldrich

Phusion HotStart DNA polymerase, NEB

Platinum Taq Polymerase, Invitrogen

Amplitaq Gold, Applied Biosystems

qPCR MasterMix Plus w/o Ung for sYBR Assay Low Rox, Eurogentec

QIAquick PCR purification kit, QIAGEN

QIAquick Gel extraction kit, QIAGEN

QIAamp DNA mini blood and tissue kit, QIAGEN

QIAamp DNA maxi blood and tissue kit, QIAGEN

Zymo DNA clean and concentrator kit, Cambridge Bioscience

Bisulphite conversion kits - EZ Methylation-Gold and EZ Methylation in single column and 96-well plate formats**,** (Zymo research)

Sybr green Low ROX, Eurogentec

RP1 purified 0.05µmol lyophilised primers (Sigma Genosys)ABI optical PCR plates

Primers were ordered from SIGMA (Desalted, scale:0.2)

Adhesive PCR film (cat# AB 0558)

MESA Blue qPCR Master Mix Plus for SYBR Assay (cat# RT-SY2X-03+WOUB 600rs)

2.1.4 Specific laboratory equipment

Accuchek aviva blood glucose monitor

Diagenode Bioruptor

Agilent 2100 Bioanalyser and DNA 1000 and HS kits

7500HT Real time PCR system

Nanodrop spectrophotometer

Quant-iT Pico green fluorometer and ds DNA BR assay kit (Invitrogen)

Dark Reader transilluminator

2.1.5 Genomic and sequencing-based equipment

Nimblegen custom-designed microarray, Nimblegen

Illumina GAIIx Genome Analyzer, UCL Genomics, Babraham Institute and Barts and the London Genome Centre

Illumina HumanMethylation 450 BeadChip, UCL Genomics and Barts and the London Genome Centre

Illumina HiScan, UCL Genomics and Barts and the London Genome Centre

Biotage PSQ HS96 instrument, UCL Genomics

2.1.6 Bioinformatic software, scripts and statistical packages

UCSC genome browser

Ensembl genome browser

Maq (Mapping and Assembly with Quality) – mapping software

BWA

BATMAN (T. Down)

USeq

Perl and JAVA scripts (written by C. Mathews, T. Down, G. Carbajosa, D. Van Heel)

R Statistical package (Ihaka R, Gentleman R, 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics Vol 5: 299-314)

Bioconductor packages: Limma

GraphPad Prism, USA

SPSS, USA

2.2 Common Experimental methods

2.2.1 DNA extraction and purification

2.2.1.1 Phenol:chloroform extraction

Frozen tissue for DNA extraction was stored at -80^{o}. One 5mm piece of tissue was added to 750uL PK buffer and 5uL proteinase K enzyme and incubated together at 55^{o} overnight. The following morning, 5uL RNAse was added and the sample incubated at 37^{o} for 60mins.

Phenol (from lower phase) was added in a 1:1 volume of phenol to sample, shaken together for 5mins and then spun at high speed for 5mins. The upper phase was removed and and put in new Eppendorf and these same steps were repeated with phenol:chloroform and then chloroform.

One-tenth of the total volume of the remaining sample (approximately 50uL) of sodium acetate was added, followed by 2.5x the total volume (usu 1.25ml) 100% ethanol. The sample was then shaken carefully until DNA became visible. The sample was then placed in a -20^{o} freezer for 1 hour and then spun at full speed for 10 minutes. The ethanol was decanted out and the DNA pellet was spun again for 10 minutes. The remaining ethanol was then removed carefully using a pipette. The tube containing the DNA pellet was covered lightly until any remaining ethanol had evaporated, but without the DNA pellet drying out.

Finally, 200uL TE buffer was added to the DNA pellet and together were heated gently at 50^{o} for 60 minutes with regular vortexing. Following this, the concentration and quality of the DNA were checked (see below).

2.2.1.2 DNA extraction and purification kits

Manufacturer's instructions were followed when using kits for DNA extraction and purification. Unless stated otherwise, an RNAase digestion was performed with all DNA extractions. DNA purification was performed using Qiagen QIAquick PCR purification kits (where elution volumes >14uL were required) or Zymo Clean and Concentrator kits (where elution volumes of <14uL were required).

2.2.2 Checking DNA concentration and quality

2.2.2.1 Agarose gels

Unless otherwise specified, 80ml agarose gels were prepared using 2% agarose with TBE and 5µL ethidium bromide for all gels to check nucleic acid integrity, size and distribution. Gels were visualised under UV light.

2.2.1.2 Checking DNA concentration

DNA concentrations were checked using a nanodrop spectrophotometer that uses UV absorbance to estimate DNA concentration. DNA samples extracted using phenol: chloroform with a salt-based precipitation method, or those with RNA contamination, were also quantified using Qubit as this fluorometry-based system is able to distinguish between DNA and other contaminants.

DNA concentration and fragment size distribution of individual samples was checked using a Bioanalyser, during MeDIP-seq library preparation.

2.2.3 Sonication

DNA was sonicated using a Diagenode Bioruptor, in 100µL H_20, with varying starting concentrations according to the experimental protocol. Sonication was performed using the 'high' setting, in 30 second on/off cycles, with ice bath changes every 10 minutes. Sonication duration varied according to the experimental protocol, and fragment size checked using an agarose gel with 100bp ladder and/or bioanalyser. Sonication was followed by DNA purification using a Qiagen kit.

2.3 Enrichment-based DNA methylation experiments

In this section, the Medip (methylated DNA immunoprecipitation) experiment will be discussed. This enrichment-based experiment, first described by Weber et al. (107), allows the isolation of a pool of methylated DNA (in <1000bp fragments) using a recombinant antibody specific to 5-methylcytosine. Coupled with a genomic platform, e.g. microarray or Next Generation Sequencing, Medip enrichment can be used to identify quantitative methylation patterns on a genome-wide of whole genome level. Two experimental approaches, Medip-chip and Medip-seq will be discussed after a description of the Medip protocol.

2.3.1 Medip

1.5µg DNA was sonicated to produce 200-700bp fragments. 50ng DNA was set aside as the 'input' control and the remaining 1.0µg was made up to 97µL in H_2O, denatured at 100^oC for 10 minutes and then cooled on ice. Subsequently, 2.5µL 5MeC-mAb was added to the DNA solution with 100µL 2xIP buffer and incubated at 4^oC on rotator at 8rpm for 2 hours. Dynabeads (10µL) were washed in 500µL 1xIP buffer, collected in a magnetic rack, and resuspended in 500µL fresh 1xIP buffer. The DNA solution was added to the beads and then incubated at 4^oC on a rotator at 8rpm for 2 hours. After this incubation, a further 3 complete washes in 500µL 1xIP buffer were performed, and finally the beads were resuspended in 100µL PK buffer. 1 µL proteinase K was added to each sample and then incubated at 50 ^oC, 20rpm. After this digestion, DNA in solution was separated from beads in a magnetic rack, and purified using a Zymo kit, eluting in 25µL H_2O.

2.3.1.1 qPCR to determine Medip efficiency

A test to determine the efficiency of MeDIP enrichment in all MeDIP-based experiments (i.e. MeDIP-chip and MeDIP-seq) was performed by qPCR. Quantitative amplification of known methylated/unmethylated regions of the genome, as determined by the Human Epigenome Project, was used on Medip and input samples. Primer sequences targeted to these regions were used, as in table 2a.

Species	Methylation enrichment status	Primer name	Primer sequence
Human	**enriched**	6583_F	CACTCACCGTCCAGCTATCA
		6583_R	CTCCCTGACCTCCATCTTCA
Human	**enriched**	11851_F	CCAAGAGGGCTCCCTAGAAG
		11851_R	ATTTGGAAGGGACCTTGCTT
Human	**enriched**	4994_F	GGGAATATAAGGAGCGCACA
		4994_R	TCGGTTAAAACGGTCAGGTC
Human	**unenriched**	8804_F	CGAGGCGTGAGTTATTCCTG
		8804_R	CTCTTGTGGCTGAGCTCCTT
Mouse	**enriched**	IAP-R	CCCCGTCCCTTTTTTAGGAGA
		IAP-F	CTCCATGTGCTCTGCCTTCC
Mouse	**unenriched**	CSA-R	GCAACATGGCAACTGGAAACA
		CSA-F	TGGTTGGCATTTTATCCCTAG
Mouse	**enriched**	H19-R	TGGCCCTTGGACATTGTCAT
		H19-F	TGCCAGAAAGCACAAAAGCC
Mouse	**unenriched**	APRT-R	AGATCCCCGAGGCTGCCTAC
		APRT-F	TGCTGTTCAGGTGCGGTCAC

Table 2a. qPCR primers used to test success of Medip enrichment

A mastermix was made for qPCR reactions, containing 3.5µl H_20 (PCR grade), 10ng DNA in 1.0µl, 6.5µl SYBR green mix and 2.5µl primer mix per reaction. Reactions were performed in duplicate, per primer pair, and mean Ct values were plotted for Medip versus input DNA. A standard qPCR cycle was used on an Agilent 7500HT realtime PCR machine, as follows:

(1) 50°C for 2 minutes
(2) 95°C for 10 minutes
(3) 40 cycles of: 95°C for 15 seconds and 60°C for 1 minute
(4) 95 for 15 seconds
(5) 60 for 15 seconds

Successful MeDIP enrichment was determined by a higher relative Ct value for Medip versus input sample at the unmethylated locus.

2.3.2 Medip-chip

This experimental approach combines the standard Medip protocol outlined above with a microarray-based genomic platform to yield quantitative methylation profiling on a genomic scale.

2.3.2.1 Amplification of Medip and input DNA

Medip and input fractions of DNA from the Medip experiment described in 2.3.1 were amplified using Sigma Whole Genome Amplification GenomePlex kits. The initial library preparation used 10ng sonicated 'input' DNA in 10µl H_2O, and 10µL 'Medip' solution, both in duplicate. To each of the 4 reactions, 2µl 1x Library preparation buffer and 1µl library stabilization solution were added, and after vortexing were denatured at 95^{o}C for 2 minutes, cooled and spun down. Subsequently, 1µl Library preparation enzyme was added, and the samples incubated in a thermal cycler at 16^{o}C for 20 minutes, 24 ^{o}C for 20 minutes, 37 ^{o}C for20 minutes and 75 ^{o}C for 5 minutes before being cooled. A mastermix was prepared, containing 7.5µl 10x Amplification Master Mix, 47.5µl nuclease-free H_2O, 5µl WGA DNA polymerase per reaction. 60µl Master Mix was added to each sample, and then the PCR incubation was performed as follows: incubation at 95 ^{o}C for 3 minutes, followed by 14 cycles of 94 ^{o}C for 15 seconds and 65 ^{o}C for 5 minutes. The amplified samples were purified using a QIAquick PCR Purification Kit, and eluted in 53µL H_2O.

2.3.2.2 Custom-designed targeted sequencing arrays for Medip-chip

A Nimblegen custom-designed microarray was used in Medip-chip experiments. The details of the array design, hybridisation, normalisation and analysis will be discussed in chapter 3.

2.3.3 Medip-seq

The experimental approach for Medip-seq library preparation is summarised in figure 2a.

.

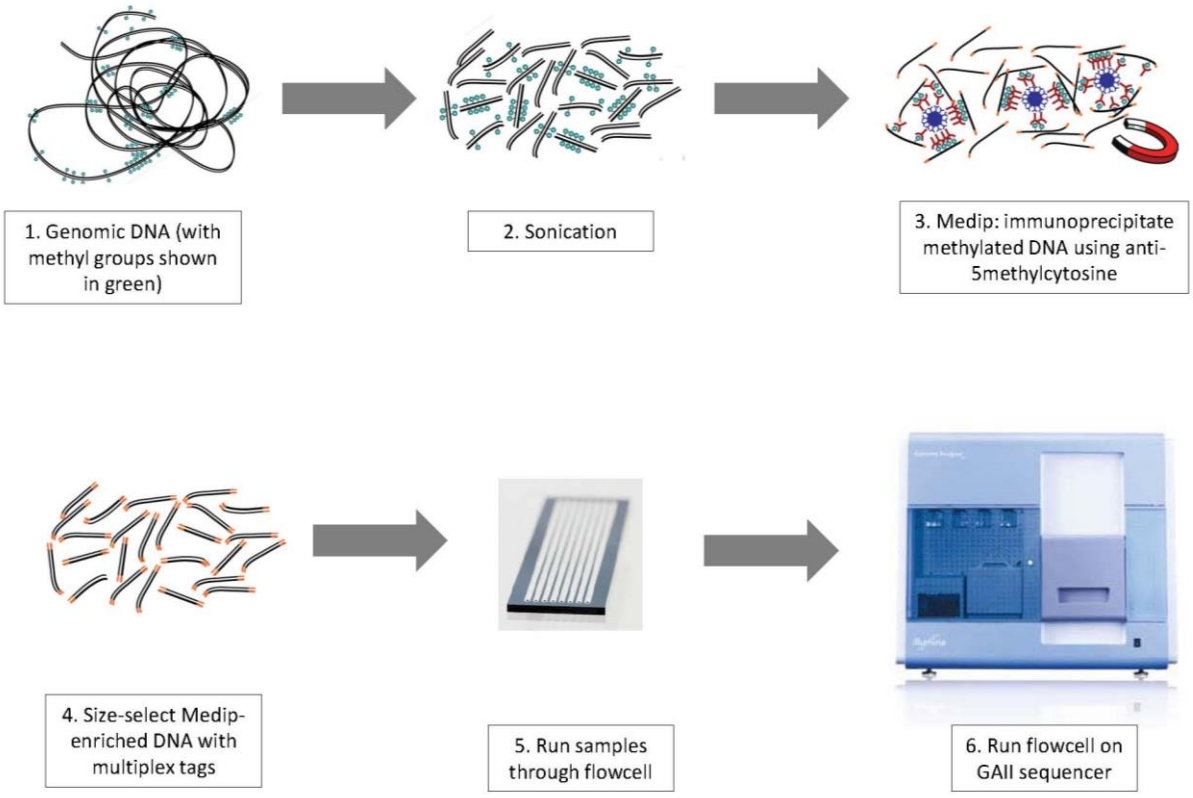


Figure 2a. Summary diagram of Medip-Seq experimental approach.

2.3.3.1 Medip-seq library preparation

Library preparation for Medip-seq experiments includes preparation of samples for Illumina GAIIx sequencing, as well as a modified Medip protocol. Compared to the Medip-chip protocol, a larger starting quantity of DNA is used to allow incorporation of sequencing adapters and careful size selection of fragments prior to sequencing. A multiplexing strategy allows a maximum of 12 samples to be combined in one library and DNA fragments to be pooled prior to size selection, thus ensuring careful size-matching of biological replicates. This strategy also allows for pooled sequencing across more than one flowcell from the same sample pool with subsequent analysis of individual or grouped samples.

Sonication**:** Three micrograms of genomic DNA were suspended in 100µl water and sonicated to median size of 150bp. Sonicated samples were purified using QIAquick PCR Purification Kit, eluting in 40µl of elution buffer (EB).

Perform end repair: blunt ends of the DNA fragments (introduced through sonication) were repaired by incubating the DNA sample (approximately 2ug DNA) with Klenow DNA polymerase (0.5µl), T4 DNA polymerase (2.5µl) and T4 PNK (2.5µl) in T4 DNA ligase buffer with 10mM ATP (5µl) and dNTP mix (2µl, 10mM) at 20°C for 30 minutes. The 3' to 5' exonuclease activity of the three enzymes removes the 3' overhangs, and the polymerase activity fills in 5' overhangs. Purification was then performed using a QIAquick PCR Purification Kit, eluting in 32µl of EB.

Add 'A' bases to 3' ends: an 'A' base is needed at the blunt 3'ends of the phosphorylated DNA fragments for subsequent ligation of the sequencing adapters, which have a single 'T' overhang at their 3' end. The DNA sample is incubated for 30 minutes at 37°C with Klenow exo 3' to 5' exo minus (3µl) and dATP (10µl 1mM) in NEB2 buffer (5µl) to create these A bases. The samples were purified using Zymo DNA Clean and Concentrator-5 kit, eluting in 12 µl of EB.

Ligate adapters to DNA fragments: paired-end sample adapter oligos (2.5µl) were ligated to the DNA samples using DNA ligase (2.0µl) in DNA ligase buffer (13.5µl) with a 15 minute incubation at room temperature. Purification was performed using the QIAquick PCR Purification Kit, eluting in 50µl of water.

After these steps, the concentrations of individual samples were checked using a Nanodrop and run on gel to ensure that no leftover adapters were present (these would form short complexes and be identified as a band of <100bp fragments).

At this stage, 50ng of DNA is set aside as an 'input' fraction, for later comparison with the 'MeDIP' fraction.

MeDIP: A modified protocol is used in the preparation of a MeDIP-seq library, and I optimised this to be able to use reduced (2ug) starting quantity of DNA. The DNA sample is suspended in 125µl water, denatured at 100 °C for 10 minutes and cooled immediately on ice for 5 minutes. 125µl of 2xIP buffer is added to the sample with 5mC Ab (4µl), followed by incubation 2 hours at 4°C with slow rotation. The sample is then added to 20µl of pre-washed Dynabeads and

incubated for 2 hours at 4°C with slow rotation. After this incubation, 300µl 1xIP buffer is added to the sample and the beads are washed in a magnetic rack in 3 cycles with 500µl 1x IP buffer, discarding the supernatant each time. After the final wash, the beads are resuspended in 100µl Proteinase K digestion buffer, proteinase K is added (2µl of 20mg/ml stock) and incubated for 2 hours at 55°C (using rotation or shaking to prevent sedimentation of the beads). The samples are purified with Zymo DNA Clean and Concentrator-5 kit and eluted in 20 µl of water.

Amplification of adapter-modified DNA fragments and introduction of multiplex tag sequences using PCR**:** both Medip and input fractions are amplified using a PCR reaction with primers designed to paired-end sample adapters. In designing the PCR in this way, amplification occurs selectively in those DNA fragments with ligated adapters. A third PCR primer is used, containing a multiplexed sequence tag that is incorporated into the DNA fragment as amplification takes place. Separate PCR reactions using one of twelve multiplex sequence tags are performed for each sample, allowing for later identification sequence data from individual samples.

The PCR reaction mix for each individual sample was as follows:

DNA (20µl of MeDIP and 50ng for input)
10µl 5X HF buffer
1µl dNTPs (10mM)
0.5µl Phusion HotStart DNA polymerase
1.0µl Index (25µM of index 1-12)
1µl PCR primer F (25µM of indexPE primer 1.0)
1µl PCR primer R (0.5µM of indexPE primer 2.0)
+ Water to 50 µl total volume.

Cycling conditions were:

(1) 30 seconds at 98°C
(2) 18 cycles of: 10 seconds at 98°C, 30 seconds at 65°C, 30 seconds at 72°C
(3) 5 minutes at 72°C
(4) Hold at 4 – 10°C

PCR products were purified using a Qiagen purification kit, eluting in 30µl of water. The success of amplification was determined by nanodrop concentration and gel visualisation to ensure that no primer-dimers had formed.

2.3.3.2 Multiplexing Medip-seq samples

The use of a multiplexed strategy restricts the number of samples that can be analysed to 12 in one library. Without multiplexing, individual samples can be run in individual lanes on the Illumina GAIIx flowcell, enabling sequencing of a larger sample number across more than one flowcell. Individual Medip-seq libraries are prepared per sample, as for multiplexed library preparation. During the PCR amplification of Medip products, the index solution was omitted, and an extra 1uL of water is added to the PCR mix. After PCR, size selection of DNA fragments for sequencing must be performed individually for each sample, with careful size matching of case and control samples in pairs to achieve a similar distribution between groups.

2.3.3.3 Size selection of finished libraries

Gel excision: Multiplexed MeDIP-seq DNA libraries were pooled and 15 µl samples were run on a 2% agarose gel (with ethidium bromide) against a 50 bp ladder. The gel was viewed using a Dark Reader transilluminator and a gel slice in the 250–300 bp range (i.e. insert size of ~ 120 – 170 bp) was excised. A Qiagen Gel Extraction Kit was used to extract the size-selected DNA fragments and purify them, eluting in 30 µl water. At this stage, the MeDIP-seq library is ready to be quantified and sequenced, provided the MeDIP efficiency has already been determined by qPCR.

2.3.3.4 qPCR to determine Medip library efficiency

This test of Medip enrichment was performed as described above in 2.3.1.1.

2.3.3.5 Illumina GAIIx sequencing

Illumina GAIIx sequencing was performed in all Medip-seq experiments by three service laboratories at the Barts and the London Genome Centre, UCL Genomics and the Babraham Research Institute. Standard methods and protocols were used at all three sites, and the choice of different sites was determined by cost and turnaround time.

This Next-Generation Sequencing platform uses a sequencing-by-synthesis approach, coupled with massive parallel read detection to generate a large quantity of high quality sequence data. The sequencing methods are as follows:

(a) Library preparation for sequencing: specific Illumina adapters are joined at each end of the DNA fragments to be sequenced. These adapters are incorporated during the

library preparation steps described above. Library preparation may also include ligation of multiplex adapters where this approach is being used.

(b) Attaching DNA to the sequencing flowcell: single-stranded DNA is added to the sequencing flowcell via capillary tubes running across each of its 8 channels.

(c) Bridge amplification: The DNA fragments join to the flowcell at each end using the adapter sequences. These 'bridges' are then amplified by the addition of nucleotides to form double-stranded DNA, and then denatured to leave single-stranded templates anchored at one end to the flowcell surface.

(d) Cluster generation: the single-stranded DNA templates are amplified to form dense clusters of DNA ready for sequencing.

(e) Sequencing cycles: the DNA template is read one base at a time in repetitive cycles. The first cycle is initiated by the addition of DNA polymerase, primers and labelled reversible terminators that recognise each of the 4 bases in the sequence. The first base is identified by the joining of the labelled terminators that emit a fluorescent signal that is captured by a laser. The reversible binding of terminators allows these to be detached from the first base to join the second, with each base being detected by another cycle of laser-identification. Sequence data is captured as image files generated from these multiple cycles.

All DNA samples were sequenced using 36bp paired-end reads, with multiplexing. Standard sequencing chemistry and protocols were used. Sequence data output files (.txt format) were accessed via the Genome Centre or remote access servers.

Paired end sequencing involves the addition of paired sequencing primers at each end of the DNA fragment. Sequencing of the paired primers occurs from either end, sequentially. By detecting the sequence from either end of the DNA fragment, the sequence of a longer insert between the two primers can be inferred by mapping the two end reads to the genome. Paired-end sequencing also enables the detection of structural variation, e.g. copy number variation, within the sequenced fragment. All sequencing performed in the experiments described used paired-end sequencing, however the short sequence read length (36 base pairs) facilitated relatively rapid sequencing and generation of a large number of reads. The addition of an index sequence for multiplexing is performed at the PCR amplification step of library preparation. This inserts an index sequence between the adapter and sequencing primers that can be read during the sequencing protocol. These steps are summarised in Figure 2b.

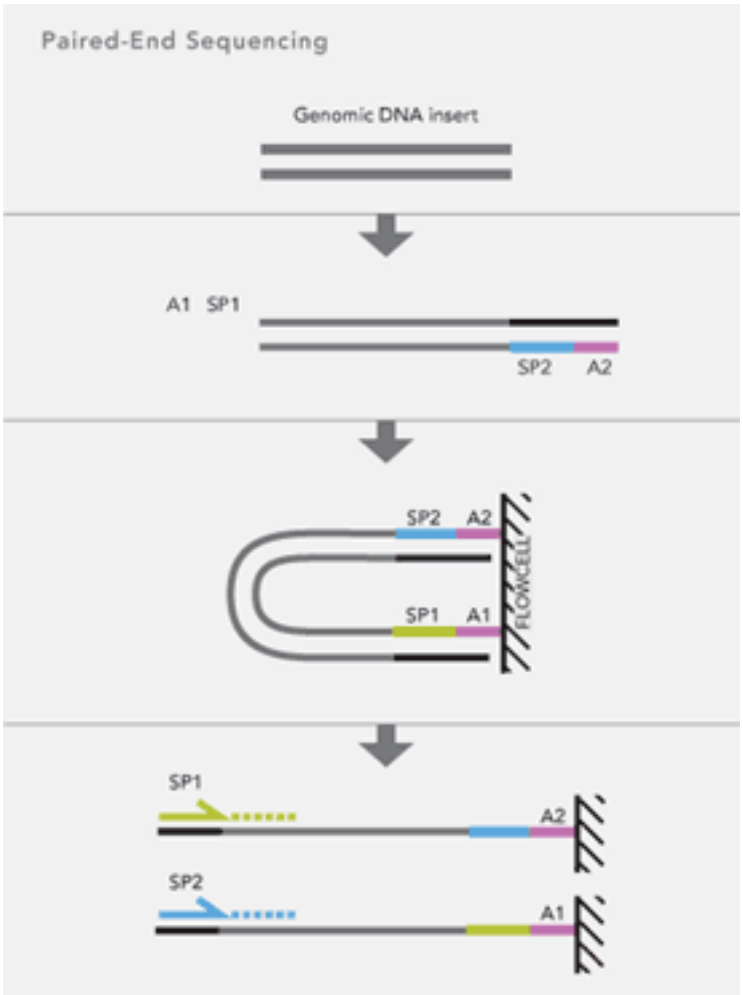


Figure 2b. Method of Illumina GAIIx paired-end sequencing using adapter fragments. Adapter fragments (A1 and A2) are ligated onto DNA fragments with sequencing primers (SP1 and SP2). DNA template multiplies into clusters using bridge amplification, and then undergoes sequencing by synthesis starting at the sequencing primers. Diagram from Illumina product brochure.

Each sequencing flowcell incorporated a PHiX control samples, derived from a bacteriophage genome. The PHiX library is a standard addition to Illumina GAIIx sequencing and is used to validate the quality of each run. Data from the PHiX control is used to generate a matrix file that is used in analysis to calculate the relative proportion of different bases in the sequenced libraries as a test of sequencing quality. PHiX control may be run in a single lane of a flowcell, or may be added as a 0.5% 'spike-in' to experimental samples; the latter obviates the need to use up a full lane of the flowcell and is therefore cheaper.

2.4 Methylation array-based experiments

2.4.1 Illumina HumanMethylation 450 BeadChip array

The Illumina HumanMethylation450 BeadChip (to be summarised hereafter as the '450k array') is a bead array-based platform designed for genome-wide DNA methylation studies in humans. This platform is recently developed and became available to researchers in 2011, and it has therefore been possible to use it to generate data from samples in the Matlab famine and Human GDM studies (chapters 5 and 7).

The 450k array is based on Illumina's Infinium assay which allows single base resolution detection of DNA methylation at pre-defined loci. Genomic DNA is first treated with sodium bisulphite, a reaction which converts unmethylated cytosines to uracil but leaves methylated cytosines intact. Following this standard conversion reaction (using Zymo EZ-methylation kits), DNA is incubated with probes, targeted to specific sites within the genome), and in two forms: one complementary to methylated DNA (M bead type), and one complementary to unmethylated DNA (U bead type). Following hybridisation, single base-pair extension of the probes occurs and a labelled ddNTP is incorporated and stained by fluorescent reagents that match either the methylated or unmethylated probes. The ratio of fluorescence from each probe is detected and used to generate a quantitative score of methylation at each probe.

Coverage of the 450k array incorporates the following genomic locations:

- All designable (approximately 99%) RefSeq genes, including promoter, 5' UTR, first exon, gene body and 3' UTR regions, without bias against those genes lacking CpG islands.

- CpG islands (approximately 96% of the total) in addition to multiple targets around these regions to incorporate CpG shores and flanking regions
- CpG sites outside of CpG islands
- Non-CpG methylated sites identified in human stem cells
- Differentially methylated sites identified in tumour versus normal (multiple forms of cancer) and across several tissue types
- miRNA promoter regions
- DNase hypersensitive sites
- Disease-associated regions identified through GWAS

Approximately 90% of the regions targeted by the 450k array were covered by its predecessor, the Illumina Human 27k Methylation array. The 27k array has been well used and results from it have been widely published (92).

Limitations of the 450k array include the lack of comprehensive coverage of Imprinting control regions (unless included in the list above) and its inability to detect hydroxymethylation due to its reliance on bisulphite conversion. In addition to these limitations, many of the array probes overlap SNPs and these can affect the efficiency of probe binding. Furthermore, the presence of SNPs in many of the probe targets could cause variable methylation patterns according to the conformation of the SNP, according to whether it creates or abrogates a CpG site. It is not yet clear how this phenomenon can be analysed in conjunction with the probes that do not include genetic variants.

2.4.1.1 Bisulphite conversion

Bisulphite conversion is the first step in the 450k array experimental protocol (see figure 2c). The bisulphite conversion reaction enables single-base discrimination of methylation status at CpG dinucleotides. In this reaction, DNA is treated with sodium bisulphite resulting in preferential deamination of unmethylated cytosines to uracil, via sulphonation, hydrolytic deamination, and desulphonation. Methylated cytosines are not converted by this reaction, and therefore a single base difference following bisulphite treatment can be detected using standard genetic sequencing to discriminate between a methylated and unmethylated CpG. This reaction is exploited in a range of targeted and epigenomic platforms designed to characterise DNA methylation status and will be discussed here in relation to the 450k methylation array and bisulphite-pyrosequencing.

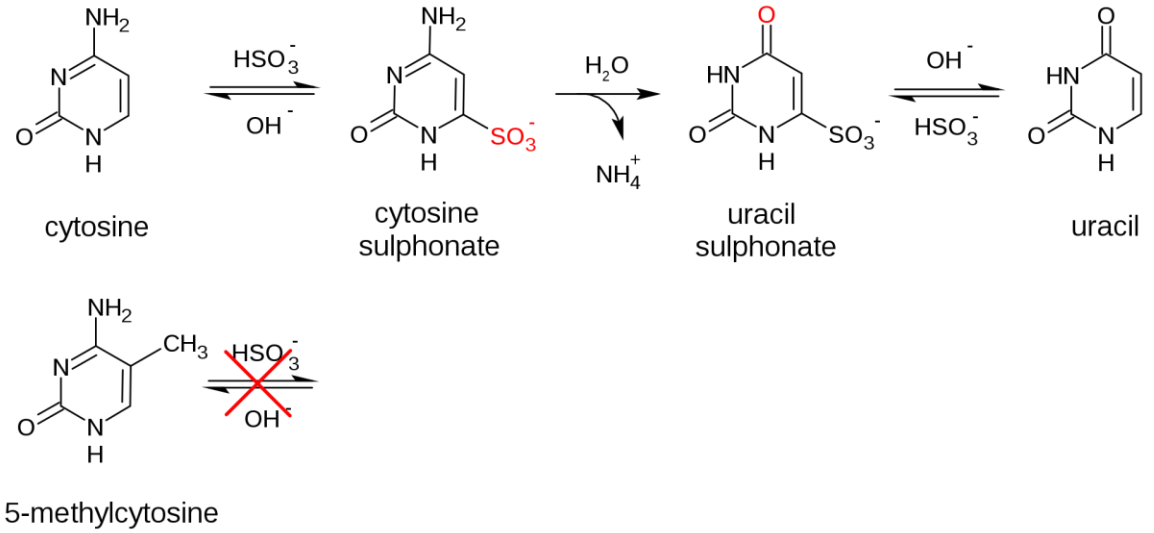


Figure 2c. Bisulphite conversion of methylated and unmethylated cytosines. Addition of sodium bisulphite (HSO^-_3) to genomic DNA enables the conversion of unmethylated cytosine to uracil through sulphonation, deamination and desulphonation. Methylated cytosines are resistant to this sulphonation and therefore do not convert to uracil, enabling downstream sequencing-based experiments to differentiate between methylated and unmethylated cytosines.

Bisulphite kits, manufactured by Zymo, have been used in the various experimental protocols to be discussed and have included both single-column and 96-well plate formats. Illumina have developed their protocol for bisulphite conversion prior to 450k array hybridisation using the standard Zymo EZ methylation kit, rather than the EZ methylation gold kit which is more routinely used (due to its considerably shorter incubation times) and has been used in later validation experiments.

The EZ methylation kit protocol requires the preparation of CT conversion reagent with 750µL water and 210 µL M-dilution buffer. DNA (200ng – 2µg) is then diluted with 5µL M-dilution buffer and water to achieve a volume of 50µL. After a short incubation at 37°C for 15 minutes, 100µL of CT conversion reagent and is combined with the DNA sample incubated at 50°C for 12-16 hours, for the bisulphite reaction to take place, followed by cooling to 4°C for 10 minutes. The sample was then added to 400µL M-Binding buffer in a spin column, inverted several times, and centrifuged at full speed for 30 seconds. 100µL wash buffer was added, and the column spun again, followed by the addition of 200µL desulphonation buffer which was allowed to wait for 15 minutes, and then spun again. A final 2 washes with 200µL wash buffer were performed, and then the sample was eluted in 10 µL of elution buffer. All samples were stored at -20°C for a maximum of 72 hours in this state or for up to 6 months at -80°C.

In experiments where large numbers of samples were being handled, the Zymo EZ methylation deep well kits for 96-well plates were used. The protocol for the plate conversion has minor differences to the single column kit, including the use of a larger volume of wash buffer (400µL) at all wash steps. The final elution step is performed in a total volume of 15µL made up with elution buffer. All centrifugation steps in this protocol were performed at 3000g for 5 minutes.

The efficiency of the bisulphite conversion using these kits is thought to be >99% when the protocol is followed correctly. At least 75% of the original DNA is recovered by the end of the conversion and cleanup. If the starting DNA concentration is low, i.e. less than the 25ng/µL required, then the difference from the required 20µL volume was made up with water.

2.4.1.1.1 Testing bisulphite conversion efficiency

Early experiments using the 450k array showed that the efficiency of the initial bisulphite conversion is crucial to the quality of the data output. As samples are processed for this array in large batches and without the use of robotics, small errors in the bisulphite conversion can easily occur and have significant consequences. Therefore, for some 450k array experiments, a QC of the bisulphite conversion was performed prior to array hybridisation. This qPCR-based QC was developed by the UCL Genomics group and was performed by Melissa Smart.

The qPCR reactions were designed against parts of the MLH and GAPDH genes (see Table 2b for primer design). The MLH primer pair is designed against the expected sequence of a region of the gene after the genetic changes brought about by successful bisulphite conversion. The GAPDH primer pair is designed to normal genomic sequence. Thus, if conversion is successful a product from the MLH1 primer pair should be seen and if it is not successful the template will not match the primer sequence and no product can result. Conversely the GAPDH primers should only produce product when conversion is NOT successful as this is when the primer sequences will properly match the template. The ratio of each of these products in a given sample can then be used to determine the conversion efficiency.

Gene	**Target**	**Primer**	**Primer sequence**
MLH1	Bisulphite converted sequence	F (5'>3')	GGAGTGAAGGAGGTTACGGGTAAGT
		R (5'>3')	AAAAACGATAAAACCCTATACCTAATCTATC
GAPDH	Unconverted sequence	F (5'>3')	CGCCCCCGGTTTCTATAAAT
		R (5'>3')	CAAAAGAAGATGCGGCTGAC

Table 2b. Primer Sequences for bisulphite conversion efficiency experiments

The bisulphite conversion has a yield of 11-13µl of converted DNA. A 2µl of this DNA is taken and diluted in 8µl water for the qPCR and the remaining DNA is set aside for the subsequent array hybridisation. A mastermix was made for qPCR reactions, containing 3.75µl H_20 (PCR grade), 1.25µl of diluted DNA, 6.5µl SYBR green mix and 1.25µl primer mix per reaction. Reactions were performed in triplicate, per primer pair, and incorporating blank wells and negative controls using unconverted DNA. A standard qPCR cycle program plus dissociation curve was used, as per that described in section 2.3.1.1.

The results of the qPCR were analysed qualitatively using the mean of triplicate Ct values for each amplicon. Where bisulphite conversion has been successful, the Ct value for GAPDH is high (reflecting a low quantity of unconverted DNA and therefore many cycles needed to amplify it) and the Ct value for MLH1 is low (as there is plenty of bisulphite converted DNA for amplification). Samples that showed less than 95% bisulphite conversion (calculated using a standard delta delta Ct calculation) were not used in array hybridisation.

2.4.1.2 Array hybridisation

The Illumina 450k methylation array uses two types of 50mer probe assays, developed using their Infinium technology. The Infinium I assay, also used by the 27k array, incorporates two beads per probe, one designed to bind to an unmethylated locus, the other to a methylated locus. In contrast, the Infinium II beadtype uses only one bead per locus, employing a single base extension step after hybridisation to differentiate converted (unmethylated) from unconverted (methylated) DNA. Where possible, Infinium II probes were used in the array design in order to maximise the possible coverage of the array.

The Infinium II chemistry uses degenerate oligonucleotides (a group of oligonucleotides designed to bind to the same genomic location but allowing for specific single base differences) designed to all possible combinations of methylation status across up to 3 CpG sites that fall within a single a probe sequence, in contrast to the Infinium I assay which assumes that all CpG sites within a single probe have the same methylation status.

Hybridisation of bisulphite-treated DNA was performed by the Genome Centre, Barts and the London, or UCL Genomics using standard protocols, and involved the following steps (also shown in figure 2d):

- Denaturation and neutralisation of bisulphite-converted DNA
- Isothermal amplification in an overnight incubation

- Enzyme-fragmentation of DNA
- DNA precipitation using isopropanol and centrifugation
- Re-suspension of DNA in hybridisation buffer and application on BeadChips (twelve samples each)
- Overnight incubation in hybridisation oven allowing for annealing to probes
- Washing of BeadChip to remove unhybridised and non-specifically bound DNA
- Single base extension of oligos using the hybridised DNA as a template, and incorporation of fluorophore labels in capillary flow-through chambers. Labels stain in one of two channels: red or green.
- Illumina HiScan detection of fluorophore labels by laser excitement of the single base extension product on each bead. The scanner collects high-resolution images of the coloured light emitted from the fluorophores as the data output.

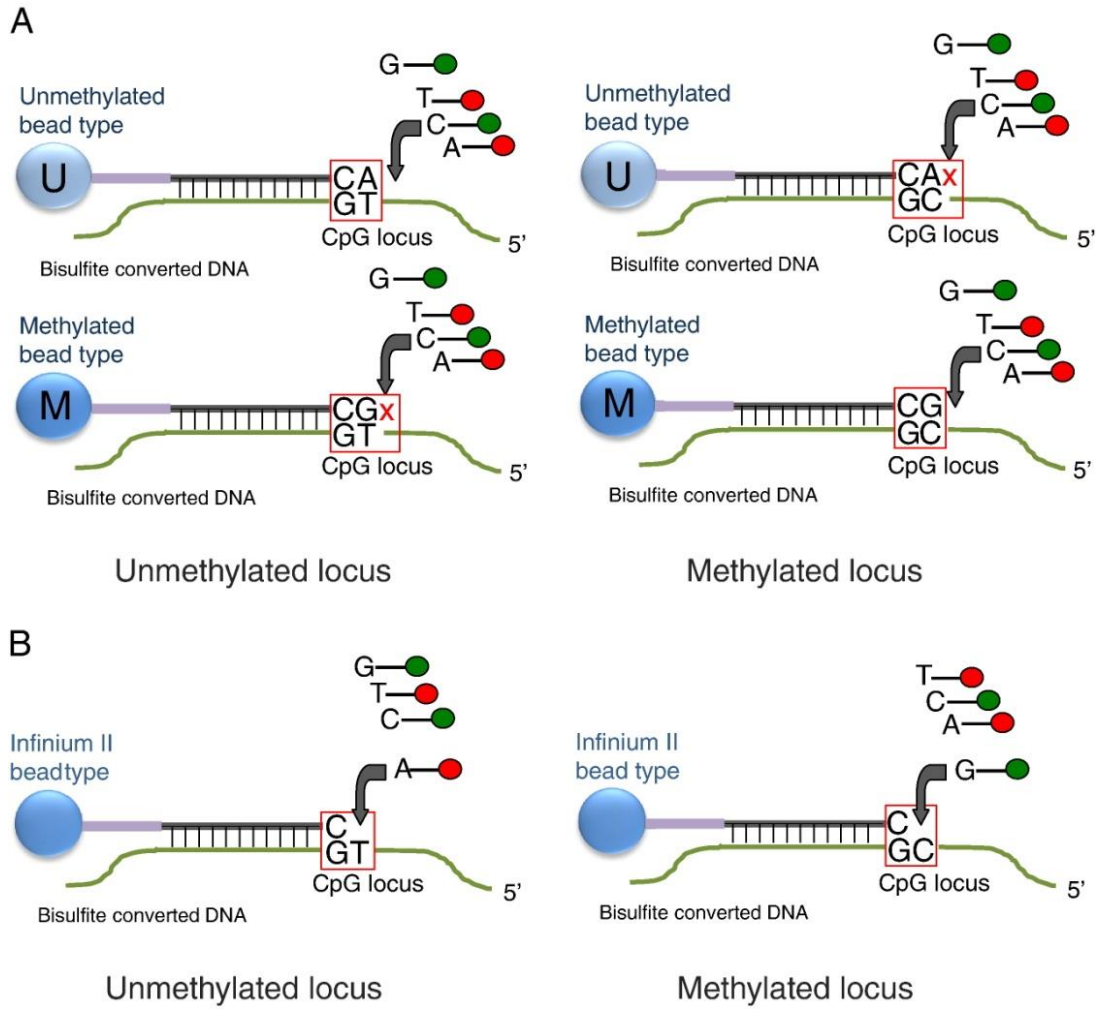


Figure 2d. Schematic of the Illumina HumanMethylation 450 BeadChip, showing the Infinium I and II bead chemistry, from Bibikova (124). The Infinium I bead chemistry is shown in (A), and show that two beads are required to differentiate between methylated and unmethylated DNA. DNA (green strands, read from 5') is hybridised to the beadchip and allows differentiation of unconverted cytosines (C) and converted cytosines (T) by the failure of single base extension beyond the mismatched primer sequence. The Infinium II bead chemistry is shown in (B) and it is seen that a single bead with a degenerate base is used to differentiate the methylation status of each individual CpG site to bind to the genomic DNA. In both beadtypes, it is the incorporation of a fluorophobe-labelled base, using single base extension that allows a colour-based detection of sequence using laser excitement from the Illumina HiScan.

2.4.1.3 Data output

Illumina-designed software is used to convert bead-level image data (.tiff files) into processed data suitable for analysis. A .idat file is produced by Illumina software and contains data from every bead on each BeadChip, summarising the data by sample, and colour channel. Conversion of raw data into .idat files incorporates some background adjustment determined by the performance of the samples in the negative and non-polymorphic array controls. These processing steps will be discussed in more detail in section 2.5.4.1. The summary data includes the number of beads (CpG sites), the mean methylation and standard deviation. The .idat files are converted into project files that can be opened in the Illumina software, GenomeStudio, or Bioconductor analysis packages, e.g. minfi, for normalisation and analysis.

2.5 Data analysis of epigenomic datasets

The broad principles underlying analysis of epigenomic data sets will be discussed here. Methods and specific details of data analysis from these epigenomic experiments will be discussed in the relevant results chapters.

2.5.1 Important considerations in the analysis of epigenomic datasets

2.5.1.1 Epigenomic profiling – what is normal?

One of the challenges in studying DNA methylation (and other epigenetic marks) on an epigenomic scale is a lack of knowledge of the basic architecture of the epigenome. General features of epigenetic marks, such as the non-random distribution of DNA methylation across the genome and its association with CpG density, are well characterised (125). However, beyond these general principles, epigenetic marks vary significantly according to many different factors, notably the tissue in which they are studied (58). Additional factors that may influence the detection of epigenetic variation include their potentially dynamic nature in different environmental conditions, relationship to gene activity and the influence of long-range genetic and structural variation that is not easily detected through standard epigenomic profiling.

Several global projects, e.g. the International Human Epigenome Consortium and NIH Epigenomics Roadmap Intitiative, have been set up in an attempt to focus resources on mapping of normal epigenomic profiles in a range of primary cell types and cell lines. It is hoped that these studies will provide important resources for researchers with which to compare their own epigenomic datasets. Such initiatives are analogous to genomic mapping projects e.g. HapMap and the 1000 Genomes Project, but also address the greater complexity of the epigenomic profile of an organism above and beyond its relatively straightforward genetic architecture.

2.5.1.2 Sample size and power calculations

The sample size required to detect epigenetic differences depends on many factors, including the size of difference expected, the amount of variation from other sources (e.g. genetic variation), and the performance of the experiment and analytic tools used to detect differences. At the current time, there is insufficient knowledge to account for all of these uncertainties and to generate accurate sample size and power calculations for study design. The majority of literature using human models to identify epigenetic variation associated with environmental exposures and/or complex phenotypes have used targeted, gene-specific epigenetic methods and the sample size calculations for these are unlikely to be able to inform discovery-based genome-wide analyses. Those genome-wide epigenomic studies that have been performed and published are mainly in the field of cancer biology where the epigenetic differences identified are often large and associated with genetic and structural variation.

2.5.1.3 Description of methylation values

Quantitative methylation values are generated by different means according to the experimental approach to their measurement. Enrichment-based methylation detection does not produce individual CpG-level quantification of methylation values. To generate a methylation level from enrichment-based experiments, it can either be estimated by computation using bioinformatic tools such as Batman, or it can be presented as 'differential methylation' in a statistical comparison of enrichment peak heights (e.g. using USeq).

The use of methylation arrays, such as the Illumina 450k array, allows direct quantification of DNA methylation at individual CpG sites. The proportion of total fluorescent signal intensity from the red and green beads in the 450k array can be converted into a beta-value to reflect

methylation. Targeted assays of methylation also give quantitative measures of DNA methylation via their PCR and sequencing-based assays.

2.5.1.4 Quality control

Early quality control of data generated from epigenomic studies is vital for subsequent analysis. One of the risks associated with handling such datasets is that their size makes it difficult to 'eyeball' data to ensure that it makes biological sense and does not introduce significant technical bias into analysis. Many genomic and epigenomic platforms incorporate control dashboards and quality control checks to enable the researcher to visualise their datasets and ensure that the technical processes have gone to plan. For example, the GenomeStudio software provided by Illumina for use with their 450k methylation array provides an inbuilt control panel dashboard where it is possible to identify technical problems with bisulphite conversion, array hybridisation and dye intensity at an early stage in analysis and using a selection of Illumina-designed control probes identical across arrays. Using such means, the researcher can make 'sanity checks' of their data prior to detailed analysis. These quality control checks include checking for expected dye intensities in array data, measures of bisulphite conversion efficiency, and read counts generated per flowcell lane during Next Generation Sequencing. Such quality control checks and potential technical biases will be discussed in more detail in due course.

2.5.1.5 Normalisation strategies

When creating large datasets from different experimental processes, e.g. different arrays or flowcells, it is important to consider how technical biases may affect the distribution of data values and downstream analysis. Normalisation uses statistical techniques to make the distribution of data from two different datasets the same. Normalisation therefore makes the assumption that different data distributions are due to technical effects, e.g. between-array differences. Using normalisation without first understanding any quality control issues could therefore mask any true differences due to technical problems and result in drawing incorrect conclusions from a dataset.

The normalisation strategies that are commonly used in analysing genomic data include quantile and Loess normalisation. Quantile normalisation uses a reference or standard distribution (the latter being based on the appropriate distribution for the data, e.g. Gaussian or Poisson) to compare the distribution of test data. Differences between the standard and

test distributions are ranked and used to correct towards the standard distribution. Loess normalisation uses multiple regression to fit multiple regression lines to a dataset. Overlap between these regression lines are then smoothed out across the entire dataset to produce a normalised set of values.

2.5.1.6 Problems of multiple testing

Another important consideration when analysing epigenomic datasets is the potential problems that arise from multiple hypothesis testing. This important statistical consideration relates to the increasing number of inferences and assumptions that are made within a dataset as more data points are compared with each other. For example, the more genomic regions that are included within an epigenomic study comparing cases and controls, the more likely the study is to show a difference between cases and controls. This incorrect assertion occurs as the more statistical comparisons that take place, the more likely they are to fall outside of the standard 95% confidence intervals or accepted hypothesis test cut-offs. (e.g. $p<0.05$). Failure to account for these factors when analysing large datasets will result in false positives being accepted.

These statistical concerns necessitate a more stringent approach to analysing epigenomic datasets. One widely accepted approach in both epigenomic and genomic studies is to use a false discovery rate (FDR, and also known as the Benjamini and Hochberg FDR) as a means of controlling and defining the expected *proportion* of false positives (i.e. incorrectly rejected null hypotheses or type I errors). The FDR states the proportion of false positives in a series of multiple comparisons is determined by the observed p value distribution of these comparisons, itself determined by the consistency of the dataset (126). An adjusted p value (or q value) is generated from the distribution of p values in an experiment, and defines the expected proportion of significant tests that will result in false positives. For example, a p value of 0.05 explains that 5% of all tests will result in false positives, but a q value describes that 5% of significant tests will occur due to false positives. An FDR threshold can be set from the q value, according to what is deemed appropriate. Unlike p values, there is no one single accepted FDR threshold in studies that incorporate multiple comparisons.

4 factors determine the FDR characteristics:

- The proportion of differentially methylated regions (DMRs) in relation to those regions that are not differentially methylated (this proportion usually increases, the more probes you have)

- The size of true differences
- Measurement of variability
- Sample size

An FDR of 0.1 means that of 100 DMRs detected, a maximum of 10 of them would be expected to be false positives. If too stringent an FDR is set, e.g. 0.05, it may reduce the false positives, but it will also reduce the ability to identify true differences (i.e. reduce the sensitivity). This is avoided by understanding the sensitivity (or false negative rate) of the dataset.

The FDR is a less conservative means of controlling for multiple comparisons than the more stringent approach using a Bonferroni correction (127). The latter is a good means of reducing false positives, but at the risk of losing true discoveries in a dataset. For this reason, the FDR is now a widely accepted tool to control for multiple hypothesis testing in epigenomic studies.

2.5.1.7 Identification of methylation variation

The amount of methylation difference that is 'significant' is commonly discussed amongst epigenetic researchers. Studies of cancer show large epigenetic differences associated with genomic rearrangements (128). However, epigenetic studies looking at environmentally induced epigenetic variation, with or without genetic susceptibility, identify methylation differences that are far smaller (89). Quantitative differences in methylation may exert their significance not only by order of magnitude, rather by the conformational changes in higher order DNA and protein structure exerted by DNA methylation. In contrast, SNP-associated variation will reflect the presence or absence of a CpG site and therefore the ability or inability to methylate, resulting in variation of 0, 50 and 100% according to the conformation of the SNP.

Identification of variants in DNA methylation between study groups are defined as either differentially methylated regions ("DMRs") or methylation variable positions ("MVPs"). The description of a DMR is applied to experiments such as Medip-chip or Medip-seq where quantitative methylation estimates are made across small regions (e.g. bioinformatically generated windows) but do not have single base resolution. In contrast, experiments that identify single base methylation differences, such as targeted bisulphite PCR or pyrosequencing, or methylation array experiments, allow the definition of an MVP as the genomic position is defined.

2.5.1.8 Integration of epigenetic and genetic data

Epigenomic datasets can be aligned to genomic data via sequence alignment to a reference genome. This epigenomic and genomic integration is performed early in data analysis pipelines and makes use of single- or paired-end sequencing methods to facilitate alignment. In contrast, array-based methylation experiments (e.g. Medip-chip, methylation arrays) have defined genomic targets incorporated into their probe design, enabling easy integration of the epigenetic and genetic data.

Custom scripts used in analysis incorporate annotation of genetic regions, including genomic position, gene names and local architecture. Scripts also allow integration of additional data sources, such as that from dbSNP.

Characterisation of DMRs and MVPs is complemented by their addition to custom views of commonly used genome browsers, such as Ensembl and UCSC. Unless otherwise specified, the Ensembl genome browser is used in the experiments presented, and the latest human genome assembly (GRCh37/hg19) was used in all analyses.

2.5.1.9 Validation and replication

It is imperative that the findings of epigenomic studies are well validated and replicated. Technical validation (repeating an experiment using a different experimental platform) is an important means by which experimental biases are excluded. Replication of studies in a larger sample size is a useful determinant of biological plausibility and can also be used to address wider questions such as the effects of sample heterogeneity, epigenetic-phenotypic associations and the influence of genetic factors. Validation and replication may be best achieved through high-throughput targeted experimental approaches. This will be discussed in more detail in the relevant chapters.

2.5.2 Analysis of Medip-chip

The principles of analysing Medip-chip data (after normalisation) include the use of a custom bioinformatic algorithm to address the potential confounding effects of CpG density on Medip enrichment, and therefore methylation estimates. Different bioinformatic algorithms exist to overcome this potential confounding; Batman is the most widely used and was developed by Rakyan and Down (58), and MEDME (129) is an alternative. Batman uses Bayesian statistics to make an estimation of absolute methylation across the genome, according to knowledge of

CpG density across the genome. The algorithm calculates a coupling factor between the CpG sites within the targeted array probe and the fraction of DNA molecules that hybridise to it. The sum of coupling factors within a probe gives the local CpG density, and this in turn is put into a statistical model that uses Bayesian probabilities to estimate the likelihood of a given methylation score (as determined by a normalised log2-array signal). To reduce the amount of computation required, the model makes an assumption that the methylation of adjacent CpGs is highly correlated across several hundred base pairs. In this paper, Down and Rakyan validated the Batman algorithm against bisulphite-PCR based quantitative methylation analysis from the Human Epigenome Project.

The estimates of absolute methylation across the targeted regions in different samples were compared using a 'sliding window' approach. Standard statistical tests, e.g. Kruskal Wallis, were then applied to identify differences in absolute methylation estimates between the experimental groups. The statistical analysis of Medip-chip data that has gone through the Batman algorithm, will be discussed in more detail in Chapter 3.

2.5.3 Analysis of Medip-seq

Similar principles were applied to the analysis of Medip-seq data and these will be discussed in more detail in Chapter 4. As with Medip-chip experiments, custom bioinformatic algorithms were used to analyse data, this time coming from Next Generation Sequencing. The volume of data generated from these experiments makes it imperative to have detailed and careful quality control, normalisation and sanity checking of data before analysis, and this will be explained in detail in the relevant chapter.

Quantitative methylation estimates are generated by mapping the sequence reads (i.e. methylated DNA) to the genome and counting the read density at different areas using size windows. The greater the read density, the higher the methylation level is estimated to be. In contrast, if a region is not methylated, it will not be contained in the enriched Medip-fraction of DNA and it will therefore not be sequenced and counted as a sequence read.

2.5.3.1 Quality control

Quality control of sequenced Medip-seq libraries takes several steps. Either side of each step, a read count is calculated per sample or multiplex tag. This read count is used to check the quality of data through the pipeline by following the predicted read loss per step that is similar

to other experiments and is similar between samples. The following steps (using custom and standard scripts) were included in the bioinformatic pipeline by way of quality control:

1. Paired end read formation: the reads from libraries that have sequenced using paired end chemistry are paired up using
2. Mapping of reads: reads were mapped to the reference genome using BWA, a standard mapping algorithm that converts reads into SAM files. This file format is common to many next generation sequencing analysis pipelines and is suitable for the subsequent steps in the QC and bioinformatic processes.
3. Removal of duplicate reads: this was performed to remove polyclonal reads which occasionally arise due to the PCR amplification step in library preparation,
4. Q10 filter: this removes reads that map poorly to the genome, using a standard Maq quality score of <10% as a cutoff.
5. Calibration plots: Custom bioinformatic scripts designed as part of the Batman algorithm are used to generate plots of read density against methylation to identify whether there is appropriate genomic coverage of the mapped reads.

2.5.3.2 DMR calling strategies

Several different approaches exist to identify regions of differential methylation from Medip-seq data. These approaches have not been extensively used nor compared and a combination of different strategies was applied to analysis of Medip-seq datasets.

2.5.3.2.1 Batman algorithm

Batman has been used to generate a 'methylome' map from Medip-seq data. However, since this data was published, concern has been raised (personal communication by T Down and V Rakyan) that the adjustments for CpG density that Batman makes may not be sufficient to represent accurately methylation differences in whole genome experiments. Batman employs a coupling factor to account for variation in CpG density corrects for CpG density but, in combination the enrichment bias of Medip (in which regions of low CpG density are over-represented), there may be over-correction of regions of low CpG density and under-correction of regions of high CpG density. Instead, it was advised that Medip-seq data should be analysed with algorithms that do not correct for CpG density.

2.5.3.2.2 Thomas Down DMR caller

Following the criticism of the Batman DMR caller, it was decided that a more simplistic approach should be taken to analysing Medip-seq data. Thomas Down therefore developed a simple and unbiased method of detecting methylation variation using minimal computation that could also offer the ability to perform pair-wise and group-wise comparisons. This algorithm uses Java-based custom scripts written by Thomas Down in Perl.

Following standard quality control checks on Medip-seq data, the 'Thomas Down DMR caller' implements size selection of fragments and fragment size normalisation in the analysis. This cautious approach reduces the chance of fragment size affecting the amount of Medip enrichment. Subsequently, calibration plots are generated (using Java-based scripts) in the Batman algorithm to identify patterns of enrichment between samples. Following this data processing, custom Java-based scripts, are used to move rolling windows across the genomic datasets and compare read counts between each other, in a group-wise or pair-wise arrangement, using simple t-test based statistics to identify regions of differential methylation enrichment. Group-wise analyses can be performed between experimental groups, e.g. cases and controls. Pair-wise analysis using this DMR caller can be applied to case-control pairs, and can allow further normalisation of data, e.g. by matching sample pairs according to Medip-enrichment efficiency.

The simplistic nature of this analytical approach confers a high chance of detecting false positive results and sensitivity being low. It can be discussed that the approach, with minimal computation based on assumptions of normal and abnormal methylation architecture, has a high specificity, or ability to detect true negative results.

2.5.3.2.3 USeq

Useq is a custom Java and R-based bioinformatic algorithm, originally designed to analyse RNA-seq and ChIP-seq data, which calculates the statistical significance of differential enrichment from RNA-seq and ChIP-seq experiments (130). USeq uses the principles of a binomial distribution to minimise the potential biases when deriving robust estimates of difference when enrichment-based experiments are coupled with Next Generation Sequencing. In their paper, Nix and Boucher discuss that the process of mapping multiple short sequence reads of unenriched DNA to the genome does not produce a random distribution of alignment (due to many technical biases in the generation of sequencing libraries). They describe that existing bioinformatic algorithms do not account for localised

systematic bias when calling differences in ChIP-seq peaks and proposed a new model to address this. This bias, as well as the difficulty of multiple hypothesis testing when handling any genomic datasets, led the authors to test a range of methods with which to identify differences in ChIP-seq peaks. Peak identification methods were compared using a control dataset and a defined sliding window approach to interrogate genomic regions. The true positive rate (TPR) and FDR were calculated with the presence of spike-in data to enable subsequent comparison of peak identification methods. Four methods of peak identification were compared, using simple comparisons of peak heights and more complex statistical interpretations. The performance of each method was evaluated using their TPR and FDR and it was concluded that the calculation of a binomial p value for binding peaks outperformed other methods.

A binomial test is used when trying to differentiate between the likelihood of two dichotomous variables occurring in an experiment with multiple testing. It is commonly applied to a situation where one variable is more or less likely to occur than the other. Examples of where this distribution could be applied are in experiments that are trying to differentiate between success and failure, or observed versus expected in a series of repeated measurements or settings. This statistical test allows for multiple observations to be compared and this can incorporate an experimental design where biological replicates are used, such as in the experiments to be presented in subsequent chapters. A negative binomial test is used in similar experimental contexts, but when trying to decide the number of tests that would be required to differentiate the two dichotomous variables.

The binomial test p values generated by USeq are calculated in 500bp windows that are applied to the mid-point of each paired-end read. The USeq algorithm also generates a 'best window' of x bp within each 500 bp window to further localise the region that is driving the most differential enrichment. At each window, an FDR is generated using Storey's q value (131). The q value is a measure of significance in terms of a false discovery rate, in contrast to the p value, a measure of significance in relation to false positive rates. The use of the q value enables more stringency in the selection of differential enrichment and reduces the number of false positive results.

The conversion of window p values to q values assumes that input DNA (i.e. unenriched) p-values would be distributed uniformly across analysis windows; as input DNA is not used in most experiments (including the ones presented here), the potential bias from this assumption controlled by filtering out windows with a low read density (<10 reads). In doing filtering out these windows, the FDR can be overestimated and therefore the sensitivity of small datasets can be reduced.

Although designed for ChIP-seq and RNA-seq, this peak detection method can be applied to other enrichment-based experiments. Researchers, e.g. Vining et al, have begun to use this statistical approach (132) to determine statistical significance of differences in peak distributions derived from Medip-seq.

2.5.3.2.4 Combined Thomas Down + USeq DMR callers

As discussed, the principles behind the Thomas Down and USeq DMR callers suggest that they offer a more robust means of calling differential methylation in Medip-seq datasets compared to Batman. In the absence of a bioinformatic algorithm for Medip-seq datasets that has been validated, the results from any of the current analytic tools should be interpreted with caution. The simple means by which the Thomas Down caller identifies differential methylation in groups or pairs with no preceding probabilistic judgment about the likelihood of differential methylation results in a low sensitivity and the over-calling of DMRs. In contrast, USeq takes an approach to identifying differential methylation in which correction for multiple testing is incorporated using binomial tests of probability. This approach is also better suited to experiments that incorporate biological replicates. FDR and p-value based statistics are generated in USeq and whilst this makes it a more stringent approach than the Thomas Down caller, it does increase the likelihood of rejecting true positives. It was therefore decided that both the Thomas Down and USeq DMR callers should be used in Medip-seq analysis, with the first used as a crude test of differential methylation with no a priori hypothesis, and the second to control for false discovery. By then collating a list of DMRs that are identified in both DMR callers, the chances of over-computation and false discovery would be minimised and provide a list of 'top hits' for validation.

2.5.3.3 Gene ontology analysis

Identification of gene ontology (GO) (133) was performed using the GOStats package. The list of Medip-seq DMRs used to generate GO terms will be described in the relevant results chapters.

2.5.4 Analysis of 450k methylation array

The projects using the 450k methylation array were analysed using the same pipeline, explained in this methods section. The custom methods described have been developed by R. Lowe, C. Mathews, G. Carbajosa with V. Rakyan.

2.5.4.1 Quality control checks

As described earlier, the data from the different bead types (Infinium I and II) in the 450k platform are summarised in .idat files generated by Illumina's custom software, GenomeStudio. Genome Studio and the minfi R package for analysis of 450k data perform a series of quality control checks on each step of the array processing using specially designed probes. These steps include sample-independent QCs: bead-staining with biotin (green) and DNP (red) of DNA; extension efficiency of nucleotide sequences to the probes; hybridisation efficiency and subsequent target removal. Sample dependent QCs include a test of the bisulphite conversion efficiency and the specificity of probe binding and primer extension. Due to the different functionality of the two Infinium probes (I and II), these QCs are performed with each probe type separately. Negative control probes are also included in the array design to define the background signal from non-specific primer extension and cross hybridisation, subsequently incorporated into data processing from raw signal to .idat files. Finally, non-polymorphic regions of the bisulphite-converted genome are used as a control to test assay performance by designing A, C, T and G probes within these non-polymorphic regions that can be compared across different samples.

2.5.4.2 Normalisation

Data is normalised using standard quantile normalisation methods. Where samples have been run in different batches, quantile normalisation is performed per batch.

2.5.4.3 Data filtering

Probes are filtered as follows:

(a) X and Y chromosome probes were removed in all studies to prevent differential methylation being called from different numbers of X and Y chromosomes in mixed sex experimental groups.

(b) Cross-hybridising probes were removed. Cross hybridisation can occur on all bead arrays due to regions of sequence homology. Specific probes susceptible to this cross hybridisation have been identified and are easily filtered out of downstream analysis.

(c) Detection P value. A p value is generated for each probe on the beadarray, describing the variance in intensity generated from the beads per probe. A high level of intensity variation implies that there is signal noise and the only probes with a p value of <0.01 are included in analysis.

(d) SNP-containing probes. 40,484 of the probes on the 450k array overlap SNPs, and it is unclear to many researchers whether this may affect the binding and therefore intensity generated by these probes, or whether they may be used to describe methylation variation in association with SNPs. In our analyses, SNP-containing probes have been identified from the 1000 genomes project data. Our analyses have been performed with and without SNP-containing probes, and this will be discussed in more detail in the relevant sections.

2.5.4.4 Methylation values

Methylation is quantified by the dye intensity emitted by methylated vs. unmethylated beads in the 450k array and is converted into beta-values of 0 (unmethylated) to 1 (fully methylated). Beta values are the most commonly used means of quantifying methylation and are used in a range of plots to start characterising data from the 450k array. Plots of beta values can also be used to sanity check data from the array to ensure that data is following the appropriate distribution of methylation values at different genomic locations.

Unsupervised cluster plots or multi-dimensional scaling (MDS) plots are frequently used to identify whether samples differ from each other in terms of beta values generated from the 450k array. The dimensions on these plots (shown on the X and Y axes) represent a calculation based on the number of probes examined and the number of samples; comparison of MDS plots require both of these variables to be constant if a quantitative estimate of difference between plots is going to be made However, clustering of samples in such hierarchical models can identify technical differences, e.g. in the case of a batch effect which clusters out from other samples, and also biological differences, and qualitative and visual assessment of these differences are usually sufficient.

Transforming beta values into an 'M value' is thought to be a more statistically sound means of representing methylation values than the beta value (134) (see figure 2e). The M value equates to the log2 ratio of the intensities of the methylated vs unmethylated probes and is

thought to account for the bimodal distribution of methylation across the genome more accurately. The variability, or so-called 'heteroscedasticity', of beta values outside of the 0.2-0.8 range prevents the appropriate use of statistical tests that rely on a Gaussian distribution of data, e.g. the t-test. In contrast, the logit transformation of beta values into M values, allows the application of statistical tests that rely on a Gaussian distribution and therefore makes the identification of differential methylation more robust.

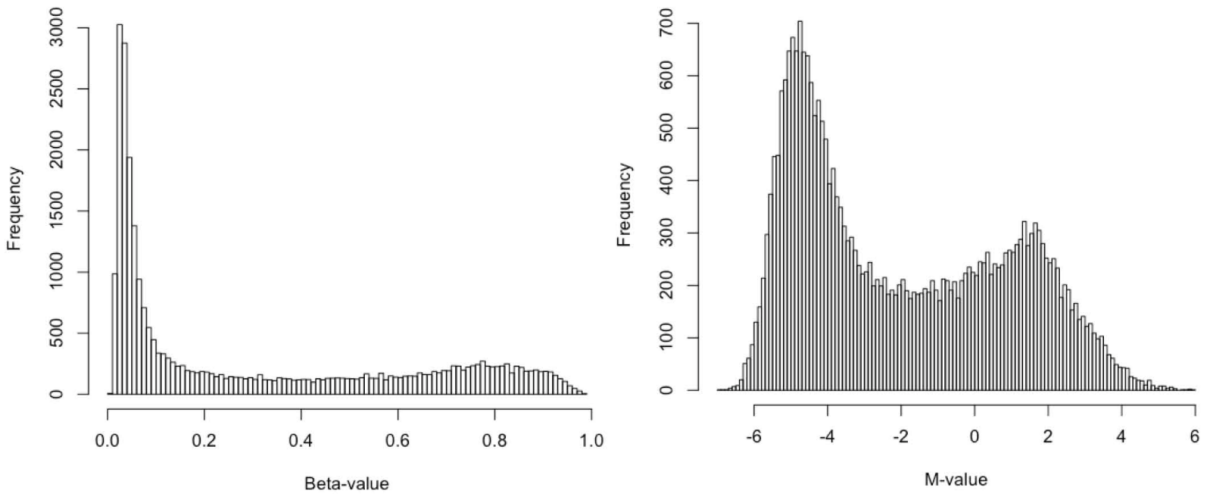


Figure 2e. Frequency histograms of beta-value (left) and M-value (right) derived from interrogating 27578 CpG sites using the Illumina 27k array. From Du et al (134).

2.5.4.5 MVP calls

Numerous techniques for analysing differences in methylation from 450k array data, mostly relying on standard statistical tests for comparing differences in mean levels of methylation at single CpG probes. The R-based Limma package was used to determine differential methylation from the 450k array, after the steps above had been carried out using Minfi.

Methylation variable positions (MVPs) are identified from comparing the methylation values (Beta- or M-values) different experimental groups and are the primary outcome from the 450k array studies presented. Parametric statistical tests that rely on normally distributed data can be used to identify MVPs after the careful processing of data outlined above.

When comparing two study groups, a t-test can be used to identify MVPs, and when comparing more than two groups, an F-test has been used. The F-test is a type of ANOVA (analysis of variance) that can identify differences between several groups without any pre-

conceived expectation of how the groups relate to each other. This test has been used to detect variation in gene expression resulting from different environmental conditions (135).

T-test and F-based statistics generate p values that are used, in the standard way, to determine the degree of significance between means. However, as these array experiments incorporate multiple testing (i.e. interrogation of each array probe is one test), a standard p value is not sufficient to control for the likelihood of false positives. A correction for multiple testing should be performed, either as an FDR-adjusted p value (or q value), or by the use of a Bonferroni correction, as described in section 2.5.1.6.

2.5.4.6 Sensitivity analysis

Sensitivity analyses are used to assess how one variable (e.g. sample within a batch of samples) controls the results of a multivariate model. Application of a sensitivity analysis to the MVPs produced in a 450k dataset allows the identification of MVPs that are driven by differential methylation in just one sample in its experimental group. Such MVPs are likely to be false positives and by excluding them from the analysis, it is possible to have greater confidence in the remaining MVPs being true positives. Although not as robust as using methods such as a false discovery rate, this approach can be useful in datasets that are not powered well enough for FDR correction.

The leave-one-out-cross-validation (LOOCV) analysis can be used as a means of performing a sensitivity analysis. In LOOCV, the dataset is re-analysed in multiple different iterations and each re-analysis leaves one successive sample out of the analysis. This identifies the MVPs driven by a large difference in methylation in just one sample, thereby allowing their exclusion from further analyses.

2.5.4.7 Identification of SNP-associated methylation variation

At this stage of analysis, it is feasible to visualise the MVPs generated from the MVP call with a simple histogram of methylation level (beta value) per sample at specific probes. Some MVPs display beta values of around 0%, 50% or 100% in different samples, consistent with the presence of zero, one or two copies of a SNP that creates or abrogates a CpG site capable of methylation. Although not conclusive, this pattern of methylation is a strong indication that this is one of the 40,000 probes overlying SNPs (in which case it can be identified from Illumina software) or that it may overlie an unknown SNP or be in *cis* with a nearby SNP.

2.5.4.8 Gene ontology analysis

Identification of gene ontology was performed using GO analysis tools that interface directly with the Limma R package. The GOstats package enables the creation of a list of GO terms from the MVPs identified in analysis. In all cases, the MVP list after sensitivity analysis was used to generate GO terms.

2.6 Validation experiments

2.6.1 Bisulphite pyrosequencing

2.6.1.1 Bisulphite conversion

Bisulphite conversion of whole blood samples was performed by C. Bell at UCL, and hypothalamus/brain samples by S. Finer. A standard bisulphite conversion kit (Zymo EZ Methylation Gold) was used in which 130µL of CT conversion reagent was combined with 20 µL DNA sample and incubated at 98^{o}C for 10 minutes, 64^{o}C for 2.5hours and then held at 4^{o}C. The sample was then added to 600µL M-Binding buffer in a spin column, inverted several times, and centrifuged at full speed for 30 seconds. 100µL wash buffer was added, and the column spun again, followed by the addition of 200µL desulphonation buffer that was allowed to wait for 15 minutes, and then spun again. A final 2 washes with 200µL wash buffer were performed, and then the sample was eluted in 10µL of elution buffer. This conversion reaction was performed for all samples individually, and the DNA was subsequently amplified using primers targeted to the Batman-derived windows of interest designed using Biotage PSQ assay design software 1.06 (© Biotage). Primer sequences are described in table 2c. Each sample was run on a 2% agarose gel to ensure amplification, using a PCR ladder to check the product size.

	Window 9, position 11-15:	**Window 9, position 12-15**	**Window 12**
Forward	GTAGATTAGGTGGGTTAAATGATA	AAATAATTCAATATCTTTAA	GGTAGGAGAATTGTTTGAAT
Reverse	TAAATTTACTCTTCCATAAACCA	AATATATGAGAAAGAAGGTAAATG	GGAGGTTGAGGTAGGAGAATTGT
Sequence	GATATTATAGTTAAATTAAG	CTAAATTTACTCTTCCATAAACCA	ATCCACACTCTTTATCATAACTTT

Table 2c. Pyrosequencing primers used to validate *FTO* 900bp window

2.6.1.2 Pyrosequencing

Pyrosequencing was performed on the amplified DNA, using the additional sequencing primer, using a Biotage PSQ HS96 instrument at UCL and at the Genome Centre, as per manufacturers instructions. SNPs within the regions sequenced were represented by their IUPAC code to enable sequencing and estimation of methylation at genetic variants.

Chapter 3. Type 2 Diabetes Study

3.1 Introduction

Recent genetic advances in the study of type 2 diabetes and related disorders have uncovered many susceptibility loci using a genome-wide association study (GWAS) approach. Despite these important advances, the common genetic variants identified explain <10% of the heritable component of disease, and little is known about their contribution to aetiology. A range of different molecular approaches is now attempting to explain the 'missing heritability' in complex diseases and address the mechanisms whereby genetic variants exert their influence in disease aetiology and pathogenesis. Some researchers are trying to identify rare and structural gene variants as well as how discovery SNPs and causal variants interact (136). Complimentary to this approach is to identify whether stable epigenetic modifications interact with genetic variants and whether they may, together, confer additional risk or pinpoint those genetic variants with functional relevance. This study has been designed to identify whether these genetic-epigenetic interactions exist at regions of known genetic susceptibility to type 2 diabetes. It incorporates contemporary knowledge of genetic susceptibility variants identified through GWAS, to a targeted study of DNA methylation. At the time this study was designed, around 30 common variant loci had been associated with T2D (137)(138). Preliminary evidence implicates epigenetic factors in the aetiology and pathogenesis of diabetes, and in some studies has shown an association with genetic susceptibility variants (103)(102). However, a genome-wide approach to identifying epigenetic-genetic interactions had not previously been taken in Type 2 diabetes when this study was performed. Given the complexity of the epigenomic landscape and its variable association with genetic features (described above), a genome-wide approach offers an unbiased means with which to interrogate the genome and identify these possible genetic-epigenetic potential.

With this broad principle in mind, this study has been designed to test the following hypotheses.

a) Variation in DNA methylation occurs at regions of genetic susceptibility to Type 2 diabetes and can be detected via a Medip-chip approach
b) Detection of epigenetic variants, and their tissue-specificity, may provide a route to identify aetiological variants identified through GWAS.

3.2 Methods

3.2.1 Sample identification and preparation

Whole blood genomic DNA from two pre-existing population studies was used. Forty UK Warren 2 T2D female participants selected from trios and had a diagnosis of Type 2 diabetes made by either current prescription of a diabetes-specific medication or laboratory evidence of hyperglycemia (World Health Organization definition). Forty female participants without diabetes were selected from the Exeter Family Study of Childhood Health and had normal fasting glucose and/or HbA1c levels. Informed consent was obtained from all participants. Hypothalamus and prefrontal cortex samples were selected from 'normal brain' post-mortem tissues from the MRC London Brainbank for Neurodegenerative Diseases, and DNA was extracted using standard phenol:chloroform techniques. Hypothalamus and prefrontal cortex tissue was selected as relevant tissues to obesity pathogenesis as well as tissues where a target gene identified in this study (see section 3.3.4) is highly expressed.

	Warren 2 T2D	**Exeter Family Study of Childhood Health**	**MRC London Brainbank for Neurodegenerative Diseases**
Number of samples	40	40	14
Disease status	Type 2 diabetes	No diabetes	'Healthy brain'
DNA source	Whole blood	Whole blood	4 hypothalamus, 10 prefrontal cortex
Mean age (range)	41 (27-56)	36 (24-44)	55 (25-96)
Sex	All female	All female	8 male, 6 female
Mean body mass index (kg/m^2) (range)	34 (23-49)	22 (15-36)	Not known

Table 3a. Source of samples and participant characteristics

Whole blood DNA, previously extracted and stored was used. Each sample was sonicated individually for 30 minutes in the first instance, using the same starting concentration and volume of DNA in solution (1.5µg in 100 µL) to minimise variation in sonication. Each sonicated sample was visualised on an agarose gel to ensure that the fragment size was between 200 and 700bp, and further sonication (in 5 minute cycles) was performed where necessary. These steps were important to achieve similar Medip efficiency between samples, and therefore to eliminate bias that could be introduced through variation in antibody enrichment between samples. Samples were cleaned up after sonication using a Qiagen QIAquick PCR purification kit.

3.2.2 Medip and whole genome amplification

The Medip protocol was optimised using test samples of DNA and different concentrations of 5-methylcytidine antibody. The quantity of antibody for optimal Medip enrichment, as determined by qPCR, was found to be 2.5ug. DNA concentrations were checked after Medip (usually between 5-10ng/uL) and after whole genome amplification (60-80ng/uL) as a preliminary means of determining their success. Whole genome amplification (WGA) was performed on Medip samples in duplicate, in order to obtain sufficient DNA quantity, but this was not required for input samples. Duplicate amplified Medip samples were pooled for all subsequent analyses.

3.2.3 qPCR

After whole genome amplification, Medip and input samples underwent a qPCR analysis to determine the success of Medip enrichment (see 2.3.1.1). Successful MeDIP enrichment was determined by a higher relative Ct value for MeDIP vs input sample at the unmethylated locus.

3.2.4 Array design

The Type 2 diabetes study used a custom-designed Nimblegen array, comprising 387,835 50-75mer probes divided into 122 regions covering 37,037,978 bases of the genome (080314_HG18_PA_Tiling), designed by C. Lindgren (Oxford). At the time this study was designed and performed, this array offered the most accessible platform with which to assay DNA methylation on a genome-wide scale; it has since been superseded by several platforms with much greater coverage. The genomic regions targeted on the array were all contemporaneously known T2D association loci (20 regions, covering the 60kb around the gene or LD block containing the association SNP) and monogenic obesity & diabetes (13 genes). In addition, the T2D chromosome 1q linkage region (1q21-24: 148.4-171.3Mb) comprising 22.8Mb identified in multi-continental populations and all known imprinted genes and imprinting control regions were covered on the array. Imprinted genes and control regions were selected from Imprinted Gene and Parent-Of-Origin Effect Database (www.otago.ac.nz/IGC) with 108 regions containing either the gene or control region +/- 5000 bp (139). These 108 regions included all those described in humans as well as regions syntenic to mouse imprinted loci that have not yet been confirmed in humans. Finally, a miscellaneous group of nine loci including Coronary Artery Disease and Stature, hyperuricaemia association regions and the *PPARGCA1* gene (described in (110)) were included. All coordinates are for genome build NCBI Hs36.1/ HG18.

3.2.5 Array hybridisation

Medip and input samples were labelled in a two-colour protocol, using Cy5 and Cy3 respectively, and co-hybridised to a single array. Colour intensity from Cy5 versus Cy3 was used to determine methylation levels in subsequent bioinformatic analysis.

3.2.6 Array normalisation

Intensity and Spatial Plots were drawn. Lowess normalisation was used in order to smooth out variation that arises due non-linear responses during labelling, hybridisation, or scanning to the two different dye colours in the data. Quantile Normalisation was performed to correct for variation in probe level intensities between arrays. The normalised and raw data are available from GEO (Gene Expression Omnibus, NCBI) under the accession number GSE20553.

3.2.7 Medip-chip bioinformatic analysis

C. Bell, G. Lewis and A. Teschendorff at University College London performed Bioinformatic analysis for this project. All genome co-ordinates for this study are given for NCBI Build Hs36.1/UCSC hg18. Scripts for analysis were written in Perl.

3.2.7.1 Estimation of Absolute Methylation

It is assumed that the array Medip/input (Cy5/Cy3) intensity is proportional to the density of methylated CpGs. The Bayesian algorithm, Batman, was used and has been described in Chapter 2 as a means of identifying differentially methylated regions (DMRs) from MeDIP-chip or MeDIP-seq experiments. The algorithm uses Bayesian theory to estimate methylation in the context of varying CpG density across the genome (as described in (114)). Estimation of the most likely methylation state is made in Batman output tiles in 100 bp windows. The algorithm was run individually on each sample.

3.2.7.2 Differential Methylation Calling

Methylation values were generated for 100 bp regions, called “proxies”, within larger regions of interest (ROI) varying from 500 to 4000 bp in length. With the knowledge that methylation is generally correlated on length scales up to 500 or 1000 bp, DMR calling was performed across entire ROIs. However, it was observed that proxies within ROIs sometimes exhibited significant variation in methylation suggesting that averaging methylation states over these

proxies was not appropriate. Therefore, an algorithm was designed that would call DMRs at the length scale of ROIs, but using the methylation values of the individual proxies within a ROI, using an adapted empirical Bayesian method (140). To correct for multiple testing, the estimated the False Discovery Rate (FDR) using two approaches: the q-value approach and a permutation-based approach whereby sample labels were permuted a large number of times (>1000).

3.2.7.3 LD Block Methylation and Sliding Windows Analysis

Linkage Disequilibrium blocks around genotyped susceptibility genes were defined as per Gabriel (141), as implemented in HAPLOVIEW v.4.1 For each block, subjects were grouped by genotype, without reference to case or control status, for each T2D susceptibility SNP. Average methylation scores per block were calculated by summating the methylation scores for all BATMAN windows across each block and dividing by the number of windows. Analysis was then performed by non-parametric (Kruskal-Wallis) and parametric (Linear Regression) means for methylation score with respect to genotype status these groupings. Permutation empirical *p*-values were calculated by retaining observed methylation scores and shuffling genotype assignment 10,000 times.

A sliding window analysis was performed across these LD blocks by utilising 100 bp BATMAN methylation output windows. This is similar in concept to sliding windows analysis performed with contiguous SNP haplotypes. Starting with a window size of one and moving one window along per calculation across the entire block, Kruskal–Wallis and Linear Regression analyses were performed for the genotype groups with respect to methylation scores. Window size was then increased by one on each pass and the analysis repeated, until the window size equalled the entire LD block. Resultant *p*-values were outputted and plotted at the midway point for each window. Compared to linear regression, the more conservative Kruskal–Wallis analysis can still be significant when there is a non-linear relation between genotype and methylation status (which is more difficult to assign biological interpretation) so the mirroring pattern of the linear regression analysis gives further robust support to the significant relationship between methylation and genotype for the *FTO* LD block. Sliding windows analysis, statistical calculation, and permutation scripts were written with the R package.

3.2.8 Chromatin studies

Chromatin studies were performed alongside this study of DNA methylation to identify common histone modifications, including H3K4me1, H3K4me2, H3K4me3, CTCF, H3K9me1 and H3K9me2, to inform the interaction between epigenetic modification and gene function. Chromatin immunoprecipitation was performed on the same T2D designed arrays (ChIP-chip) in normal human skeletal muscle cells, by P. Akan (Sanger Institute) as part of a larger experiment using a standard protocol (142) and detection of ChIP peaks using MPeak software. Duplicates were performed for all, with 90-95% agreement between replicates for all antibodies.

3.2.9 Gene expression

RNA-Seq data generated by an Illumina Genome Analyser from RNA derived from post-mortem cerebellum tissue samples, collected within 24 hours of death, of six anonymous unrelated donor males was available from Wang *et al* (143). C. Bell and G. Lewis aligned these data using TopHat 1.0, which incorporates the Bowtie aligner and additionally generates splice junction reads. SAM output files were visualised with the Integrative Genomics Viewer (version 1.4.01, http://www.broadinstitute.org/igv).

3.3 Results

3.3.1 Sample preparation

Sonication of genomic DNA was successful and produced DNA fragments of a uniform size distribution (see fig 3a).

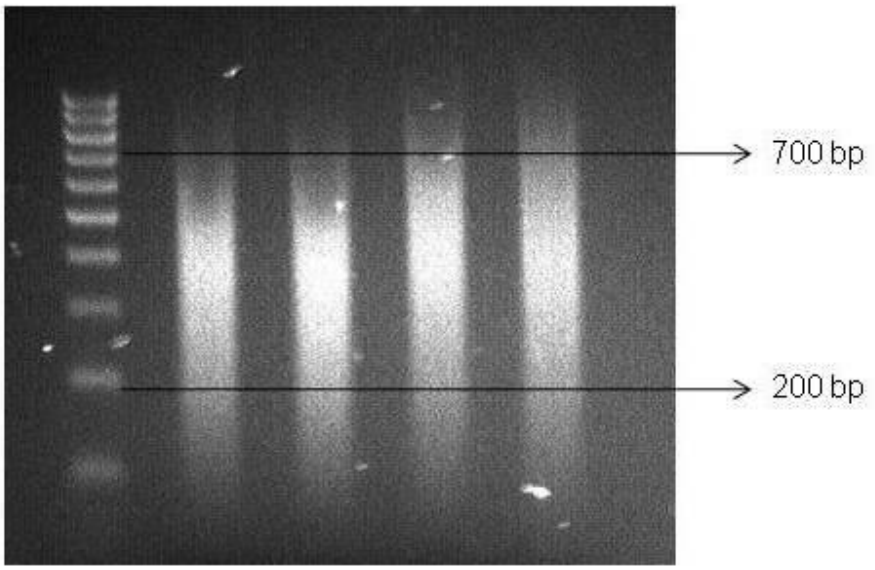
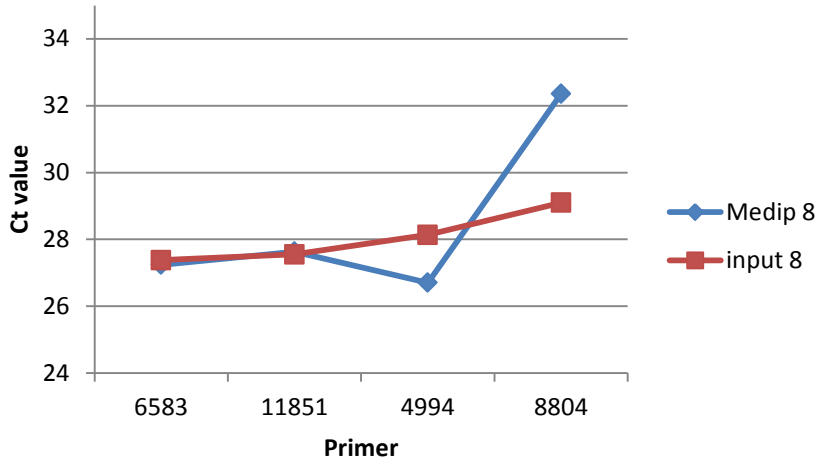


Figure 3a. Sonicated DNA run on an agarose gel showing size range of DNA fragments

After optimising the amplification of Medip DNA, the qPCR quality control check showed uniformly successfully Medip enrichment in all samples (see figure 3b). Figure 3b demonstrates that the Ct (cycle threshold) values of the Medip samples are similar across the amplicons targeted at known methylated regions (6583, 11851, 4994). In the reaction using an amplicon targeted to a known unmethylated region (8804), amplification of the Medip DNA has a much greater Ct value, reflecting a greater number of cycles to amplify the DNA that has been successfully enriched for methylation by Medip. This Ct value, and the fold change between it and the 3 previous targets, provides evidence that the Medip enrichment has been successful. Furthermore, the input DNA (i.e. total DNA) shows equivalent Ct values across all 4 amplicons as it has not been enriched for DNA methylation and therefore amplification of methylated and unmethylated regions are equally possible.



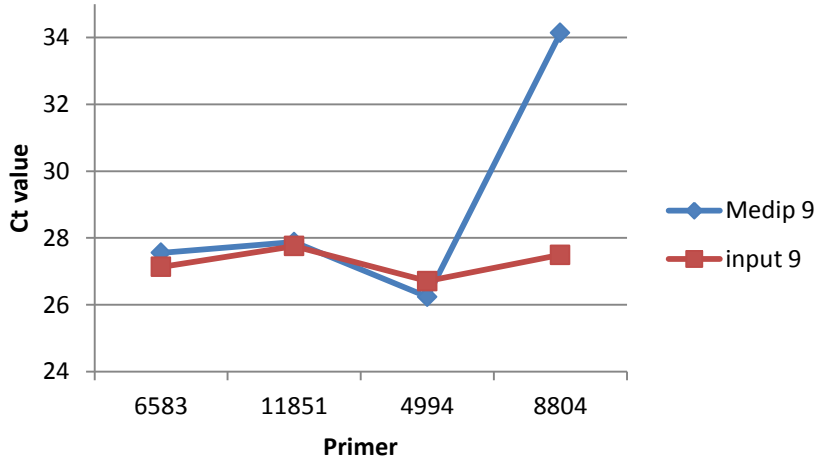


Figure 3b. Plot of qPCR analyses from 2 samples demonstrating good Medip enrichment. Successful enrichment is determined by the visible fold change required to amplify the MeDIP DNA fraction using an unmethylated primer (8804), compared to methylated primers (6583, 11851, 4994).

The integrity of amplified Medip and input samples was checked on an agarose gel and with a nanodrop concentration, and confirmed that all samples had reached the target amplification to a concentration of 60ng/uL. Subsequently, the 60 amplified samples (30 Warren 2 T2D and 30 EFSOCH) were labelled with barcodes and sent for hybridisation to Nimblegen arrays.

3.3.2 Nimblegen array

Array hybridisation was performed on samples and standard QC checks performed by Nimblegen did not identify any problems.

3.3.3 DNA Methylation Analysis within T2D Association SNP LD Blocks

Bioinformatic analysis for this study was performed by C. Bell and G. Wilson, UCL. Array data underwent Lowess normalisation to smooth out any variation arising from differences in colour labelling. Using the Batman algorithm DNA methylation from individual samples that had passed through this MeDIP-chip approach was quantified. Whilst the exact methylation states of individual CpGs are not measured with the MeDIP technique, an estimate of the relationship between genetic haplotypes and DNA methylation was made. BATMAN window methylation estimate scores across each T2D association SNP LD block were summated and averaged; in doing so, an average 'methylation load' was calculated for all 60 individuals, for each LD block. Individuals then were grouped into three genotypic sets, by their respective genotype for each T2D-association SNP (or tagging SNP in perfect LD with association SNP) and Kruskal-Wallis and Linear Regression analyses were performed. Average methylation scores for each genotype group and uncorrected *p*-values for the Linear Regression analysis are shown in Table 3b.

					Average Methylation			
Chr	LD Block		Genotyped SNP	Gene/ Locus	11	12	22	*p*-value
1	120236149	120398430	rs2934381[a]	*NOTCH2*	0.496	0.495	0.502	0.187
2	43529937	43617946	rs7578597	*THADA*	0.492	0.512	0.504	0.564
3	12298413	12372392	rs1801282	*PPARG*	0.512	0.509	0.511	0.640
3	64673853	64705161	rs4607103	*ADAMTS9*	0.477	0.472	0.481	0.760
3	186971576	187031377	rs4402960	*IGF2BP2*	0.502	0.493	0.502	0.016
4	6317902	6363877	rs10010131	*WFS1*	0.581	0.604	0.590	0.982
7	28147081	28175361	rs864745	*JAZF1*	0.501	0.492	0.494	0.317
8	118252732	118254914	rs13266634	*SLC30A8*	0.333	0.350	0.303	0.865
9	22122209	22126489	rs10811661	*CDKN2A/CDKN2B*	0.543	0.512	0.512	0.389
10	12367941	12368040[b]	rs12779790	*CDC123/CAMK1D*	0.611	0.590	0.671	0.913
10	94426831	94467199	rs1111875	*HHEX/IDE*	0.480	0.483	0.483	0.369
11	17350649	17365206[c]	rs5219[c]	*KCNJ11*	0.501	0.502	0.498	0.540
12	69942990	69949369	rs7961581	*TSPAN8*	0.457	0.461	0.470	0.289
16	52357008	52402988	rs8050136	*FTO*	0.497	0.510	0.531	9.397×10^{-4}
17	33170413	33182480	rs757210	*HNF1B*	0.423	0.430	0.427	0.382

Table 3b. Average methylation levels within each LD block, per genotype. *P*-values are calculated by Linear Regression and are shown uncorrected (11 – homozygote common, 12 – heterozygote, 22 – homozygote rare allele). a) r^2=1 with rs10923931, b) Not in LD block – single Batman window of 100 bp utilised, c) LD block of associated SNP rs5215 used as rs5219 not typed in HapMap, r^2 =0.995 with rs5219. Methylation is given as average BATMAN scores across regions (0 = unmethylated, 1 = fully methylated).

3.3.4 Sliding Windows & Permutation Methylation Analysis of the *FTO* LD Block

In the integrated epigenomic-genomic analysis the large 46 kb LD block of the *FTO* gene (Figure 3c) was the only locus to reach nominal statistical significance by Kruskal-Wallis analysis ($p = 0.014$). This result would not have been significant with multiple testing correction, so a permutation analysis was performed by shuffling genotype assignment with the observed methylation scores 10,000 times and achieved a significant empirical *p*-value of 0.012. Subsequent analysis by Linear Regression alone was highly significant with $p = 9.40 \times 10^{-4}$, and empirical *p*-value calculation by 10,000 permutations of 1.0×10^{-3}. Age was also included as a factor in this later analysis, as there was variation in ages of participating subject, and this has been identified as an independent factor in determining methylation at certain loci (63, 118, 119) but was excluded as a significant confounder ($p = 0.676$). The *FTO* SNP rs8050136 risk allele A homozygotes were shown to have a higher average level of methylation of 0.531, with heterozygotes at an intermediate level at 0.510 and G homozygotes with 0.497 (Table 3c).

The initial association for *FTO* with BMI was identified by SNP rs9939609 (144), though a number of subsequent studies using differing SNPs have all, with equal significance, replicated this finding as they are all capturing the same common haplotype (HapMap CEU frequency = 0.425) (Figures 3c and 3d). This LD Block encompasses just under half of the first intron, exon 2 and the beginning of the second intron of the major *FTO* isoform.

	***FTO* LD Block (46 kb)**	**Broad Peak (7.7 kb)**	**Narrow Peak (900 bp)**
Region	52,357,008-52,402,988	52,371,700-52,379,399	52,378,500-52,379,399
AA	0.531	0.529	0.603
AC	0.510	0.503	0.564
CC	0.497	0.478	0.507
***p*-value**	0.014	8.69×10^{-6}	5.630×10^{-5}

Table 3c. Results shown for the entire *FTO* LD block, the broad 7.7 kb 60-window peak (figure 3f) and the narrow 900 bp 9-window peak (figure 3e). Methylation is given as average BATMAN scores across regions (0 = unmethylated, 1 = fully methylated).

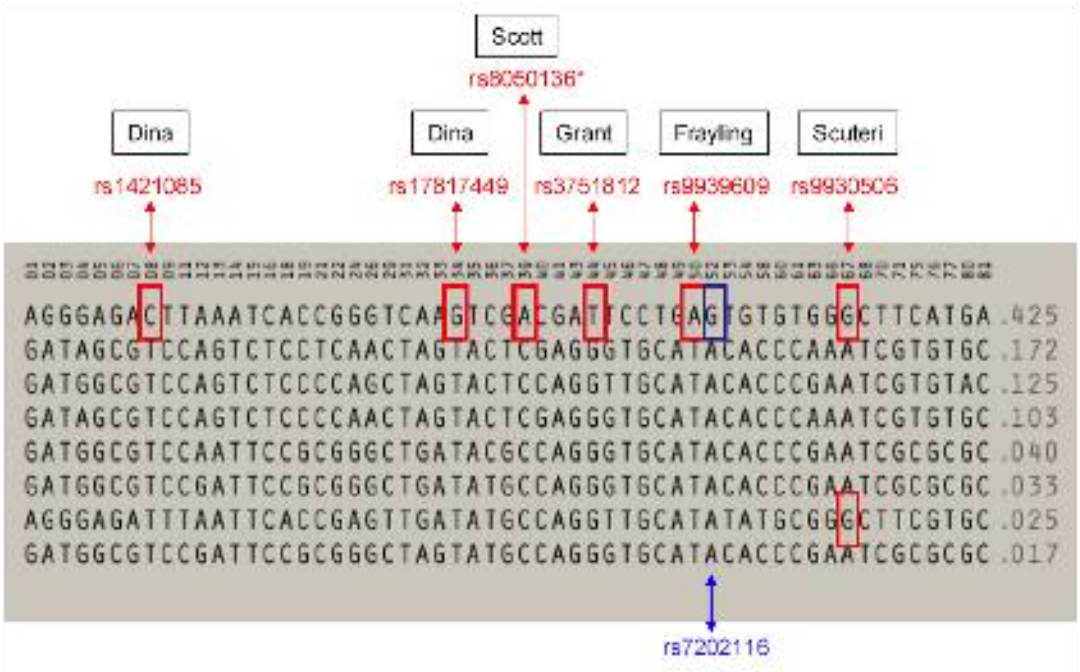


Figure 3c. Haplotypes for the *FTO* suscepbility LD block from HapMap CEU. This figure summarises the SNPs identified in published studies of this haplotype. The SNP rs8050136 genotypes were used as the haplotype tagging SNP for this study's subjects. SNP 52 (rs7202116) creates or abrogates a CpG site according to allele.

To investigate whether this methylation difference was being driven by a distinct region within the 46 kb LD block and, if so, to isolate this location, a sliding windows analysis utilising BATMAN methylation scores (methylation estimates are tiled in 100 bp windows) was performed. The sliding window analysis started with a window size of 1 (100 bp) and increased iteratively by one on each pass, up to the maximum window number of 334, moving across the LD block one window at a time. A central peak of difference in methylation was most prominent in the 9 window (900 bp wide, at windows 161-169, chr16:52,378,500-52,379,399) analysis with Kruskal-Wallis $p = 5.630 \times 10^{-5}$, empirical $p = < 1 \times 10^{-5}$ with 10,000 permutations and Linear Regression analysis $p = 1.94 \times 10^{-5}$ (Figure 3e). Methylation averages were again highest for the A homozygotes (AA 0.603, AC 0.564, CC 0.507). In addition, plotting the slope of linear regression analysis for the 9 window across the LD block indicates that the *p*-value peaks all co-locate with the same negative regression slope direction. From these findings, it can be inferred that all of these regions could be contributing, with varying strength, to the increased methylation signal.

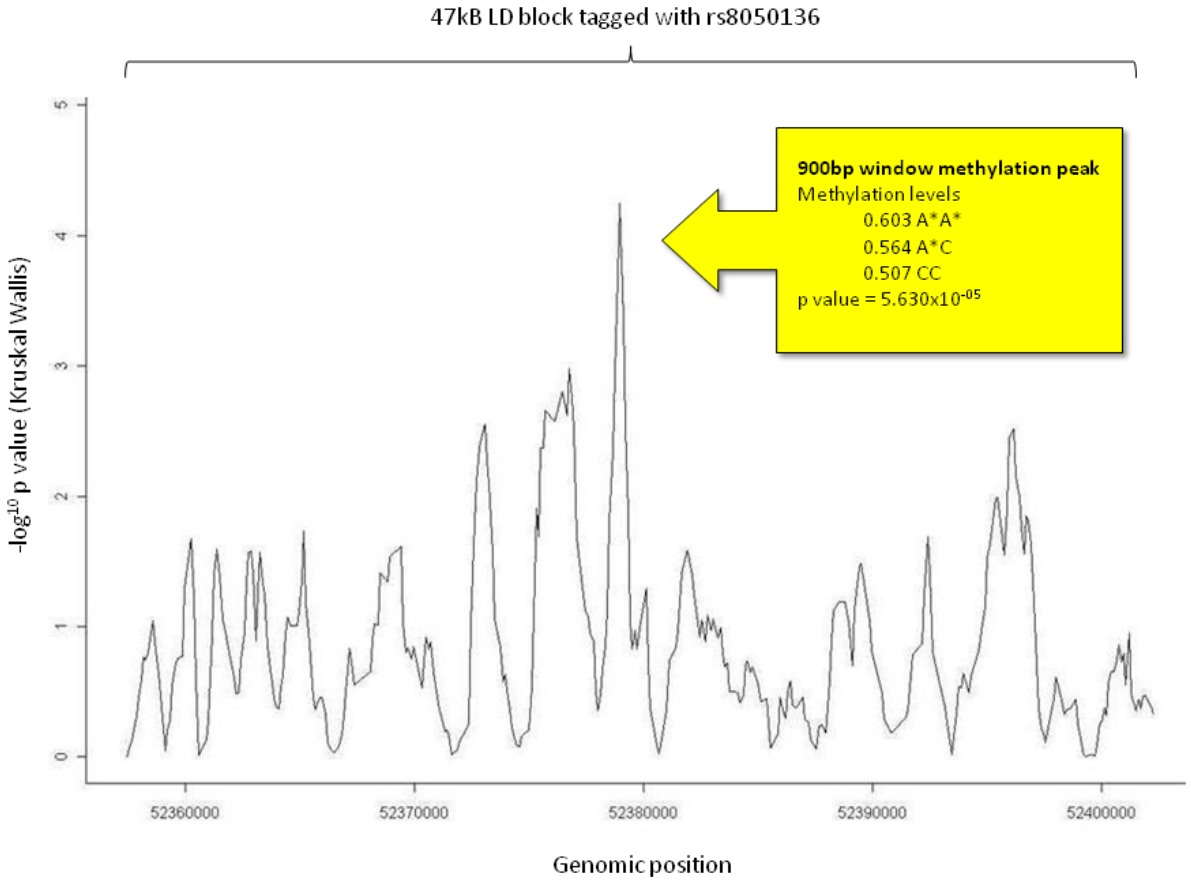


Figure 3d. Narrow 9 window (900bp) peak of methylation derived by sliding windows analysis across *FTO* LD block

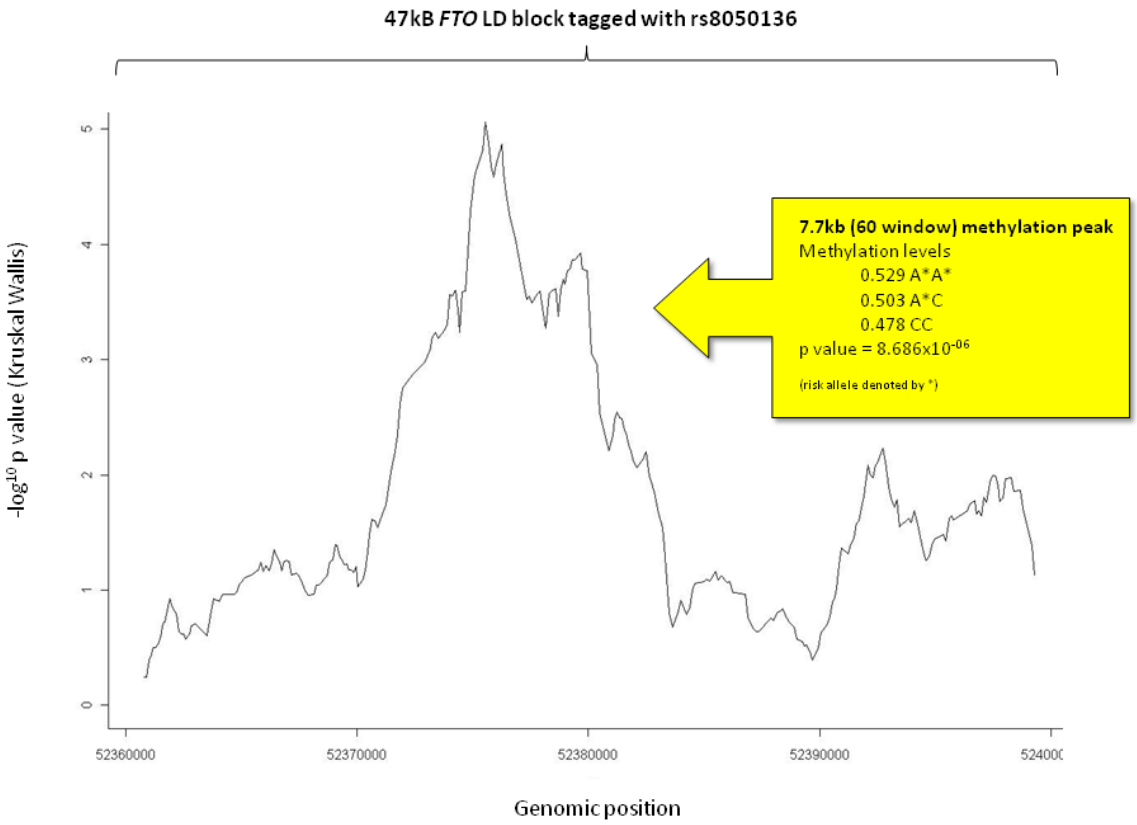


Figure 3e. Broad 60 window (7.7.kb) peak of methylation

The most significant result was identified with a window size of 60 (7.7 kb - as repeat regions excluded from BATMAN analysis - windows 110-169, chr16:52,371,700-52,379,399) with Kruskal-Wallis $p = 8.69 \times 10^{-6}$, empirical $p = < 1 \times 10^{-5}$ with 10,000 permutations, Linear Regression analysis $p = 1.33 \times 10^{-7}$ (Figure 3f) and age $p = 0.444$. This 7.7 kb window becomes most significant just at the point when it encapsulates the 900 bp window at its downstream edge. The same trend in the average methylation scores was again identified (AA 0.529, AC 0.503, CC 0.478). The sliding windows analysis was in addition performed across the other T2D association LD blocks, but none showed any significant increasing *p*-values with window sizes greater than 10 (1 kb) (data not shown).

The findings at the *FTO* LD block, using an integrated epigenomic-genomic analysis, give a strong estimate of the relationship between haplotype and DNA methylation level extending across the 7.7 kb region. Since this approach is calculated in 100 bp windows it is unable to define the actual methylation state of individual CpG sites and their relationship to SNPs. We therefore determined the genetic and epigenetic architecture of the 900 bp peak window in order to explore its contribution to the signal.

3.3.5 Investigation of the Genetic Architecture Underlying the 900 bp Window Peak

The central 9 window peak of 900 bp contains only seven CpG sites in the reference sequence, although of the eight common SNPs within this region, three create or abrogate additional CpG sites (being YpG or CpR SNPs, IUPAC: Y = C or T, R = G or A). These SNPs (rs7206629, rs7202116 and rs7202296) were all included in the AFD-EUR European panel (dbSNP). In this dataset, the SNPS all possess identical minor allele frequencies implying high likelihood of residing on the same haplotype (Table 3d). Only rs7202116 is included in the HapMap CEU data set, and the methylated state allele, G of a CpG, as opposed to the A alternate allele, is present on the susceptibility haplotype, and is in absolute, and near perfect LD with the susceptibility SNPs identified (Figures 3c and 3d). Therefore those with the obesity-susceptibility haplotype would possess ten CpG sites capable of methylation, compared with seven, within this peak region.

		R^2					
SNP	**rs1421085**	**rs17817449**	**rs8050136**	**rs3751812**	**rs9939609**	**rs7202116***	**rs9930506**
rs1421085		0.927	0.931	0.931	0.931	0.965	0.835
rs17817449	0.963		1	1	1	0.964	0.833
D' **rs8050136**	0.965	1		1	1	0.966	0.841
rs3751812	0.965	1	1		1	0.966	0.841
rs9939609	0.965	1	1	1		0.966	0.841
rs7202116*	1	1	1	1	1		0.871
rs9930506	0.963	0.962	0.964	0.964	0.964	1	

Table 3d. LD relationship for *FTO* Association SNPs and rs7202116. Results are given for D' and R^2 in CEU HapMap population. *Indicates the methylation critical SNP within 900 bp window peak.

3.3.6 Validation of Methylation Differences by Pyrosequencing

To validate the methylation differences identified between the susceptibility haplotype and non-susceptibility haplotypes a quantitative DNA methylation pyrosequencing experiment was performed. Bisulphite-pyrosequencing was performed on 80 samples, using the initial 60 plus an additional 20 samples from the same cohorts. Within the 900 bp window four CpGs were investigated, including one of the SNP dependent CpGs (rs7202296). An additional six CpGs were examined that lie directly beneath the broad 7.7 kb peak, including one SNP dependent CpG (rs11075988).

Cytosine methylation levels at the CpG at rs7202296 CpR were, as expected, completely dependent on the presence of CpG in the genetic sequence, with average methylation levels of 87% for A homozygotes, 55.7% for AC heterozygotes and 11.5% for C homozygotes, with respect to rs8050136 genotype (Table 3e and Figure 3g). On examination of the sequence information from the pyrosequencing data, one individual homozygous for the A rs8050136 allele was in fact homozygous CpA and another was heterozygote and this was due to non-perfect LD between rs8050136 and rs7202296. Excluding these two individuals, the methylation level was in fact 98.6% for those who were CpG at rs7202296. The additional cytosines examined within this window though showed no significant differences that could be accounted for by allele-specific methylation (ASM) or evidence of a *cis*-methylation effect on the surrounding non-polymorphic CpGs (120). Of note is the ~50% methylation values of the cytosine at 52,379,254, implying that whilst no evidence of ASM, there may be parent of origin specific methylation imprinting of this CpG, although *FTO* is not known to be imprinted. The three in-phase CpG-creating SNPs therefore would easily account on their own for the ~10% difference seen in methylation levels in this 900 bp window peak region identified in the initial

MeDIP array experiment. The additional CpGs investigated beneath the peak of Window 60 showed no evidence of ASM, although there were also methylation changes associated with a CpR site at rs11075988. For this SNP the allele that abrogated the ability to methylate was in LD with the rs8050136 susceptibility genotype A, *i.e*. it was on the alternate phase (Data not shown).

	CpG Location			
rs8050136 Genotype	52379190*	52379221	52379251	52379254
AA	87	95.1	96.7	49.6
AC	55.7	95.5	97	51.6
CC	11.5	95.1	96.6	49.9

Table 3e. Bisulphite-pyrosequencing validation assay of 900bp window quantifying cytosine methylation (%) at CpG sites within the 900 bp 9-Window. *CpG-creating SNP rs7202296 dependent.

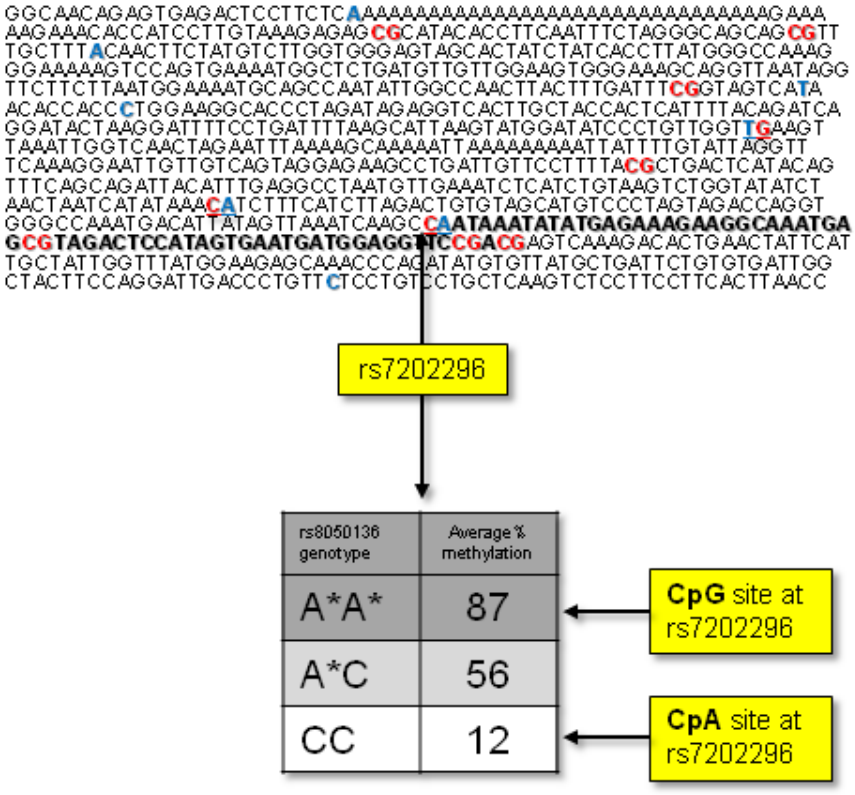


Figure 3f. Single-base pictorial representation of the whole 900bp 9-window with pyrosequenced region (in bold) showing CG sites (red), SNPs (blue) and SNP-CpG sites (red and blue). Methylation at rs7202296 is shown (according to genotype of the tag SNP (rs8050136).

Fourteen healthy brain samples (4 hypothalamus and 10 prefrontal cortex) were examined with the same bisulphite-pyrosequencing arrays to identify whether there might be allele-specific methylation in tissues that show high *FTO* expression and have a known role in central energy balance. The genotype of rs7202296 was determined from pyrosequencing, and it was assumed that it held the same LD relationship to the tag SNP found in whole blood samples. The same findings at the CpG-creating SNPs described above were identified (Table 3f), and no allele-specific effects on the surrounding CpG sites were found.

	CpG Location			
rs8050136 Genotype (n)	52379190*	52379221	52379251	52379254
AA (1)	96.0	100.0	100.0	84.0
AC (8)	53.3	97.3	100.0	80.3
CC (5)	3.2	96.0	100.0	82.2

Table 3f. Bisulphite-pyrosequencing validation assay in 14 healthy brain samples, quantifying cytosine methylation (%) at CpG sites within the 900 bp 9-Window. *CpG-creating SNP rs7202296 dependent.

3.3.7 Evolutionary Analysis

This analysis was performed by C. Bell, UCL, to elucidate possible evolutionary context in which these genetic-epigenetic interactions would function. Examination of the three critical CpG-creating SNPs within the narrow peak (rs7206629, rs7202116 & rs7202296) reveals that the latter two have both transitioned from CpA to CpG and gained the ability to methylate from their ancestral state (in *Pan Troglodytes*, *Pongo pygmaeus* and *Macaca mulatta)* (Table 3g). This particular transition is a rarer mutational event than reverse transition in which deamination of methylated cytosines occurs (145). Analysis of dbSNP data (release 129) indicates that whilst ~60% of the set of 193,133 validated SNPs of known ancestral state on chromosome 16 co-occur at a CpG site, only ~12% of them have gained the ability to methylate since the common ancestor, and less than 0.5% lie within at least 75 bp of each other, as per rs7202116 and rs7202996. One of the three CpG-creating SNPs, rs7202116, is located within a short 49 bp mammalian evolutionarily constrained element (Chr16:52,379,085-52,379,132) as identified by Genomic Evolutionary Rate Profiling (GERP) (146).

SNP		Major	Minor	EUR		CHN		AFR		Ancestral
rs7206629	YpG	T	**C**	0.609	0.391	0.833	0.167	0.587	0.413	C
rs7202116	CpR	A	**G**	0.609	0.391	0.833	0.167	0.674	0.326	A
rs7202296	CpR	A	**G**	0.609	0.391	0.833	0.167	0.674	0.326	A

Table 3g. Allele frequency for CpG-creating SNPs within the 900bp window. CpG creating alleles are shown in bold. EUR = European, CHN = Asian, AFR = African American from dbSNP. CpG-creating sites (YpG or CpR SNPs, IUPAC: Y = C or T, R = G or A).

Further analysis of the haplotypes present in humans was performed using the 420 HapMap Phase II phased SNP haplotypes for CEU, YRI and ASN and indicated that this region comprises 60 unique haplotypes. An inferred ancestral haplotype included in a Median-Joining Network analysis (147) indicates the methylation ability of the susceptibility haplotype is close to the ancestral with respect to the number of mutations between them. This suggests that the ability to methylate was gained early and maintained on the susceptibility haplotype, whilst subsequently lost in the younger haplotypes present in humans.

3.3.8 Enhancer Activity Evidence within the 7.7 kb Region

A ChIP-chip study, using the same T2D array, as part of a larger experiment performed by P Akan (Wellcome Trust Sanger Institute), investigated normal skeletal muscle cells and identified evidence of enhancer activity within the 7.7 kb window, with the classic signature of H3K4me1 and no other strong signal. These findings are consistent with those of Ragvin (148) who have recently identified that the susceptibility locus within the first intron of *FTO* contains long range enhancers influencing the expression of the *IRX3* gene located ~470 kb downstream of this region and that *FTO* expression is not affected in this process. One of the two critical Highly Conserved non-coding Elements (HCNEs) implicated to affect *IRX3* expression by enhancer activity and overlap with risk variants, designated as element 6, in Ragvin *et al*. is located at Chr16:52,377,745 - 52,378,287 and resides centrally within the broad 7.7 kb peak (Figure 3f). This 542 bp element is also only 213 bps 5' of the 900 bp window.

Examining the Sanger GENEVAR HapMap CEU expression data identified no significant differences with respect to *FTO* haplotype status in *FTO* or *IRX3* expression from lymphoblastoid cell lines in children or adults (149) and insufficient expression in untreated and activated CD4+ T cells (150). We also used an available RNA-Seq dataset derived from the cerebellum of six male individuals, but were unable to identify allele-specific expression of *IRX3* or *FTO* due to low coverage across the former, and the presence of only one common but uninformative SNP in the latter.

3.3.9 T2D-DMR Analysis

A case-control analysis across all the assayed regions of the array (T2D and obesity related phenotype candidates, 1q linkage region and imprinted regions) was performed to determine any disease-related DMRs. There were no disease-associated DMRs with statistical significance post FDR correction, indicating a lack of power in the study.

3.4 Discussion

In this study, the power of large-scale GWAS has been harnessed to look for genetic-epigenetic associations that may occur in conjunction with known genetic susceptibility variants. A methylation association with the strongly replicated disease haplotype of the *FTO* gene, tagged by SNP rs8050136 has been identified using a Medip-chip approach. This association was confirmed and validated at single-base resolution using bisulphite pyrosequencing and it was found to be a genetically-led phenomenon occurring due to the phase of three CpG-creating SNPs in LD within a narrow 900 bp window peak, creating or abrogating sites for DNA methylation. Work in murine strains supports the inference that inherited genetic variability is a major determinate on epigenetic variability (68) and the significant contribution of allele-specific methylation at CpG islands has been further characterised by Zhang *et al*. (151). However, how much of this effect is driven by CpG-creating SNPs directly, or additionally their influence on surrounding CpG methylation, or other genetic polymorphism effects on the methylation machinery or a combination of all of above is not known. The absence of allele-specific methylation differences in CpG sites downstream of the CpG-SNPs in this study lends further support to these findings being a haplotype-driven effect, rather than one driven by a primarily epigenetic phenomenon. Had this study been performed in a larger sample set, allowing sufficient power for case-control analysis, it may have been possible to define our methylation differences erroneously as a disease-associated ‘‘differentially-methylated region’ and overlook a purely genetic event.

A recent paper by Toperoff (152) has provided some external validation of the results of this study in larger sample numbers (710 with T2D and 459 controls) with an assay DNA methylation in approximately 3 million of the 28 million CpG sites across the genome. Using a

microarray-based assay, they identify a CpG site located 11bp upstream of the known T2D/obesity variant rs1121980 that is hypermethylated in a pooled set of samples of individuals carrying the risk allele A identified hypermethylation compared to those containing the G allele. However, the authors present conflicting data from an analysis of cases vs. controls, identifying relative hypomethylation (-2.55%, $p=2x10^{5}$, q=0.0003) at the same CpG site in those with T2D. In an independent sample cohort from a longitudinal study of diabetes, the authors identify hypomethylation at this same CpG site in individuals sampled prior to their known onset of diabetes compared to those unaffected. They hypothesise that this gives support of this methylation signature being causative, however they do not rule out the methylation differences being due to pre-clinical phenotypes. More importantly, they do not study the risk alleles in the individuals included in this study. This is a notable omission if trying to examine the precise relationship between genotype and DNA methylation at this locus, and especially given that this was a study of relatively young individuals with diabetes that might therefore be expected to have a high genetic risk. It seems likely that the conflicting association between the phase of the rs1121980 risk allele and the adverse phenotype with methylation at this particular CpG site is due to the presence of another *cis*-controlling SNP that has not been characterised. The data presented in this chapter, specifically the presence of a nearby SNP (rs11075988) in the alternate phase to the SNP-CpG of interest would support this.

Additional support for our findings is derived from a recent paper by Almen (153) that identifies methylation differences associated with *FTO* genotype in *cis*- and *trans*. In this study, a genome-wide methylation array (Illumina 27k HumanMethylation array) was used to identify an interaction between the *FTO* risk allele and the methylation state at other loci. Interestingly, they identify long-range control over methylation at 3 distant loci as well as two loci on chromosome 16 where *FTO* is located. The gene loci that are associated with the *FTO* risk allele have functions in transcriptional regulation that could potentially have wide-ranging phenotypic effects.

It is important to consider these associations of *FTO* epigenotype and phenotypic outcome in the light of uncertainty over the genetic function of *FTO*. The *FTO* disease haplotype has been well studied since it was first identified as containing T2D risk variants in a GWAS. *FTO* has a common haplotype located across intron 1, exon 2 and intron 2 captured equivalently by various SNPs. Despite the initial association of the *FTO* risk haplotype with T2D, it has since been shown to be mediating its disease susceptibility effect through obesity (144) (154). The effect size of this risk variant is small, with adult homozygotes displaying an increase in weight

of 2-3kg, but it is a common variant and its effects are seen in early life (155). Early studies on *FTO* identified phenotypic similarities observed with mutations in the nearby *Ftm* gene raised uncertainty as to the direct role of *FTO* in mediating obesity and energy balance. Human studies of *FTO* have also added complexity to these theories, and do not elucidate a clear functional role for *FTO* in the pathogenesis of obesity. Loss-of-function mutations in the *FTO* gene have been recognised in a rare autosomal-recessive lethal syndrome accompanied by severe growth retardation (OMIM #612938) (156). The ability to translate the findings of these loss-of-function models to the effects of the common *FTO* sequence variants has proved difficult, not least as the latter are intronic and therefore their effects are expected to be subtle effects on gene regulation. Attempts to identify expression differences in *FTO* in human skeletal muscle and subcutaneous adipose tissue with respect to risk SNPs have been unsuccessful to date (157) and no evidence of allele-specific expression in immortalised lymphoblastic cell lines was able to be established (158). However, *FTO* heterozygous loss-of-function mutations have been found in both obese and lean subjects, further clouding its causative role (159).

Recent murine models of *fto* deficiency lend support to *FTO* having a causal role, with reduced body weight and fat mass observed in the *fto* knockout mouse from an early developmental stage (160). Heterozygote $fto^{-/-}$ mice display presumed haploinsufficiency, with a similar phenotype seen in heterozygous and wildtype mice. A similar, but lesser, obesity-protective effect has been observed in the hypomorphic *fto* missense I367F mutant (161) which is considered to better replicate of human models as the mutation affects intron 1, the location of human disease-associated SNPs. In contrast to the knockout model, a dominant-negative effect of the fto^{1367F} is suggested by the phenotypic similarity between heterozygous and homozygous mice. Both animal models display an increased metabolic rate and sympathetic over-activity in the presence of *fto*-deficiency, suggesting it has a role in altering energy expenditure. *FTO* is ubiquitously expressed, but at high levels from the hypothalamus, suggesting that is mechanism may be in altering central energy balance.

The mechanisms by which *FTO* acts on a cellular level are being elucidated. In vitro studies identify *FTO* as a member of the Fe(II)- and 2-oxoglutarate-dependent oxygenase superfamilies, which have multiple roles in oxygen sensing and fatty acid metabolism, but recent studies cast suspicion over this being its mechanism of action (162). Murine studies suggest a role in the catalysis of nucleic acid demethylation (reviewed by (163)) and recent studies show that a methylated base (m^6A) contained in cellular mRNA may be the target of *FTO*-mediated oxidative demethylation with potentially wide-ranging consequences (164).

These convincing studies of *FTO* function do now provide the evidence to support its direct functional role and are also beginning to elucidate the exact mechanisms by which it acts. The identification of stable epigenetic modifications associated with genotype may help to elucidate the precise mechanisms of action, but also highlight another possible layer of complexity in that distant regulators may affect the function of *FTO* itself. Recent work has hypothesized that the lack of evidence for *FTO* expression modulation by the susceptibility SNPs may be due to this region having effects on distal surrounding genes including *IRX3*. Ragvin et al. used comparative genomics to identify highly conserved non-coding elements (HCNEs) and overlying genomic regulatory blocks, and has proposed that enhancers in the first intron susceptibility region exert long range regulatory effects on expression of the developmental transcription factor gene *IRX3*, Iroquois Homeobox 3, located in a gene desert ~170 kb 3' of *FTO*. Enhancers are predominately located in intergenic or intronic regions and may act as regulators of gene transcription over long distances (165), have an activating function on chromatin structure (166), are sensitive to CpG methylation (167) and have an important role in developmental processes. Of two HCNE-containing elements with enhancer effects implicated with a metabolic role, one is located within the 7.7 kb methylation window (chr16:52,371,700-52,378,500). Methylation differences, driven by SNP-CpG variation, of this enhancer may impede its action in terms of enhancer-specific transcription factor recruitment, subsequent chromatin DNA looping, enhancer-promoter interaction and enhanceosome formation with subsequent down-regulation of *IRX3* expression. Furthermore, this HCNE is just over 200 bp away from the 5' of the 900 bp window (chr16:52,378,500-52,379,399), placing it within a 2 kb 'shore' region of this enhancer. It may be possible that these 'enhancer shores' act in a similar fashion to 'CpG Island shores' (regions 2 kb either side of CGIs) which have been identified with a more dynamic role in regards to DNA methylation change effects around promoter-associated CGIs (168). The ChIP-chip data from skeletal muscle indicates a H3K4me1 signature within this 7.7 kb region, as well ChIP-Seq data from cell lines confirms a 5K block of H3K4me1 enrichment completely encapsulated here (http://bioinformatics-renlab.ucsd.edu/enhancer) (166) and a recent examination of histone modifications in pancreatic islets also identified this enhancer marker 1.2 kb wide over rs8050136 within the region (112). The inability to determine allele-specific expression in two different datasets used reflects the complexity of studying the functional sequelae of genetic and epigenetic differences and is likely, in part, to be due to insufficient sample size and use of tissues from which *FTO* is not highly expressed. Earlier data presented also highlights the need to consider the role of the fed/fasting state, circadian clock and other environmental conditions when designing functional studies. The absence of correlation between allele and gene expression in

this and previous studies (157) may also suggest a role via *IRX3* enhancer activity rather than *cis*-effects.

Evidence against the potential trans-acting role of *IRX3* over *FTO* in causing the obesity phenotype comes from two main sources. First, the *fto* knockout mouse described earlier targeted exon 2 and 3, only ~1 kb into intron 129), and therefore did not remove any of the putative enhancers. Second, the function of *IRX3* is poorly understood, but a role in pancreatic β insulin- and α glucagon-secreting cells and increased ghrelin-producing ε cells has been suggested through the study of orthologous irx3a in a zebrafish knockdown model. This putative functional mechanism of *IRX3* in pancreatic development conflicts with the knowledge that most obese individuals display an increase in pancreatic beta cell mass as a compensatory response to the peripheral insulin resistance that co-exists (169) and the finding that most previously-identified obesity genes are involved in neuronally-mediated central energy balance (170). However the evidence of functional enhancer capability of this conserved non-coding region is the crucial finding, as its downstream target may have changed or evolved to take on a more complex role over time. It is possible that the role of *IRX3* in mammalian posterior forebrain development, including the hypothalamus, may in fact be critical (171) (172) and underlie mechanisms of energy balance, or a complex interaction between multiple genes in the region.

Loss or gain of CpG dinucleotides over evolutionary time leading to a genetically driven variation in DNA methylation and subsequent higher variance has been proposed to be a major driver in evolutionary adaption as well as disease susceptibility (173). The abrogation of CpG-SNPs can have considerable effects on regional methylation (66) and this may underlie their adaptive evolutionary role (174). The functional consequence of this deamination of methylated CpGs in forming of transcription factor binding sites has been noted in the study of p53 and its critical role in regulation of insulin resistance (175).

In conclusion, genetically driven methylation differences have been identified in association with the *FTO* obesity susceptibility locus, driven by the creation or abrogation of SNP-CpG sites within the risk haplotype. Genetic association studies have identified numerous SNPs of equivalent importance across this region and the difference in CpG methylation capability could highlight the aetiological variants within this susceptibility haplotype. Detailed analysis of the methylation signal and pyrosequencing validation indicate the genetic phase of CpG-creating SNPs are a strong influence in this finding and an understanding of this is a relevant consideration in large-scale studies that associate epigenetic variants with disease. The most significant results in this study identifies a 7.7 kb region containing enhancer sequence and the

observed increased methylation ability of this enhancer region would be highly likely to reduce the efficiency of this regulatory element. Thus the investigation of epigenetic variation may be very useful in narrowing down significant regions in large LD association blocks and proposing functional hypotheses for subsequent follow-up from GWA studies. The functional relevance of these findings remains undetermined, and preliminary studies highlight the complexity of integrating genomic datasets on a sequence, epigenetic and gene expression level.

Chapter 4: Pune Maternal Nutrition Study

4.1 Introduction

4.1.1 Pune Maternal Nutrition Study

The Pune Maternal Nutrition Study (PMNS) is a prospective study of a pre-pregnant population in rural India, led by Ranjan Yajnik (Pune, India) (43) that offers an exciting opportunity to investigate the role of epigenetic factors in fetal programming of type 2 diabetes and obesity. PMNS recruited non-pregnant women of reproductive age from 6 villages near Pune, India in the early 1990s and is summarised in Figure 4a. These women were seen regularly with monitoring of their body anthropometry and menstrual cycle. This close observation enabled careful dating of incipient pregnancy, and when pregnancy was confirmed, women were followed up at 18 and 28 weeks' gestation. At these study visits, women underwent further anthropometry, questionnaire-based assessment of diet and physical activity and blood sampling. Assays of circulating one-carbon cycle micronutrients (see figure 7a) were performed, including red cell folate (its substrate) vitamin B12 (its co-factor), as well as homocysteine (a one-carbon cycle intermediary) and methylmalonic acid (a marker of functional B12 status in tissue). Anthropometry of offspring born to these mothers was performed within 72 hours of their birth. These children were followed up at 6 years of age with anthropometry, body composition (using whole-body dual-energy X-ray absorptiometry, or DEXA), glucose tolerance testing, the HOMA-IR (homeostatic model assessment of insulin resistance) measure of insulin resistance (www.dtu.ox.ac.uk/homacalculator/index.php) and blood sampling for nutritional and metabolic assays. An extremely high follow-up rate was achieved with 700 of the 762 live births studied at 6 years. Follow-up of offspring at 12 years has also recently been completed.

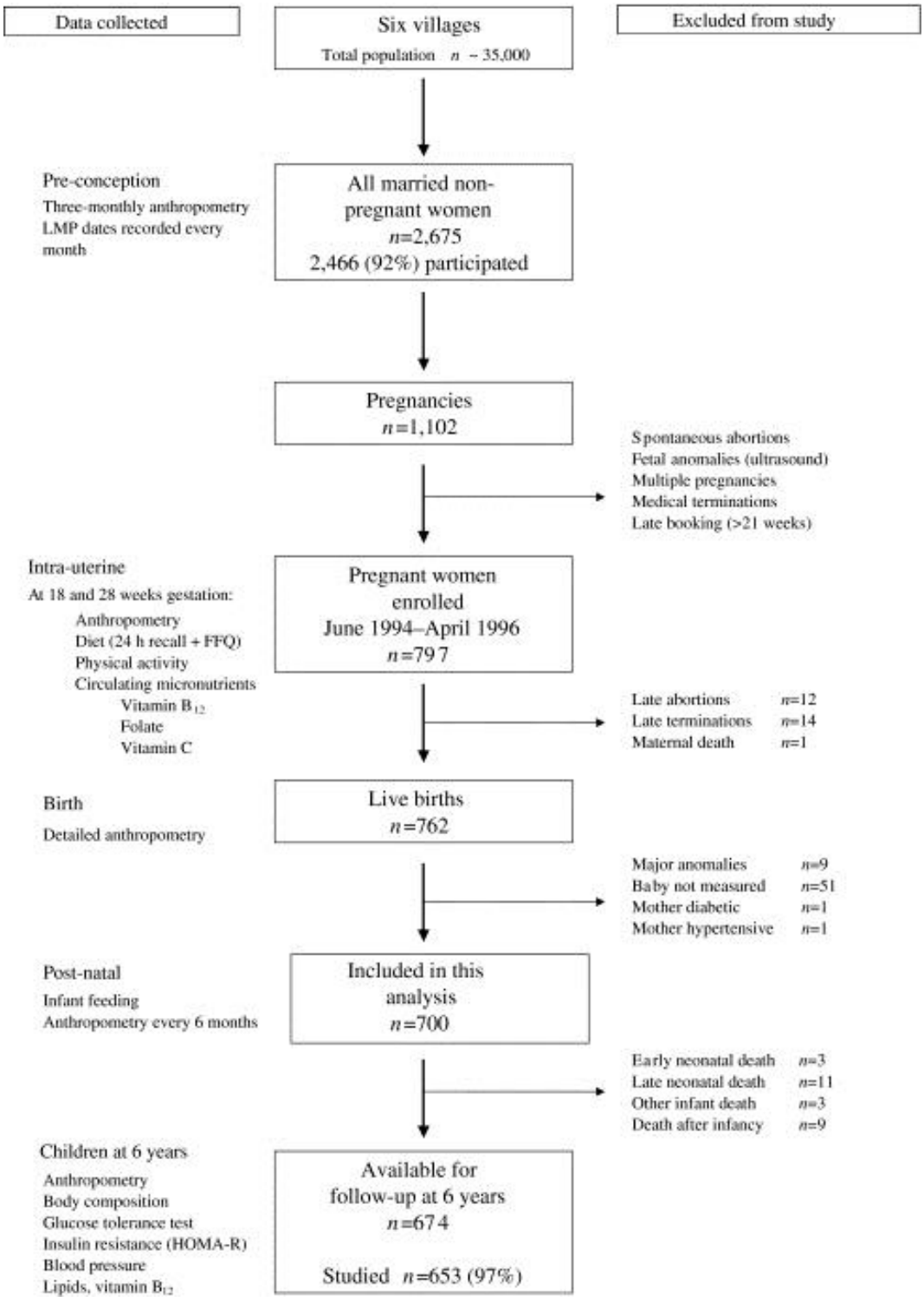


Figure 4a. Pune Maternal Nutrition Study. Summary of participants, recruitment and study design (from (43)).

Over 60% of women had a low concentration of vitamin B12 (<150 pmol/L) when measured at 18 and 28 weeks' gestation, and an inverse correlation between B12 and homocysteine and methylmalonic acid (MMA) provided evidence for this reflecting true deficiency. In contrast, only one woman out of the 618 studied had a low red cell folate concentration. Offspring born to these women were studied at 6 years of age and an association between low maternal vitamin B12 at 18 weeks with higher HOMA-IR in the children was found (p=0.03). In addition, maternal erythrocyte folate concentrations at 28 weeks predicted higher offspring adiposity and higher HOMA-IR (both p<0.01) at 6 years of age (Figure 4b). Regression modeling showed that the offspring of mothers with a combination of high folate and low vitamin B12 concentrations were the most insulin resistant, even when adjusted for the child's own fat mass and plasma B12 concentration.

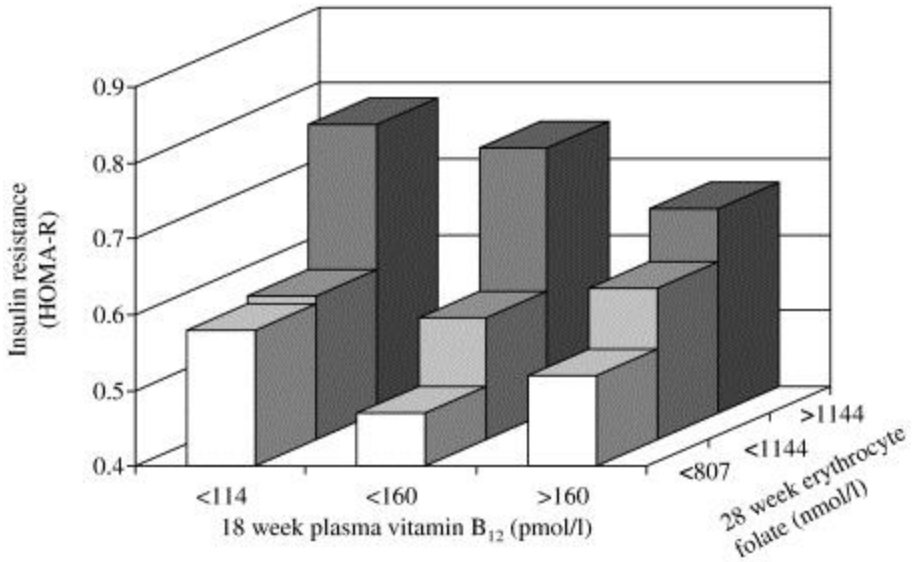


Figure 4b. Relationship between maternal B12/folate status and offspring insulin resistance in the Pune Maternal Nutrition Study. (from (43)).

The notable finding of this study was an association between maternal B12 deficiency with a phenotype of insulin resistance and obesity in offspring. The unusual finding that normal/high folate concentrations were found in association with these other nutritional deficiencies was thought to be due to women of conceptual age being given folate supplementation from pre-conception into their pregnancy. This supplementation is routine throughout the world, but in India was being given at much higher doses than are considered necessary. In the UK, 400mcg

folic acid is recommended to all women who are planning a pregnancy, but in contrast, up to 10mg was being given routinely in India. The 'methylfolate trap' has been proposed as a mechanism by which folate repletion can exacerbate co-existent B12 deficiency via the impaired conversion of homocysteine to methionine (176).

As discussed in the introduction, the one-carbon cycle uses folate as its substrate and vitamin B12 as cofactor to convert homocysteine to methionine; in doing so, one-carbon cycle regulates production of methyl groups for DNA methylation. Preliminary evidence exists showing the role of one carbon cycle disorders in DNA methylation and a programmed phenotype, and this has been discussed in Chapter 1. This pilot study was therefore designed to investigate the environmental-epigenetic interaction between disorders of maternal one-carbon status and DNA methylation patterns in programmed offspring. The study was developed in collaboration with Ranjan Yajnik and Giriraj Chandak (Hyderabad, India), to try and characterise differences in DNA methylation across the genome in offspring according to varied maternal one-carbon status.

4.1.2 Specific objectives

This project sought to identify whether maternal vitamin B12 and folate levels in pregnancy programme insulin resistance and obesity in offspring via epigenetic modifications and had the following objectives:

a) Characterisation of epigenetic variation in offspring associated with maternal vitamin B12 and folate levels
b) To use whole genome DNA methylation profiling techniques to produce an unbiased view of epigenetic variation, with no *a priori* hypothesis as to where this variation would be occurring (other than that determined by the presence of CpG sites capable of methylation).
c) To use Medip-seq as the primary tool for methylation profiling, coupled with the 450k array platform as an experimental validation platform.
d) To gain insight into the gene-environment interactions involved in fetal programming of insulin resistance/cardiometabolic syndrome with which to develop further hypothesis-driven studies.
e) To demonstrate the feasibility of using whole blood to detect differences in DNA methylation associated with *in utero* micro-nutrient deficiencies

4.2 Methods – common to all experiments

4.2.1 Sample selection

Samples collected from the 6-year old offspring born to the Pune Maternal Nutrition Study were chosen for this project. The following selection criteria were used:

1. Age of offspring: It was considered better to use samples from offspring at age six rather than age twelve as at the offspring in the former group had not entered varied stages of pubertal development. In addition, at 6 years of age, the offspring had been exposed to less environmental variation (and therefore potential confounding postnatal effects) than at 12 years of age.
2. Maternal exposure: the two experimental groups were selected from the extremes of maternal exposure, i.e. the 'high risk' exposure of low vitamin B12 and high folate and the 'low risk' exposure of high vitamin B12 and low folate. 'High' and 'low' were defined by tertiles from the whole study population.
3. Childhood phenotype: after selecting for maternal exposure, samples were selected according to the phenotype associated with exposure, i.e. for the 'high risk' maternal exposure, offspring with the highest index of insulin resistance were chosen, and vice versa. This aim of enriching for phenotype was to maximize the potential to identify epigenetic variation in association with the programming.
4. Sex: after selection for extremes of maternal exposure and the programmed phenotype, it was possible to select 6 male offspring from each group to form the first experiment. In subsequent experiments (2 and 3) where larger sample numbers were used, it not possible to select single sex groups for study, nor was it possible to balance the numbers of males and females across the groups.

After defining these selection criteria, the database of study participants was searched by Charu Joglekar (Pune, India) to identify appropriate participants.

- Cases: offspring born to mothers with low B12/high folate levels in pregnancy (in whom the highest risk of insulin resistance was observed).
- Controls: offspring born to mothers with high B12/low folate levels in pregnancy (in whom the lowest risk of insulin resistance was observed).

For the first Medip-seq experiment, 6 individuals from each group were chosen to provide what was thought to be adequate sequencing coverage across two flowcells within the budget available. Experiment 2 (also Medip-seq) used 8 individuals from each group and did not overlap Experiment 1 completely; the reason for this will be explained below. For the cheaper array-based experiment (Experiment 3), all 20 samples were used. Table 4a summarises the samples that were used in these different studies.

Cases = Low B12 and high folate						
Sample ID	**Sex**	**Insulin Resistance**	**Age at sampling**	**Experiment 1: Medip-seq Tag**	**Experiment 2: Medip-seq ID**	**Experiment 3: 450k array**
30	Male	**1.88**	6 Yr	Tag 1 – ATCACG	f_1_s_1	Pune 5
33	Male	**1.88**	6 Yr	Tag 2 – CGATGT	Not used	Not used
161	Male	**0.72**	6 Yr	Not used	Not used	Pune 1
348	Female	**0.51**	6 Yr	Not used	Not used	Pune 10
456	Female	**0.92**	6 Yr	Not used	f_2_s_5	Pune 8
470	Male	**0.75**	6 Yr	Not used	f_2_s_6	Pune 2
533	Male	**1.90**	6 Yr	3 – TTAGGC	f_2_s_7	Pune 6
621	Male	**1.34**	6 Yr	4 – TGACCA	f_1_s_6	Pune 3
622	Female	**1.98**	6 Yr	Not used	f_2_s_8	Pune 9
637	Male	**1.48**	6 Yr	5 – ACAGTG	f_1_s_7	Pune 4
657	Male	**0.69**	6 Yr	Not used	Not used	Pune 7
687	Male	**1.42**	6 Yr	Tag 6 – GCCAAT	Not used	Not used
Controls = High B12 and low folate						
Sample ID	**Sex**	**Insulin Resistance**	**Age at sampling**	**Experiment 1: Medip-seq Tag**	**Experiment 2: Medip-seq ID**	**Experiment 3: 450k array**
25	Male	**0.30**	6 Yr	Tag 7 – CAGATC	Not used	Not used
43	Male	**0.40**	6 Yr	Not used	f_2_s_1	Pune 14
48	Female	**0.34**	6 Yr	Not used	f_2_s_2	Pune 15
71	Female	**0.73**	6 Yr	Not used	Not used	Pune 20
78	Male	**0.87**	6 Yr	Not used	f_2_s_3	Pune 13
183	Male	**0.19**	6 Yr	Tag 8 – ACTTGA	Not used	Not used
205	Male	**0.17**	6 Yr	Tag 9 – GATCAG	Not used	Not used
211	Male	**0.18**	6 Yr	Tag 10 – TAGCTT	Not used	Not used
270	Male	**0.19**	6 Yr	Tag 11 – GGCTAC	Not used	Not used
317	Male	**0.32**	6 Yr	Not used	f_2_s_4	Pune 12
330	Female	**0.41**	6 Yr	Not used	f_1_s_2	Pune 18
410	Female	**0.37**	6 Yr	Not used	f_1_s_8	Pune 17
417	Female	**0.37**	6 Yr	Not used	f_1_s_4	Pune 16
432	Female	**0.52**	6 Yr	Not used	Not used	Pune 19
562	Male	**0.19**	6 Yr	Tag 12 – CTTGTA	f_1_s_5	Pune 11

Table 4a. Summary of case and control samples used in Pune Medip-seq and 450k experiments. Insulin resistance was calculated using the HOMA-IR algorithm (www.dtu.ox.ac.uk/homacalculator/index.php).

4.2.2 Sample collection and preparation

Whole blood genomic DNA samples from the twelve participants above were used. Genomic DNA (unamplified) had previously been extracted using standard phenol: chloroform techniques and stored in Pune, India. The quality and concentration of the samples was checked with visualisation on an agarose gel and Nanodrop in India, with a repeat Nanodrop concentration measure performed after transit to the UK.

4.3 Methods (Experiment 1)

4.3.1 Medip-seq library preparation (multiplexed)

Medip-seq libraries were prepared using the 12 samples, as per the protocol in section 2.3.2. A qPCR of Medip enrichment was performed as a quality control of enrichment efficiency. In this experiment, each sample was assigned an individual index (or tag), ligated to the library fragments during the final PCR, allowing the preparation of a multiplexed library. After indexing, the 12 libraries were pooled and nanodrop concentrations were used to achieve a similar quantity of each sample in the final pooled library. The pooled library was run on an agarose gel in a single well, allowing a single band to be cut out at 250-300 bp therefore having equal fragment size distributions of each sample within the pool.

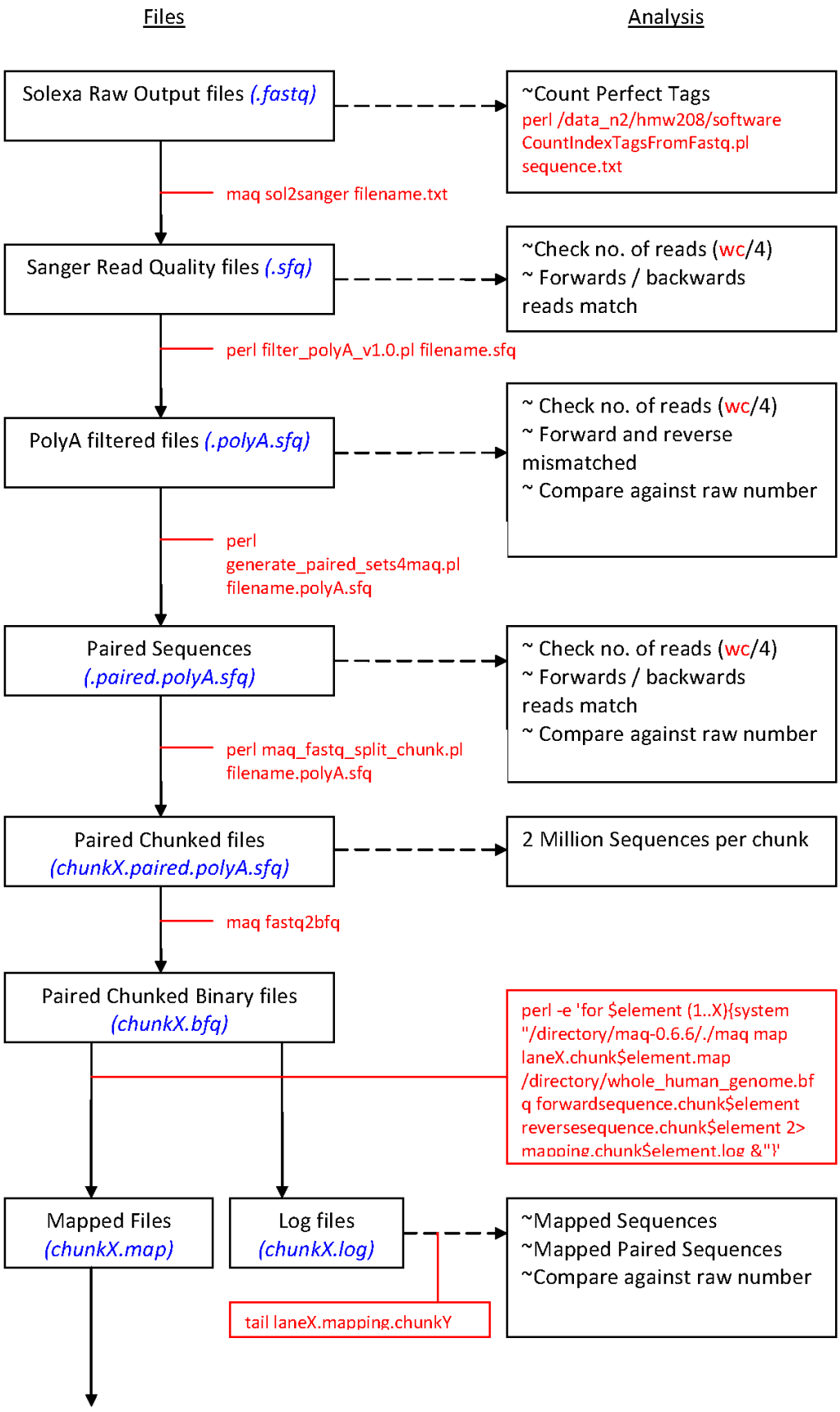
4.3.2 Illumina GAIIx sequencing

The pooled Medip-seq library underwent 36bp PE sequencing on the Illumina GAIIx at the Genome Centre, Barts and the London. A total of 10 lanes of sequencing were performed, split across two flowcells. A single lane of PhiX control (a control library composed from a small viral genome, used routinely in Illumina GAIIx sequencing), was run on each flowcell.

4.3.3 Sequence read processing

Sequence data output from sequencing runs was provided at .txt files. Bioinformatic scripts used in data processing were written by Thomas Down using Perl. The following steps were performed to process the sequence data for downstream analysis and identification of DMRs and is summarised in Figure 3:

1. Converting sequence files: the Illumina GAIIx (also known as Solexa) pipeline provides sequence files in a .txt format unsuitable for downstream processing. The .txt file format was converted to .fastq and subsequently to .sfq, a Sanger sequence format, which enables the counting of reads and mapping.
2. Filters: sequence files undergo a 'PolyA' filter to remove the polyA tails that are a technical artifact associated with the Illumina GAIIx platform.
3. Pairing: as paired-end sequencing was performed, sequence reads were paired for mapping and further analysis.
4. Mapping: sequence reads were mapped to the reference genome using Maq software
5. Q10: a quality filter was applied to the mapped reads, designed to remove the worst 10% of reads.
6. De-multiplexing: mapped sequence reads were filtered by multiplex tag and saved as individual sample files.
7. Sanity-checking of sequence: read counts before and after each step above were checked and plotted to ensure that the processing was proceeding in the expected manner, and to determine the quality of sequence data. After all the above steps had been performed, calibration plots of Medip enrichment were generated using the 'Batman' algorithm, and fragment size distribution was plotted to ensure uniformity across the multiplexed samples.


Files
Analysis
Solexa Raw Output files (.fastq)
~Count Perfect Tags
perl /data_n2/hmw208/software
CountIndexTagsFromFastq.pl
sequence.txt
maq sol2sanger filename.txt
Sanger Read Quality files (.sfq)
~Check no. of reads (wc/4)
~ Forwards / backwards
reads match
perl filter_polyA_v1.0.pl filename.sfq
PolyA filtered files (.polyA.sfq)
~ Check no. of reads (wc/4)
~ Forward and reverse
mismatched
~ Compare against raw number
perl
generate_paired_sets4maq.pl
filename.polyA.sfq
Paired Sequences
(.paired.polyA.sfq)
~ Check no. of reads (wc/4)
~ Forwards / backwards
reads match
~ Compare against raw number
perl maq_fastq_split_chunk.pl
filename.polyA.sfq
Paired Chunked files
(chunkX.paired.polyA.sfq)
2 Million Sequences per chunk
maq fastq2bfq
Paired Chunked Binary files
(chunkX.bfq)
perl -e 'for $element (1..X){system
"/directory/maq-0.6.6/./maq map
laneX.chunk$element.map
/directory/whole_human_genome.bf
q forwardsequence.chunk$element
reversesequence.chunk$element 2>
mapping.chunk$element.log &"}'
Mapped Files
(chunkX.map)
Log files
(chunkX.log)
~Mapped Sequences
~Mapped Paired Sequences
~Compare against raw number
tail laneX.mapping.chunkY

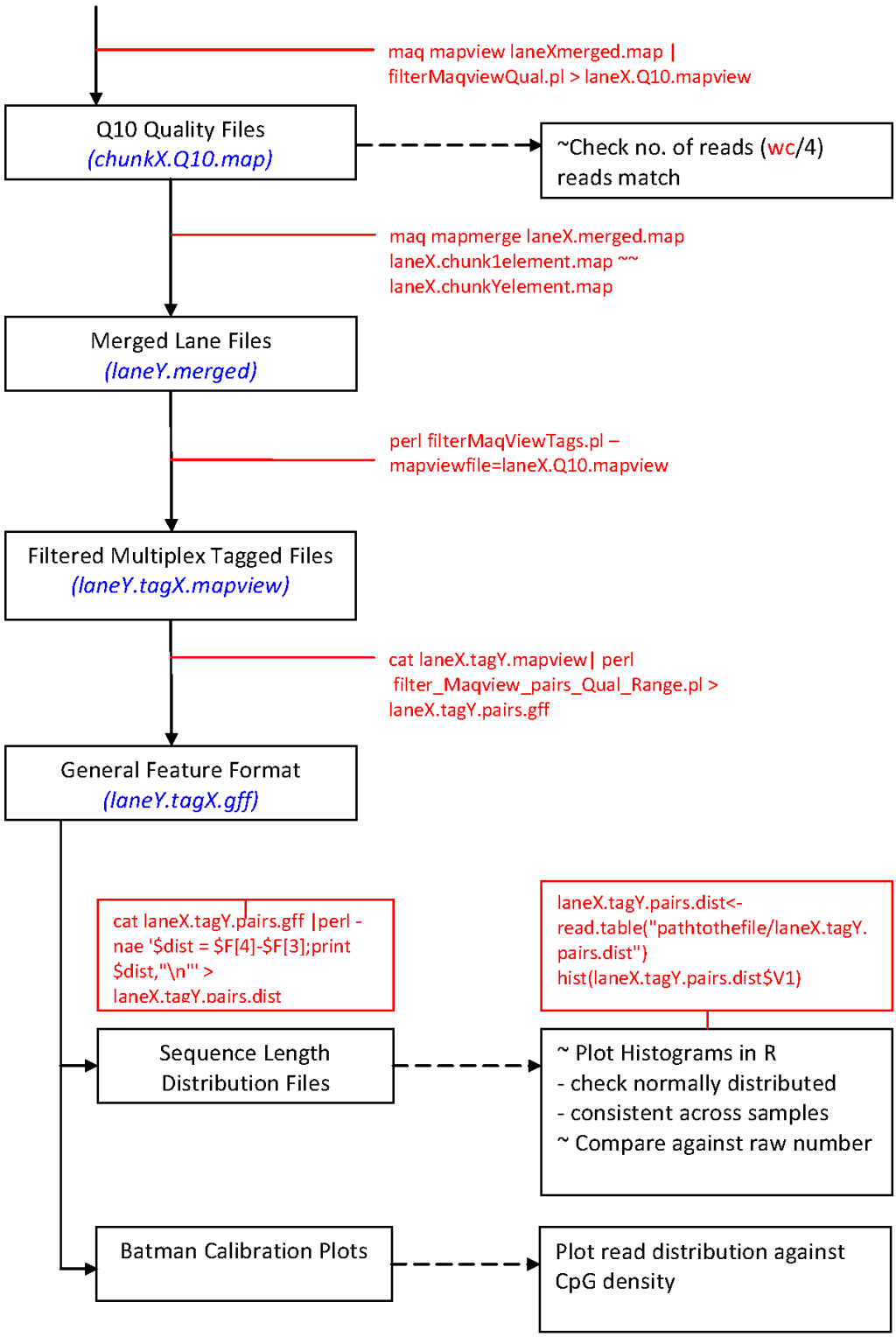


Figure 4c. Summary of bioinformatic processing steps to convert raw data into sequence data for analysis.

4.3.4 Bioinformatic analysis

Following these quality control steps, Thomas Down's DMR caller (as described in section 2.5.3.2.2) was used to identify differential methylation in cases versus controls.

4.3.5 Short Tandem Repeat (STR) panel analysis

In this experiment, additional techniques were used to interpret the results of the DMR call. These experiments applied STR panels to the original DNA samples to try and identify whether there was any sample contamination. STR panel experiments were performed by Denise Syndercombe-Court and David Ballard (Forensic DNA laboratory, Blizard Institute) using the Promega PowerPlex 16 HS System. This system is commonly used in forensic and paternity testing and includes sex-specific STRs in the 16 markers it targets. One such sex-specific marker is targeted to the amelogenin gene, present on both X and Y chromosomes in different quantity, thereby allowing distinction between males and females by the presence of one or two peaks (respectively). The autosomal markers on this panel include CSF1PO, D13S317, D16S539, D18S51, D19S433.

4.4 Results (Experiment 1)

4.4.1 Medip-seq library preparation

Medip-seq libraries were prepared for each sample individually using the protocol described above and with the appropriate checks of sample concentration and quality control during the experiment. DNA samples did not undergo any significant change in concentration (as tested by nanodrop) between India and the UK, as seen in table 4b.

Sample code	Group	ng/ul (India)	ng/ul (UK)
30	Case (low B12 and high folate)	65.99	81.6
33		81.37	113.5
533		149.73	197.3
621		94.64	102.4
637		61.5	53.7
687		84.92	109.5
25	Control (high B12 and low folate)	72.71	122.9
183		50.36	76.2
205		93.29	97.7
211		86.55	78.8
270		129.58	132.0
562		55.49	60.8

Table 4b. Summary of samples and DNA concentrations used in Medip-seq (experiment 1)

After careful optimisation and practice of the library preparation protocol, samples were processed in two batches of six, containing equal numbers of cases and control selected at random to avoid any experimental bias.

4.4.2 Test of Medip enrichment using qPCR

As described in 2.3.1.1, a qPCR is performed to test Medip enrichment efficiency using primers to stable regions of the genome that are known to be either fully methylated or unmethylated. These regions were defined by the Human Epigenome Project and have had primers designed to them by V Rakyan. Three fully methylated regions (and therefore enriched after Medip) and one unmethylated region (i.e. should not enrich after Medip) were used in the qPCR described in detail in the methods section. Each reaction (per sample and amplicon) is performed in duplicate and the mean Ct value is plotted.

	6583			11851			4994			8804		
	Ct duplicates		**mean Ct**	Ct duplicates		**mean Ct**	Ct duplicates		**mean Ct**	Ct duplicates		**mean Ct**
Medip	24.2	24.4	**24.3**	25.2	25.3	**25.25**	23.7	24.4	**24.05**	31.1	31.1	**31.1**
input	25.5	25.8	**25.65**	26.3	26.5	**26.4**	25.5	25.6	**25.55**	25.6	25.8	**25.7**

Table 4c. Mean Ct values (from duplicates) for sample 5 qPCR. Methylated primers: 6583, 11851, 4994, unmethylated primer: 8804

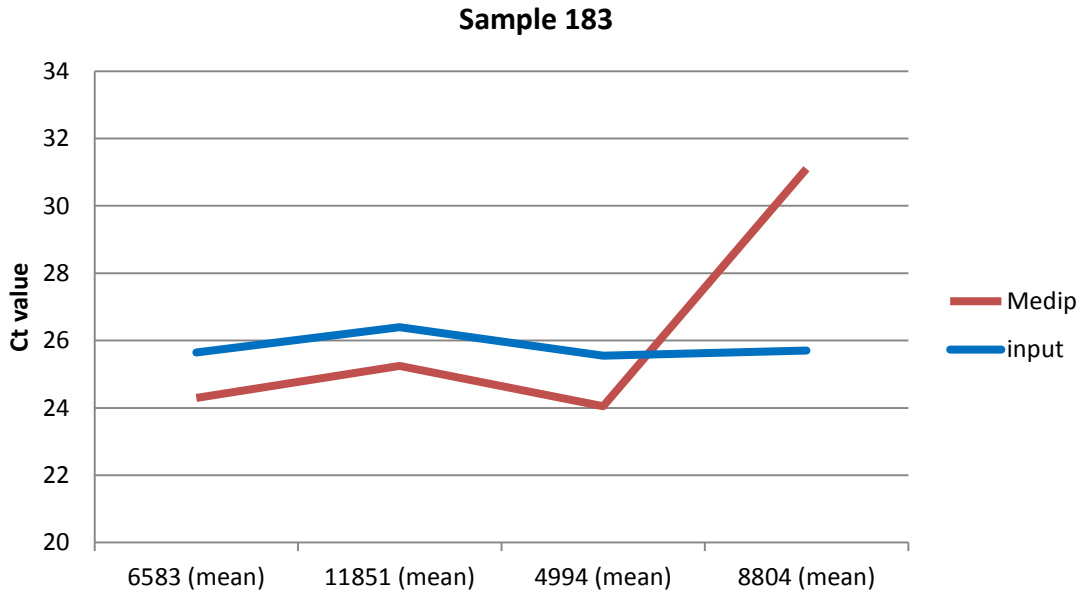


Figure 4d. Plot of qPCR results for sample number 5, showing successful MeDIP enrichment. Successful enrichment is determined by the visible fold change required to amplify the MeDIP DNA fraction using an unmethylated primer (8804), compared to methylated primers (6583, 11851, 4994).

As is seen in table 4c and figure 4d, the Ct values of the Medip samples are similar across the amplicons targeted at known methylated regions (6583, 11851, 4994). In the reaction using an amplicon targeted to a known unmethylated region (8804), amplification of the Medip DNA has a much greater Ct value, reflecting a greater number of cycles to amplify the DNA that has been successfully enriched for methylation by Medip. This Ct value, and the fold change between it and the 3 previous targets, provides evidence that the Medip enrichment has been successful. Furthermore, the input DNA (i.e. total DNA) shows equivalent Ct values across all 4 amplicons as it has not been enriched for DNA methylation and therefore amplification of methylated and unmethylated regions are equally possible. The final observation in this plot is that the input Ct values are slightly higher than those where Medip DNA is amplified with

primers to known methylated regions. This is due to the presence of both methylated and unmethylated DNA in the input fraction and therefore the need for a greater number of qPCR reaction cycles to amplify just one region alone.
This qPCR was performed for all samples prior to further steps in this experimental protocol. Although fold change can be calculated to ensure the success of Medip enrichment, it was deemed unnecessary to quantify this and inspection of the plots was sufficient to proceed to the next stage of the experiment.

4.4.3 Pooling of samples

Sample fragment size distribution and concentration was determined for individual samples using a bioanalyser to guide pooling of samples to ensure that equal quantities of each sample were used. Pooled samples were then run in two lanes of an agarose gel, and a band 250-300bp was cut out to size select fragments. DNA was extracted from the gel, forming the completed size-selected, pooled sample ready for sequencing (figure 5).

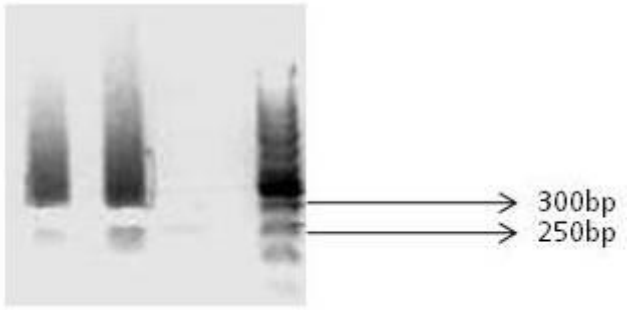


Figure 4e. Agarose gel showing DNA bands cut out after size selection.

4.4.4 Illumina GAIIx sequencing

The standard sequencing protocols were followed and met all routine QC checks in the process. Raw sequence read counts seemed consistent with successful sequencing and were interpreted in more detail during the bioinformatic workflow.

4.4.5 Bioinformatic analysis

Bioinformatic analysis was performed by Chris Mathews, with guidance from Guillermo Carbajosa and Thomas Down.

4.4.5.1 Sequence reads pre-processing

Sequence reads were filtered by multiplex tag so that their relative proportions at each stage of the pre-processing could be determined from within the pooled sample run (table 4d).

Tag	Tag sequence	Sample	Raw reads	Raw (% of total reads)	Q10	% reads maintained after Q10 filter
Tag 1	ATCACG	30	5578823	9.6	1654321	29.7
Tag 2	CGATGT	33	6022974	10.4	2361672	39.2
Tag 3	TTAGGC	533	4536517	7.8	3296593	72.7
Tag 4	TGACCA	621	5092573	8.8	2034835	40.0
Tag 5	ACAGTG	637	3514627	6.1	2780752	79.1
Tag 6	GCCAAT	687	4170958	7.2	2539663	60.9
Tag 7	CAGATC	25	6445443	11.1	2507863	38.9
Tag 8	ACTTGA	183	4210016	7.3	2100440	49.9
Tag 9	GATCAG	205	5343653	9.2	2803444	52.5
Tag 10	TAGCTT	211	4008342	6.9	3031010	75.6
Tag 11	GGCTAC	270	3543938	6.1	2275514	64.2
Tag 12	CTTGTA	562	5549317	9.6	2084369	37.6
		Total	**58017181**	**100**	**29470476**	**50.8**

Table 4d. Number of Medip-seq sequence reads at each stage of processing, pre-bioinformatic analysis in experiment 1.

As can be seen from table 4d, there is considerable variation between samples in proportion of reads that passed the Q10 filter stage. The possible explanation for this is variation in Medip enrichment due to different fragment sizes, however there was no clear correlation between longer fragment sizes and higher quality mapping when comparing the bioanalyser data with the plot above. Furthermore, the size distribution of fragments was consistent across all samples (figure 4f is an example of one of these plots and shows the same distribution as all other samples) as the Medip-seq libraries were pooled and cut out of the same gel. Variation could occur in the relative proportion of each sample, and this could affect the proportion of differently sized fragments between samples. A final explanation could be variation in sequencing quality, but this is unlikely in a pooled sample set such as this.

Batman calibration plots were drawn for each sample (e.g. Figure 4g) to identify the number of mapped sequence reads against CpG density within the reference genome. These plots (exemplified using sample 33) show that the most MeDIP enrichment (and therefore mapped reads) occurs at genomic regions with 20-30% CpG density. This finding is consistent with published data describing that the majority of DNA methylation occurs at regions of low CpG density (177). Published data on Medip-seq shows a similar pattern in calibration plots, further supporting this finding (58).

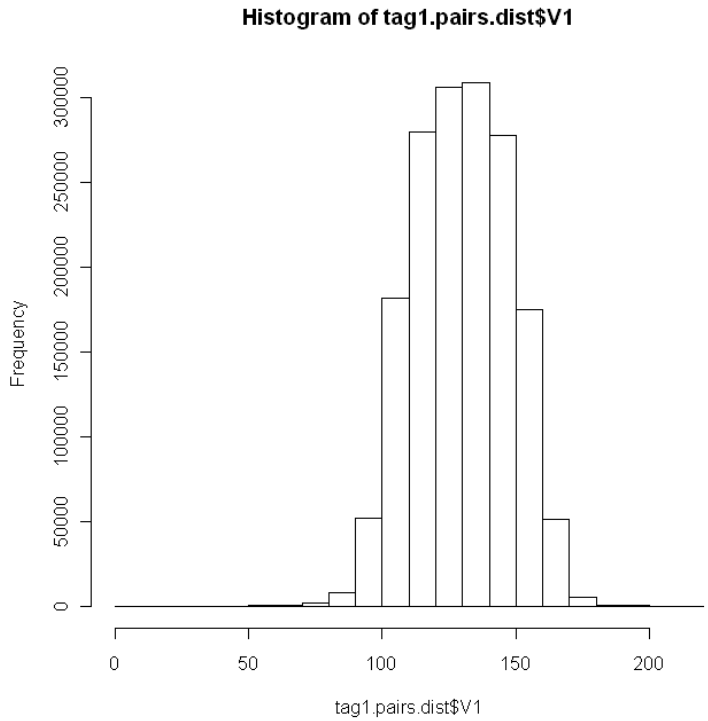


Figure 4f. Size distribution plot of raw sequence reads for sample number 30. The X axis shows the fragment size (bp) and Y axis frequency.

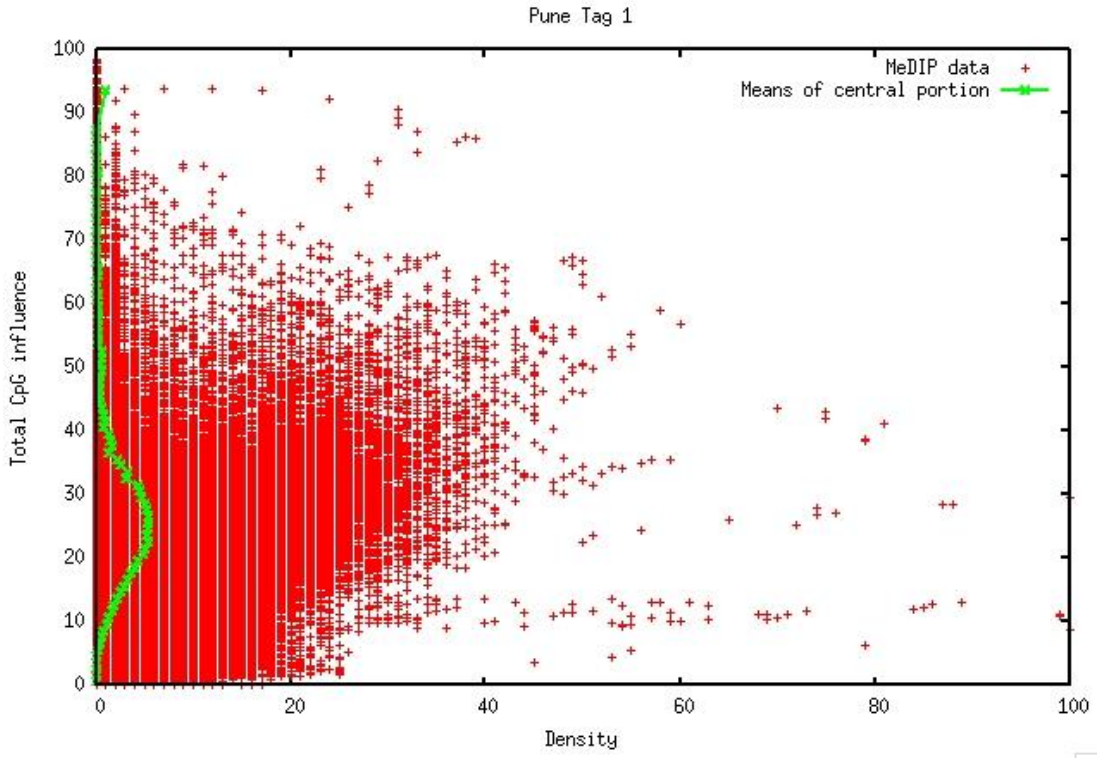


Figure 4g. Batman calibration plot. This plot shows the density of sequence reads (X axis) in relationship to CpG density (Y axis) for sample number 30.

4.4.5.2 DMR calling

DMRs were called using the Thomas Down DMR caller, and 2320 were identified. On first inspection of the DMRs that were generated, it was observed that a high proportion (20%) of these DMRs were located on the X chromosome. This number appeared to be disproportionate to other chromosomes, given our hypothesis that we would see genome-wide effects from this potential environmentally mediated programming event. The methylation data tracks from Medip-seq were loaded into the Ensembl genome browser for visualisation and 'sanity checking'. On close inspection of the data tracks, uploaded per multiplex tag, it was evident that two samples were driving most of the X chromosome DMRs (Figure 4h).

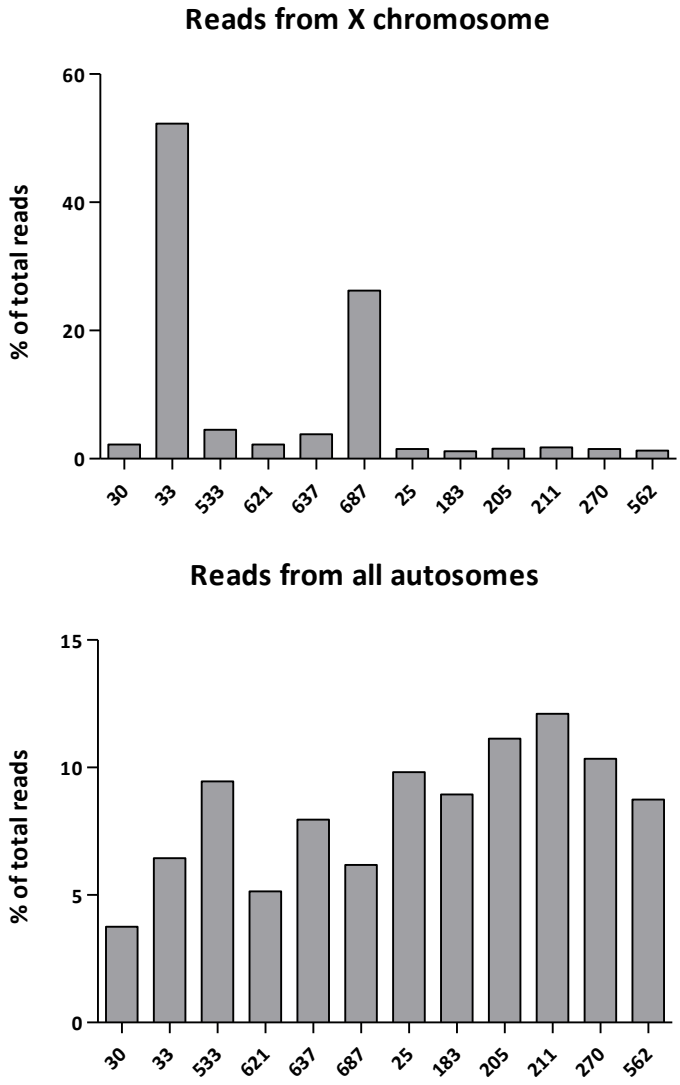


Figure 4h(i) and 4h(ii). Plots of sequence read counts (Y axis) per sample (X axis). Plot (i) shows the read counts from the X chromosome reads and (ii) shows read counts from autosomes.

These findings suggested a technical concern with the 2 samples driving these X chromosome differences, rather than a true biological process. A known cause of widespread X chromosome methylation differences is that between female and male X chromosomes due to the epigenetically mediated X inactivation in females. The most likely cause of the large and widespread X chromosome DMRs was thought to be from inclusion of female DNA in what was thought to be male DNA samples, either due to sample confusion or contamination. The quantitative estimates of methylation these X chromosome DMRs were also consistent with the evidence that X inactivation is associated with hypermethylation at gene promoters and hypomethylation at gene bodies (178).

4.4.6 STR panel analysis of samples

Due to the concern of sample contamination, we performed careful analysis of all samples with STR panels and confirmed that samples 2 and 8 contained a mixture of male and female DNA, the likely cause being sample contamination.

A summary of data from 4 STR markers applied to 4 of the experimental samples is shown in Table 5. Samples 33 and 687 are male but contaminated with female DNA, shown by the high peak height in the X allele marker (sample 33) and the low peak heights for the Y allele (samples 33 and 687) compared to the other male samples. The presence of 3 alleles in samples 33 and 687 with the D13S317 and D16S539 confirms the presence of DNA from a contaminant sample. Marker D18S51 does not show contamination and samples 621 and 183 did not display any evidence of contamination in these markers or the other 12 samples on the panel. The remaining 10 samples were also free of contaminant DNA. In the light of this confirmed sample contamination, we halted further analysis of the data from these Medip-seq experiments.

Sample name	Marker	Allele 1	Size 1	Height 1	Peak Area 1	Allele 2	Size 2	Height 2	Peak Area 2	Allele 3	Size 3	Height 3	Peak Area 3
33	**AMELOGENIN**	**X**	**107.71**	**5138**	**31220**	**Y**	**113.22**	**1122**	**6649**				
621	AMELOGENIN	X	107.66	3044	18598	Y	113.12	2976	17732				
687	**AMELOGENIN**	**X**	**107.63**	**3072**	**18114**	**Y**	**113.15**	**1315**	**7996**				
183	AMELOGENIN	X	107.74	3654	21707	Y	113.14	3187	18860				
positive control	AMELOGENIN	X	107.66	1833	11032								
water control	AMELOGENIN												
33	**D13S317**	**10**	**228.01**	**1293**	**7469**	**11**	**232.28**	**482**	**2880**	**13**	**240.58**	**1401**	**8241**
621	D13S317	9	223.72	866	5016	13	240.29	713	4370				
687	**D13S317**	**11**	**232.2**	**333**	**2068**	**12**	**236.4**	**384**	**2312**	**13**	**240.49**	**250**	**1552**
183	D13S317	9	223.65	835	5068	12	236.25	550	3318				
positive control	D13S317	11	232.08	5148	29821	11	232.08	5148	29821				
water control	D13S317												
33	**D16S539**	**11**	**278.62**	**1099**	**6717**	**12**	**282.34**	**351**	**2124**	**13**	**286.06**	**738**	**4744**
621	D16S539	8	267.3	1441	9004	8	267.3	1441	9004				
687	D16S539	9	271.06	219	1564	12	282.34	405	2720				
183	D16S539	10	274.87	716	4731	11	278.62	446	3033				
positive control	D16S539	11	278.59	2515	15333	12	282.39	2883	18141				
water control	D16S539												
33	D18S51	12	285.53	464	2699	14	293.39	484	3038				
621	D18S51	13	289.39	372	2431	13	289.39	372	2431				
687	D18S51	15	297.33	84	689	25	339.52	53	231				
183	D18S51	13.3	292.35	357	2505	20	318.38	104	761				
positive control	D18S51	15	297.27	2858	17466	19	314.04	2231	13927				
water control	D18S51												

Table 4e. Summary of 4 STR panel markers performed on 4 samples from experiment 1. Contaminated samples are highlighted in red.

4.5 Methods (Experiment 2)

4.5.1 Sample selection

New DNA samples were, taken from stored aliquots of blood that had not previously been used from our collaborators in India. These samples were put through 2 STR panel analyses – a 10 STR panel in Hyderabad, and the 16 STR forensic panel described in Experiment 1, both of which confirmed that the samples were uncontaminated.

We were unable to select single sex groups for analysis within our primary selection criteria (tertiles of B12 and folate; insulin resistance index) so we matched sex across groups as much as possible. During the time taken to obtain new samples, it had become evident that Illumina GAIIx sequencing was becoming more efficient and a greater number of sequencing reads were being processed per flowcell, thus we were able to include a larger sample number (8 vs. 8) in the forthcoming experiments.

Cases = Low B12 and high folate				
Sample ID	**Sex**	**Insulin Resistance**	**Age at sampling**	**Medip-seq ID**
470	Male	**0.75**	6 Yr	f_2_s_6
621	Male	**1.34**	6 Yr	f_1_s_6
637	Male	**1.48**	6 Yr	f_1_s_7
30	Male	**1.88**	6 Yr	f_1_s_1
533	Male	**1.9**	6 Yr	f_2_s_7
456	Female	**0.92**	6 Yr	f_2_s_5
622	Female	**1.98**	6 Yr	f_2_s_8
348	Female	**0.51**	6 Yr	f_1_s_3
Controls = High B12 and low folate				
Sample ID	**Sex**	**Insulin Resistance**	**Age at sampling**	**Medip-seq ID**
562	Male	**0.19**	6 Yr	f_1_s_5
317	Male	**0.32**	6 Yr	f_2_s_4
78	Male	**0.87**	6 Yr	f_2_s_3
43	Male	**0.4**	6 Yr	f_2_s_1
48	Female	**0.34**	6 Yr	f_2_s_2
417	Female	**0.37**	6 Yr	f_1_s_4
410	Female	**0.37**	6 Yr	f_1_s_8
330	Female	**0.41**	6 Yr	f_1_s_2

Table 4f. Pune samples used in Experiment 2, including colour codes that indicate subsequent sample pairing according to fragment size pairing (see figure 4i).

Sample collection and preparation was performed as per the methods in Chapter 2 and the colour coding of samples in table 4f relates to the fragment size pairing that will be described in the results section.

4.5.2 Medip-seq library preparation (non-multiplexed)

Samples were randomly assigned to one of two batches, each containing equal numbers of samples from group 1 and group 2 for library preparation. Multiplexing of samples was not possible as the technology is only suitable for 12 samples in one batch. It was decided that each samples would be run in a separate lane and that two flowcells would be used. PhIX control DNA (5%) was spiked into each lane, rather than being run as a full lane per flowcell. Medip-seq libraries were prepared as described in the methods section, but with no multiplex adapter used in the PCR amplification of the Medip-enriched libraries.

4.5.3 Size selection of samples

As a multiplexing strategy was not used, it was important to perform careful size selection of DNA fragments between samples, to ensure no bias in downstream analysis of samples from different sequence read lengths. Each completed Medip-seq library was split into two, and one half was run on a gel, and a band was cut out after UV-guided marking, aiming at 250-350bp. After DNA extraction from the gel band, samples were run on a Bioanalyser High Sensitivity DNA chip to determine the DNA fragment size. Samples were matched by size in case-control pairs, with a mean intra-pair difference of <5bp. In some cases, the second half of the Medip-seq library was run on a gel to re-extract size bands and to achieve optimal size matching.

4.5.4 Illumina GAIIx sequencing

Samples were re-randomised to run on 2 flowcells at the Babraham Institute, Cambridge, using standard reagents and protocols, running one sample per lane. This genomics facility was chosen in favour of our local facilities because of time constraints.

4.5.5 Sequence read processing

This was performed as per Experiment 1. Close attention was paid to the fragment lengths across cases and control samples as differences could result in variable Medip enrichment efficiency.

4.5.6 Bioinformatic analysis

Identifying case-control sample pairs with equivalent Medip enrichment efficiency reduced the possible confounding role of Medip-enrichment bias. Pairs were defined by drawing plots of enrichment efficiency in all possible case-control pair arrangements and then selecting the pairs with the most similar enrichment at low CpG densities (where the greatest read density is located).

Bioinformatic analysis incorporated combined DMR calling with the Thomas Down DMR caller and USeq. The Thomas Down DMR caller incorporated groupwise analyses (i.e. cases vs. controls) and pairwise analyses (i.e. case-control pairs matched for Medip enrichment efficiency).

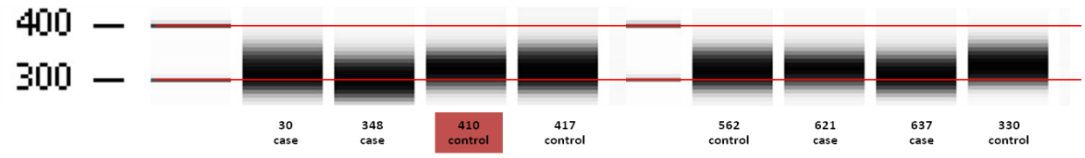
4.6 Results (Experiment 2)

4.6.1 Library preparation

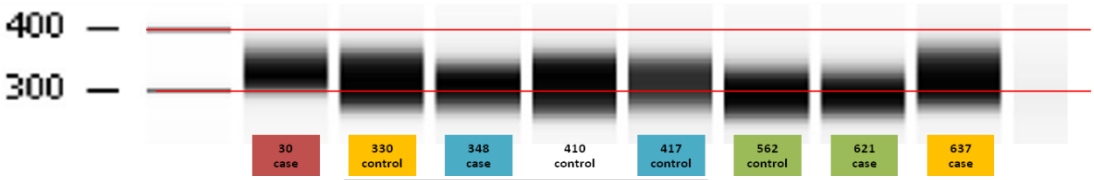
Library preparation of the two sample batches was performed. Case and control sample pairs were assigned with matched fragment sizes to one of two sequencing flowcells. Fragment sizes were approximated from the bioanalyser output, shown in figure 4i.

Flowcell 1

Gel 1 (8th May 2011)



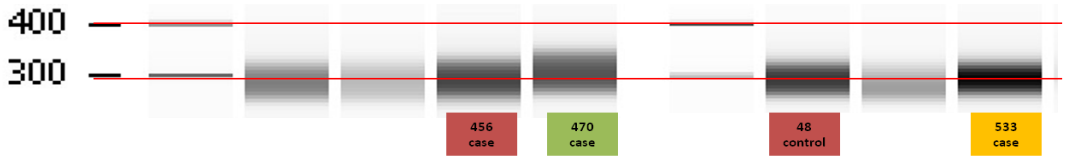
Gel 2 (9th May 2011)



Flowcell 2

Gel 1 (11th May 2011)

Gel 2 (12th May 2011)



Gel 3 (12th May 2011)

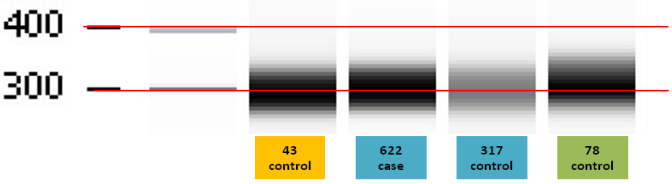


Figure 4i. Bioanalyser gel pictures of finished Medip-seq libraries. Samples were paired according to fragment size distribution seen in these gel pictures. Sample pairs are highlighted by matching colours within each flowcell.

4.6.2 Illumina GAIIx sequencing

The standard sequencing protocols were followed and met all routine QC checks in the process. Raw sequence read counts seemed consistent with successful sequencing and were interpreted in more detail during the bioinformatic workflow.

4.6.3 Bioinformatic analysis

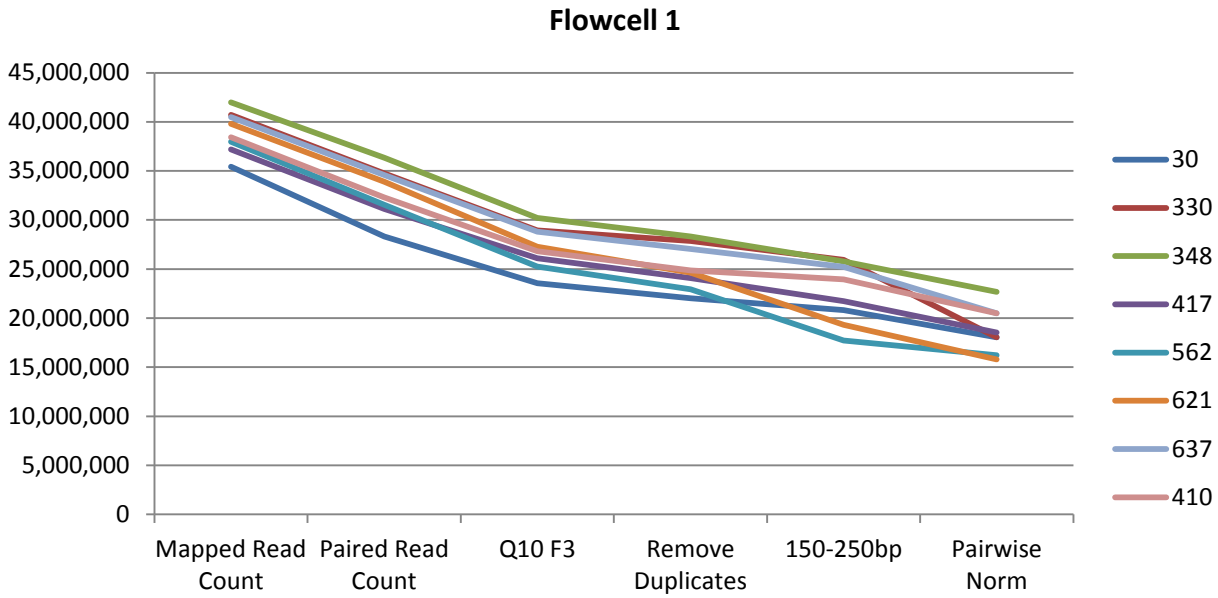
Chris Mathews performed the bioinformatic analyses with support from Guillermo Carbajosa and Thomas Down.

4.6.3.1 Sequence reads pre-processing

The standard sequence read processing steps were performed and showed predictable and comparable losses at each filtering step (table 4g and figure 4j). After mapping, pairing, quality filters were applied, as well as removal of PCR duplicates, all but one sample showed similar rates of attrition with over 60% of mapped reads suitable for analysis. Three samples (622, 456 and 43) on flowcell 2 that did not map as well as others performed well through the remaining quality filters resulting in a similar number of reads suitable for subsequent analysis. One sample (317) seemed to perform considerably less well than others through the quality filter, losing proportionately more reads at each QC step than other samples, resulting in only 50% high quality reads suitable for downstream analysis. This sample had been processed no differently to others and there was no suggestion from the various library preparation QCs (e.g. concentration checks, gel visualisation) that there was a particular problem with it. Furthermore, calibration and enrichment efficiency plots involving this sample did not highlight a specific cause for concern and therefore it was decided that it should remain in the analysis.

		Read length statistics (pre-processing)						Read processing						
Sample	Case/ control	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Mapped Read Count	Paired Read Count	% Pair	Q10 F3	%Q10	Remove Duplicates	% Passed
30	case	40	189	203	204	220	363	35,421,751	28,325,202	80	23,555,429	67	22,012,589	62.1
330	control	38	168	187	189	210	384	40,723,488	34,757,501	85	28,937,958	71	27,867,336	68.4
348	case	40	164	180	179	197	339	41,973,330	36,327,296	87	30,198,912	72	28,333,459	67.5
417	control	39	164	181	182	201	359	37,197,975	31,139,309	84	26,088,189	70	24,090,630	64.8
562	control	37	151	168	169	187	334	37,964,441	31,531,623	83	25,243,807	66	22,939,705	60.4
621	case	40	152	167	167	184	319	39,806,483	33,898,716	85	27,259,823	68	24,637,453	61.9
637	case	37	170	189	191	212	388	40,489,197	34,570,624	85	28,808,015	71	27,050,721	66.8
410	control	39	180	195	194	212	348	38,436,470	32,306,832	84	26,810,441	70	24,846,490	64.6
	Mean	**39**	**167**	**184**	**184**	**203**	**354**	**312,013,135**	**262,857,103**	**84**	**216,902,573**	**70**	**201,778,383**	**64.7**
43	control	37	153	168	168	186	327	29,113,826	27,496,959	94	22,806,397	78	19,456,002	66.8
48	control	40	153	168	167	183	308	41,938,761	39,574,711	94	32,030,655	76	26,924,591	64.2
78	control	40	159	175	176	194	350	41,178,009	38,150,640	93	31,305,264	76	27,235,074	66.1
317	control	40	154	171	171	190	332	40,820,099	37,558,129	92	29,866,853	73	20,910,454	51.2
456	case	41	149	165	165	183	327	34,460,691	28,232,679	82	27,032,753	78	23,590,031	68.5
470	case	39	159	177	177	196	347	40,345,590	36,798,432	91	29,693,253	74	24,923,691	61.8
533	case	37	154	168	167	182	302	39,600,334	36,637,299	93	28,799,156	73	24,506,535	61.9
622	case	39	158	173	174	191	328	33,525,971	31,004,329	92	25,759,611	77	23,437,304	69.9
	Mean	**39**	**155**	**171**	**171**	**188**	**328**	**300,983,281**	**275,453,178**	**92**	**227,293,940**	**76**	**190,983,680**	**63.5**
	Mean cases	**39**	**162**	**178**	**178**	**196**	**339**	**38202918**	**33224322**	**87**	**27638369**	**72**	**24811473**	**65**
	Mean controls	**39**	**160**	**177**	**177**	**195**	**343**	**38421634**	**34064463**	**89**	**27886195**	**73**	**24283785**	**63**

Table 4g. Medip-seq sequence read counts (n) per sample and per flowcell through each data pre-processing step.



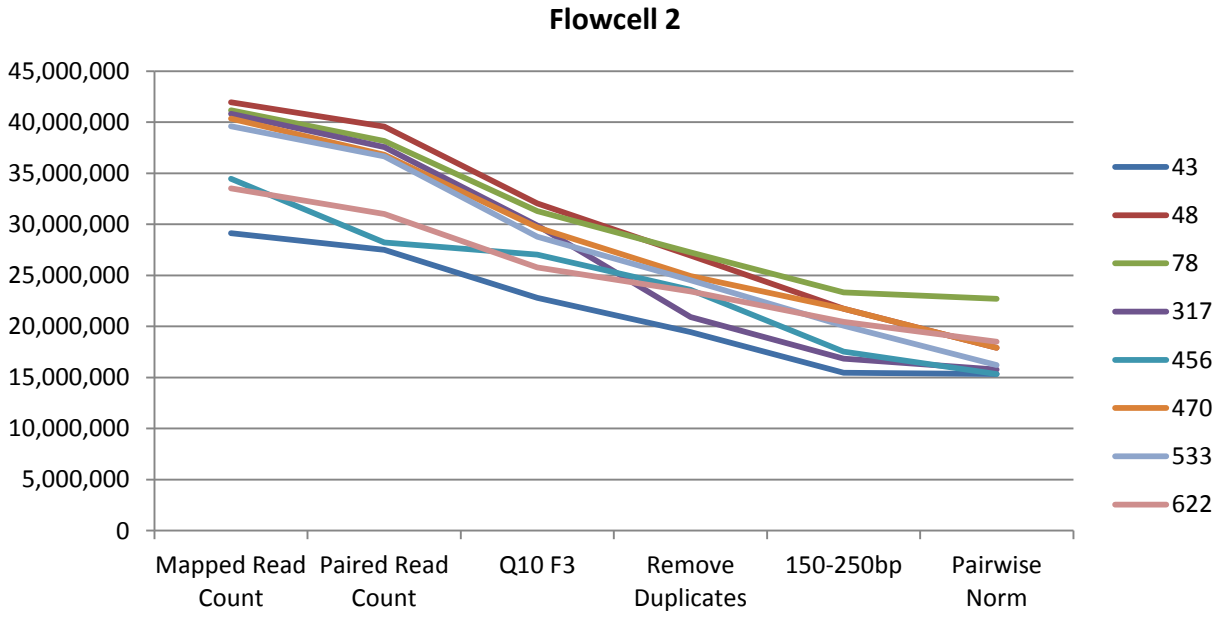


Figure 4j. Medip-sequence read counts per sample and per flowcell that passed each stage of pre-processing.

BATMAN calibration plots were drawn for all samples and showed appropriate patterns of read depth across regions of high and low CpG density. Prior to DMR analysis, reads from the X and Y chromosomes were filtered out due to the mixed sample groups precluding analysis of them. In addition, a final size filter was applied to remove reads that were outside the 150-250 bp size range and case and control samples were paired according to Medip enrichment efficiency (figure 4j). The read count in each size-matched pair was normalised. Previous sample arrangements resulted in matched numbers of male and female samples as far as possible. The result of these careful normalisation steps was that of size-matched and enrichment-matched case and control groups with equal numbers of reads, suitable for DMR calling (see table 4h).

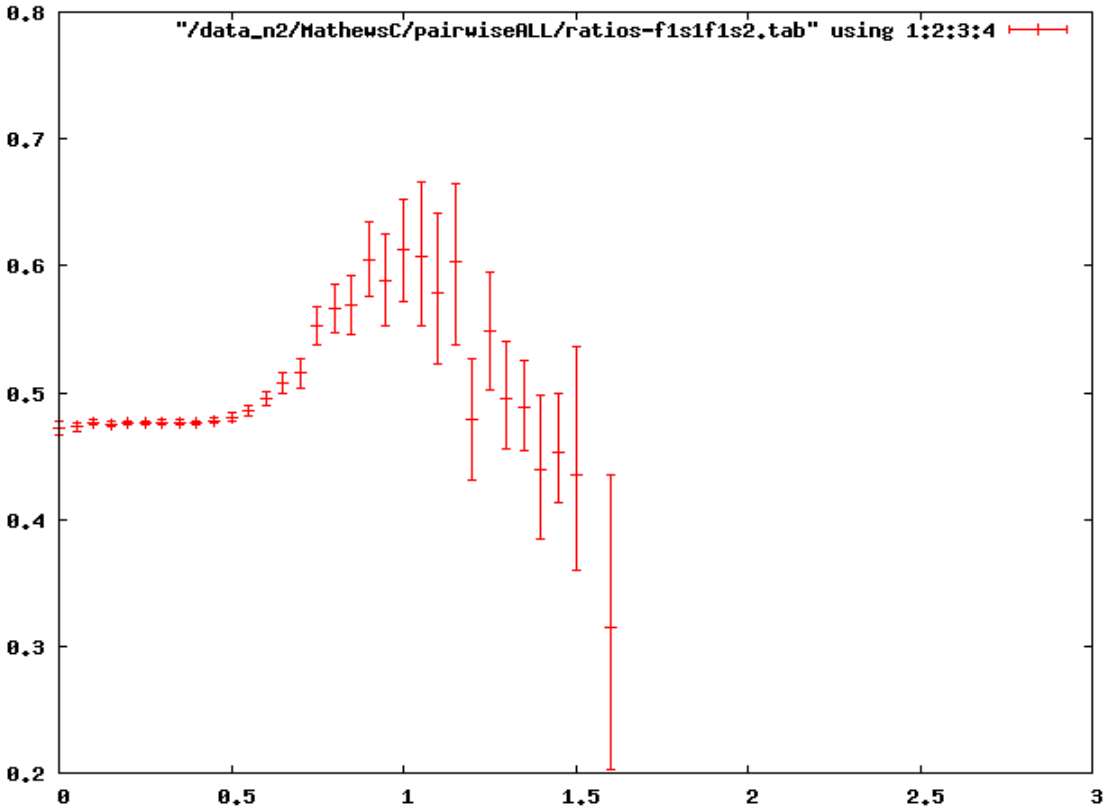


Figure 4k. Medip enrichment sample matching. This figure shows sample 30 matched to sample 330 with equivalent enrichment of samples (estimated methylation of sample 1:sample 2 on Y axis) at regions that have 60% CpG density or less (X axis). Variable enrichment is seen at regions of the genome that are have high CpG density due to the fact that these are far fewer (hence the wide error bars) and exhibit much more variable methylation.

Enrichment matched pair	Read length (bp)	Normalised reads (n)	**Case**	**Control**	Normalised reads (n)	Read length(bp)
1	204	18,052,736	**30***	**330***	18,052,736	189
2	179	22,690,640	**348***	**78**	22,690,640	176
3	167	15,776,059	**621***	**317**	15,776,059	171
4	191	20,502,217	**637***	**410***	20,502,217	194
5	165	15,339,645	**456**	**43**	15,339,645	168
6	177	17,918,483	**470**	**48**	17,918,483	167
7	167	16,205,735	**533**	**562***	16,205,735	169
8	174	18,518,541	**622**	**417***	18,518,541	182
	1424	**145,004,056**	Total	Total	**145,004,056**	1416
	178	**18,125,507**	Average	Average	**18,125,507**	177

Table 4h. Normalised and fragment-size matched case-control pairs with read counts, ready for DMR calling.
*= flowcell 1, and highlighted colours indicate male (blue) and female (pink).

4.6.3.2 DMR calling

DMR calling of was performed in a systematic fashion using a combination of bioinformatic techniques that have been discussed in the methods section.

The Thomas Down DMR caller was used to generate a long list of DMRs, first in a group-wise (case-control) analysis, and subsequently in a pair-wise analysis using the fragment size and Medip-enrichment normalised sample pairs in table 9. Following this DMR call, USeq was applied to the processed and normalised data in a new DMR call.

	DMR caller	Method	Size Normalisation	Enrichment matched	Test used	DMRs Called
Single method	Thomas Down CaseControl	Pairwise Caller (pooled case and control groups)	Yes	No	T-test	**30,000**
	Thomas Down Pairwise	Pairwise Caller (individual case/control samples paired using enrichment plots). DMRs common to all pairs.	Yes	Yes	T-test	**16,000**
	Useq	Useq	No	No	Negative Binomial distribution	**1800**
Combined methods	Final Combination	Thomas Down CaseControl + Pairwise + Useq	n/a	n/a	n/a	**48**

Table 4i. DMR calling strategies and numbers of DMRs

As could be predicted, the most DMRs were identified in a group-wise comparison between cases and controls using the Thomas Down caller, and this number (~30,000) was reduced considerably when differences that existed between single case-control pairs were called (~16,000). The groupwise call will have picked up differences that are driven by single samples and has no capacity for eliminating these potentially spurious DMRs, e.g. via a sensitivity analysis. The pairwise call is likely to call more differences when comparing two samples, but by combining the results of all pairwise calls and selecting those which are common to all, the number of differences is considerably reduced. The USeq call generated the least number of DMRs (1800) due to its control of individual false discovery rates at each DMR. When the DMRs identified by all 3 methods were compared, only 48 were common to all, and these have been inspected and analysed in further detail.

4.6.3.3 Sanity-checking of DMRs

DMRs were uploaded as custom tracks to the UCSC genome browser and inspected visually across individual samples and are available via the following URL: *http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=Chris%20Mathews&hgS_otherUserSessionName=Pune*. An example of this upload is shown in figure 4l with individual samples represented as individual tracks. This visual assessment did not suggest any technical issues, such as all DMRs being driven by specific samples, as had previously been noted when sample contamination had occurred. The UCSC upload was then used for further characterisation of the Medip-seq DMRs.

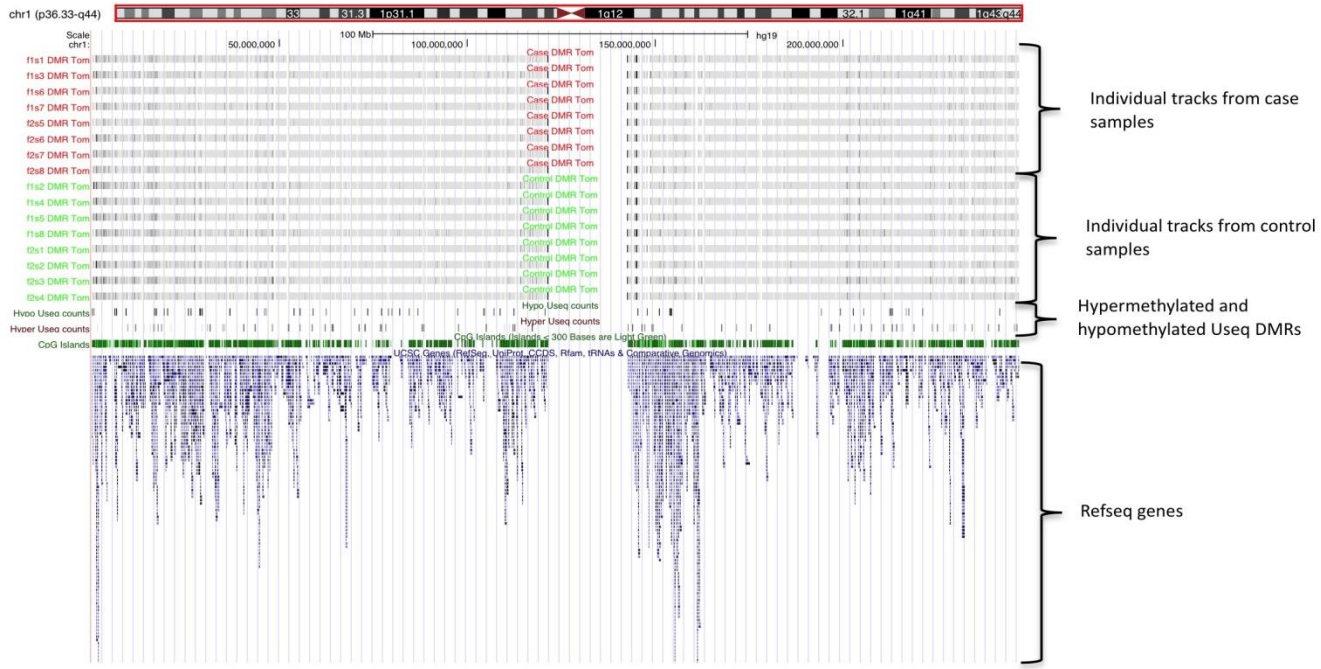


Figure 4l. Example UCSC upload of Medip-seq data used for sanity checking of DMRs.

4.6.3.4 DMR characteristics

The 44 DMRs identified by the combined DMR call were studied in more detail using the data generated from the DMR callers and the UCSC upload. Of the 48 DMRs, 27 were hypermethylated and 21 were hypomethylated in cases compared to controls. The DMRs identified are shown in Table 4m.

Thomas Down DMR caller							Useq		Genomic features		
Chromosome	**Start BP**	**Finish BP**	**P-Value**	**t-stat**	**DMR**	**# Windows**	**FDR**	**Log2**	**Gene**	**Location**	**Gene details**
chr1	**1596601**	**1597200**	0.00002	4.88	hyper	2	5.97	0.84	***CDC2L1/CDK11,* *SLC35E2B* and multiple other transcripts**	intron, also 3'UTR of other transcripts	Cell division cycle 2-like 1. p34Cdc2 kinase family essential for eukaryotic cell cycle control.
chr1	**58692501**	**58693500**	0.00010	4.62	hyper	6	3.59	1.54	***DAB1***	intron, also 3'UTR of other transcripts	Disabled homolog 1 (drosophila) - involved in cerebral cortex development
chr1	**117299101**	**117299700**	0.00001	5.18	hyper	2	5.07	1.77	***CD2***	intron 2	Surface antigen of human T-lymphocyte lineage, expressed on all peripheral blood T cells.
chr1	**149160501**	**149162600**	0.00018	4.70	hyper	16	10.98	1.20	**no**	repeat-rich region	
chr1	**167704301**	**167705100**	0.00038	3.71	hyper	4	5.89	1.64	***MPZL1*, 20kb downstream of *ADCY10***	intron 1	Myelin protein zero-like 1: cell surface protein involved in signal transduction via tyrosine phosphorylation, widely expressed including in haematopoetic cells and fetal liver
chr1	**175662701**	**175663300**	0.00004	4.64	hyper	2	11.93	1.66	***TNR***	intron 1	Tenascin-R: extracellular matrix protein expressed primarily in CNS and in other tissues in different times of development
chr2	**230519501**	**230520100**	0.00016	4.06	hyper	2	4.89	1.58	***DNER*, 50kb downstream ofTRIP12**	intron 1	Delta & notch-like epidermal growth factor-related receptor: present on cell membrane. Mediates neuron-glia interaction
chr3	**77303401**	**77304300**	0.00024	4.28	hyper	5	9.29	1.55	***ROBO2***	intron 2	Roundabout axon guidance receptor, homolog 2: (immunoglobulin superfamily) encodes for protein SLIT2, functions in axon guidance.
chr3	**162413401**	**162414000**	0.00012	4.12	hyper	2	9.92	1.86	**no**		
chr3	**178411601**	**178412400**	0.00032	3.62	hyper	4	4.18	1.18	***KCNMB2***	intron 1	Potassium large conductance calcium-

											activated channel, subfamily M, beta member 2: a voltage- & calcium-sensitive potassium channel used in smooth muscle and neuronal excitability
chr3	**195084601**	**195085600**	0.00004	4.55	hyper	6	4.88	1.18	***ACAP2***	intron	ArfGAP with coiled-coil, ankyrin repeat and PH domains 2: mediate diverse protein-protein interactions
chr4	**47741501**	**47742400**	0.00003	4.59	hyper	5	6.16	1.76	***CORIN***	intron	Serine petidase: a transmembrane protein (trypsin superfamily). Encoded protein converts pro- to atrial natriuretic peptide – role in blood volume and pressure.
chr4	**105137701**	**105138400**	0.00004	4.61	hyper	3	3.45	1.38	**no**		
chr5	**93820001**	**93821000**	0.00025	3.91	hyper	6	4.15	1.20	***KIAA0825/C5orf36***	introns + exon	
chr5	**109881501**	**109882500**	0.00027	3.72	hyper	6	3.76	1.46	***TMEM232***	intron	Transmembrane protein 232
chr6	**9025101**	**9026300**	0.00020	5.67	hyper	8	19.04	1.12	**no**		
chr6	**99929001**	**99929900**	0.00012	4.19	hyper	5	5.05	1.42	***USP45***	intron	Ubiquitin-specific peptidase 45
chr8	**3945201**	**3946300**	0.00012	5.42	hyper	7	20.65	1.27	***CSMD1***	intron	Ubiquitin-specific peptidase 45
chr8	**36946101**	**36946600**	0.00001	5.19	hyper	1	4.03	1.44	**no**		
chr8	**43092301**	**43097700**	0.00001	24.20	hyper	49	32.87	0.85	**no**	repeat-rich region	
chr8	**79909701**	**79910700**	0.00017	4.40	hyper	6	8.26	1.81	**no**		
chr10	**9780501**	**9781100**	0.00001	4.90	hyper	2	9.92	1.86	**no**		
chr10	**48607301**	**48607900**	0.00003	4.58	hyper	2	5.47	1.74	**no**		
chr10	**131119601**	**131120600**	0.00005	4.87	hyper	6	8.11	1.00	**no, 100kb upstream of *MGMT***		
chr13	**47344301**	**47345000**	0.00030	3.81	hyper	3	14.66	1.97	**no, but 1kb upstream of *ESD* 3'UTR**		Adjacent to ESD, a gene encoding a serine hydrolase and used as a genetic marker for retinoblastoma and Wilson's disease
chr17	**26916601**	**26917100**	0.00003	4.55	hyper	1	3.22	1.17	***SPAG5***	intron	Sperm associated antigen 5. This gene encodes a protein associated with the mitotic spindle apparatus.

chr18	**60348901**	**60349700**	0.00028	3.92	hyper	4	3.93	1.60	**no**		
chr1	**26462301**	**26463200**	0.00017	4.20	hypo	5	8.97	-1.32	**no**		
chr1	**79073301**	**79074000**	0.00018	3.87	hypo	3	6.42	-1.58	**no, 10kb upstream of *ILI44L***		
chr2	**72486601**	**72487400**	0.00015	4.48	hypo	4	7.09	-1.69	***EXOC6B***	penultimate intron, 100kb from CYP26B1	Exocyst complex component 6B
chr2	**215564601**	**215565600**	0.00009	5.16	hypo	6	20.28	-1.88	**no**	likely hypervariable region	
chr3	**36865101**	**36866100**	0.00029	3.92	hypo	6	7.88	-1.51	**no, 1kb uptream of *TRANK1***		
chr3	**131406201**	**131407300**	0.00026	4.41	hypo	6	7.73	-1.75	***CPNE4***	intron	Copine IV. Calcium-dependent membrane-binding proteins may regulate molecular events at the interface of the cell membrane and cytoplasm
chr4	**4265601**	**4266600**	0.00034	3.60	hypo	6	7.58	-1.18	**no, 2kb upstream of *LYAR***		
chr4	**47949001**	**47950000**	0.00003	5.11	hypo	6	7.04	-1.03	***CNGA1***	intron	Cyclic nucleotide gated channel alpha 1. The protein encoded is involved in phototransduction. Defects in this gene are a cause of retinitis pigmentosa autosomal recessive (ARRP) disease
chr6	**35316301**	**35317200**	0.00002	4.95	hypo	5	8.21	-1.39	***PPARD***	intron 2	Peroxisome proliferator-activated receptor delta. Potent inhibitor of ligand-induced transcription activity of PPARa & PPARg gamma. Integrator of transcription repression and nuclear receptor signaling. Heterodimerises with RXR to regulate gene expression
chr6	**35740101**	**35740800**	0.00004	4.69	hypo	3	3.39	-1.29	**no, 5kb upstream of *Corf127***		

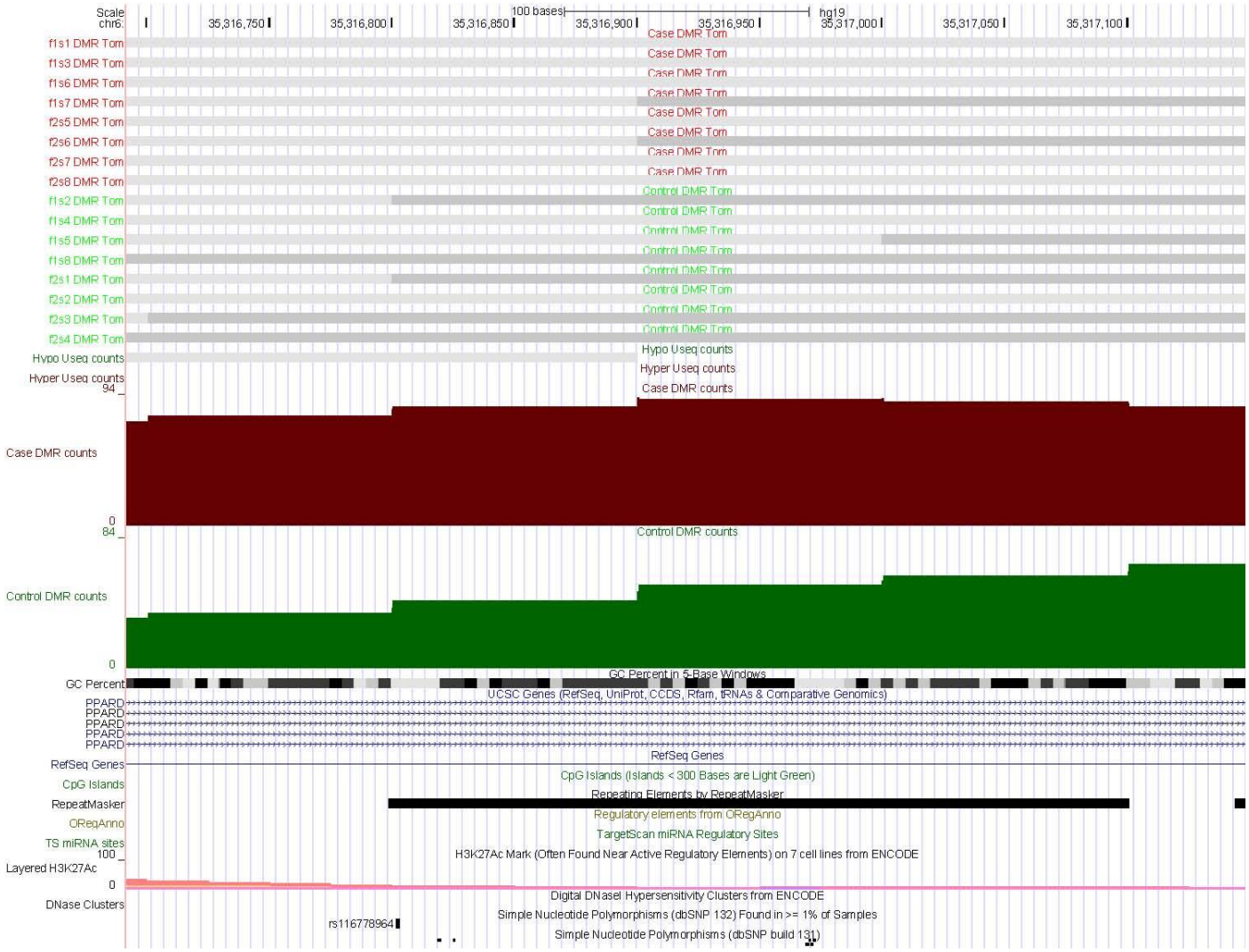
chr10	**91250801**	**91251600**	0.00019	3.86	hypo	4	5.41	-1.49	***SLC16A12***	intron 1	Solute carrier family 16, member 12. Encodes a transmembrane transporter with a role in monocarboxylic acid transport. A mutation is associated with juvenile cataracts with microcornea and renal glucosuria.
chr11	**24768401**	**24769100**	0.00010	4.04	hypo	3	3.38	-1.65	***LUZP2***	intron 4	Leucine zipper protein 2
chr13	**45665501**	**45666400**	0.00035	4.46	hypo	5	8.74	-1.41	**no**		
chr15	**31183601**	**31184600**	0.00019	4.47	hypo	6	4.45	-1.10	**no, 10kb upstream of *FAN1/MTMR15***		
chr15	**42095701**	**42096600**	0.00030	3.77	hypo	5	3.12	-1.30	***MAPKBP1***	intron 3	Mitogen-activated protein kinase binding protein 1
chr15	**102108801**	**102109300**	0.00005	4.28	hypo	1	4.08	-1.12	**no**		
chr16	**12739201**	**12740100**	0.00003	4.65	hypo	5	8.19	-1.39	**no, 10kb upstream of *CPPED1* 3'UTR**		
chr17	**7758101**	**7758800**	0.00008	4.35	hypo	3	10.56	-0.88	***TMEM88***	5'UTR	Transmembrane protein 88
chr17	**8247301**	**8248000**	0.00008	4.69	hypo	3	12.04	-1.27	***ODF4***	intron 1	Outer dense fiber of sperm tails. Encodes a protein that is localized in the outer dense fibers of the tails of mature sperm and is involved in sperm structure, sperm movement and general organization of cellular cytoskeleton.
chr17	**80105501**	**80107300**	0.00014	8.78	hypo	12	109.40	-2.17	***CCDC57***	intron	Coiled-coil domain containing 57
chr19	**54727901**	**54728900**	0.00035	3.63	hypo	4	4.46	-1.59	***LILRA6/LILRB3***	intron	Leukocyte immunoglobulin-like receptor, subfamily A

Table 4k. Summary table of Medip-seq DMRs identified using a combination of Thomas Down and USeq DMR calling methods. The table summarises the data derived from the Thomas Down caller in the first 7 columns.

A noticeable feature of many DMRs was that they overlap contiguous windows across genomic regions, giving a mean size of each DMR of 650 bp. Several DMRs crossed more than 1000bp and these tended to be in repeat-rich areas. Some of the DMRs are in regions of specific genetic interest to the programmed phenotype, e.g. *PPARD, LUZP2* (shown in figure 4m(i) and (ii)) and *CORIN*. Several DMRs overlie genes that have role in neuronal development/myelination the relevance of which may be supported by the knowledge that vitamin B12 deficiency is associated with nerve degeneration. However, other studies (92) have identified neural pathways as highly variable in their methylation status without such influences and therefore this may not have biological relevance. Table 4j summarises the genomic features at the DMRs called. The DMRs identified also appear to overlap some regulatory elements, as suggested by CTCF binding, DNase hypersensitivity, Pol2 binding sites and regions of enrichment of specific histone marks. The repeat-rich nature of many DMRs is consistent with the findings of other studies that suggest variable DNA methylation is associated with repeat elements (179).

Genomic features	**N**	**%**
Genic	14	29
Intergenic	34	71
Intronic (of 14 genic DMRs)	10	71
Conserved across species	28	58
CpG island	2	4
Repeat-rich	42	88
Overlies regulatory features and/or regions of histone enrichment	41	85

Table 4j. Summary of genomic features of the 48 DMRs identified through the combined DMR call.


Scale
chr6:
100 bases
hg19
35,316,750
35,316,800
35,316,850
35,316,900
35,316,950
35,317,000
35,317,050
35,317,100
f1s1 DMR Tom
f1s3 DMR Tom
f1s6 DMR Tom
f1s7 DMR Tom
f2s5 DMR Tom
f2s6 DMR Tom
f2s7 DMR Tom
f2s8 DMR Tom
f1s2 DMR Tom
f1s4 DMR Tom
f1s5 DMR Tom
f1s8 DMR Tom
f2s1 DMR Tom
f2s2 DMR Tom
f2s3 DMR Tom
f2s4 DMR Tom
Hypo Useq counts
Hyper Useq counts
Case DMR counts
Control DMR counts
GC Percent
PPARD
RefSeq Genes
CpG Islands
RepeatMasker
ORegAnno
TS miRNA sites
Layered H3K27Ac
DNase Clusters
rs116778964
Case DMR Tom
Control DMR Tom
Hypo Useq counts
Hyper Useq counts
Case DMR counts
94
0
Case DMR counts
84
0
Control DMR counts
GC Percent in 5-Base Windows
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
RefSeq Genes
CpG Islands (Islands < 300 Bases are Light Green)
Repeating Elements by RepeatMasker
Regulatory elements from ORegAnno
TargetScan miRNA Regulatory Sites
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
0
Digital DNaseI Hypersensitivity Clusters from ENCODE
Simple Nucleotide Polymorphisms (dbSNP 132) Found in >= 1% of Samples
Simple Nucleotide Polymorphisms (dbSNP build 131)

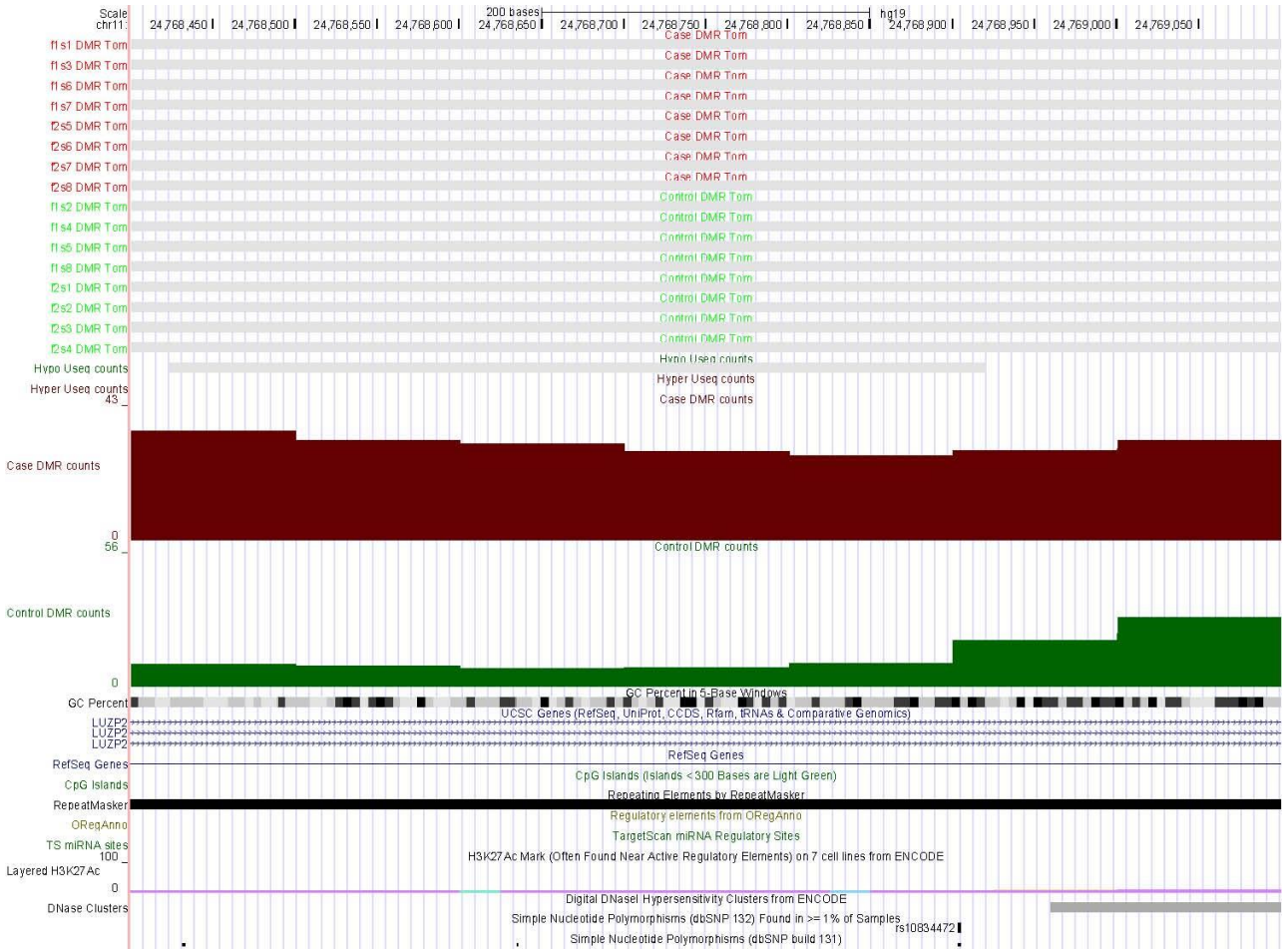


Figure 4m(i) and (ii). UCSC screenshots of Medip-seq DMRs at *PPARD* (i) and *LUZP2* (ii). Individual sample methylation tracks are shown in addition to Useq-derived DMR counts of case (dark red) and control (green) samples. Both DMRs are intronic, overlap repeat-rich areas, and contain SNPs within them. *LUZP2* overlaps a DNase hypersensitivity site .

4.7 Methods (Experiment 3)

4.7.1 Sample selection

DNA from all 10 case and 10 control samples were analysed in this experiment (see table 4a), designed as a technical validation of the Medip-seq data from Experiment 2. An unbalanced mixture of male and female samples was present across cases and controls.

4.7.2 Illumina 450k methylation array

The protocols for bisulphite conversion and qPCR test of conversion efficiency were followed. Array experiments were performed by UCL Genomics, within a randomised batch of containing samples from 2 other studies.

4.7.3 Data analysis

QC and analysis of data from the 450k array was performed in accordance with the methods described in the Methods Chapter. In view of the mixed sex samples, X and Y probes were removed via a bioinformatic filter.

4.7 Results (experiment 3)

4.7.1 Quality control checks

Data generated from the 450k array experiments was subjected to the standard quality control checks described in the methods section. Sample-independent QC (staining, extension, hybridisation and target-removal) showed good performance of the array at the control probes (data not shown). Sample-dependent QC (bisulphite conversion, specificity, negative controls and non-polymorphic controls) also showed good performance of the samples through the array experiments. The bisulphite conversion QC plots are shown in figure 4n and 4o and represent both colour channels at both Infinium bead types (I and II). In the Infinium I

probe QC plots (figure 4n), controls C1, C2 and C3 assess the performance of bisulphite-converted DNA and this is monitored in the green channel. The signal intensity produced from C1, C2 and C3 should be lower than their unconverted control sample counterparts (U1, U2, U3) due to difference in extension at the control probes. The same pattern should be observed in C4/U4, C5/U5 and C6/U6 control probes in the red channel. The Infinium II probe QC plots (figure 4o) are different due to the different probe design with a single probe per bead that hybridises both methylated and unmethylated sequence. In these plots, a successful bisulphite conversion is represented by higher intensities across all control probes (II1, II2, II3 and II4) in the red channel compared to the green channel.

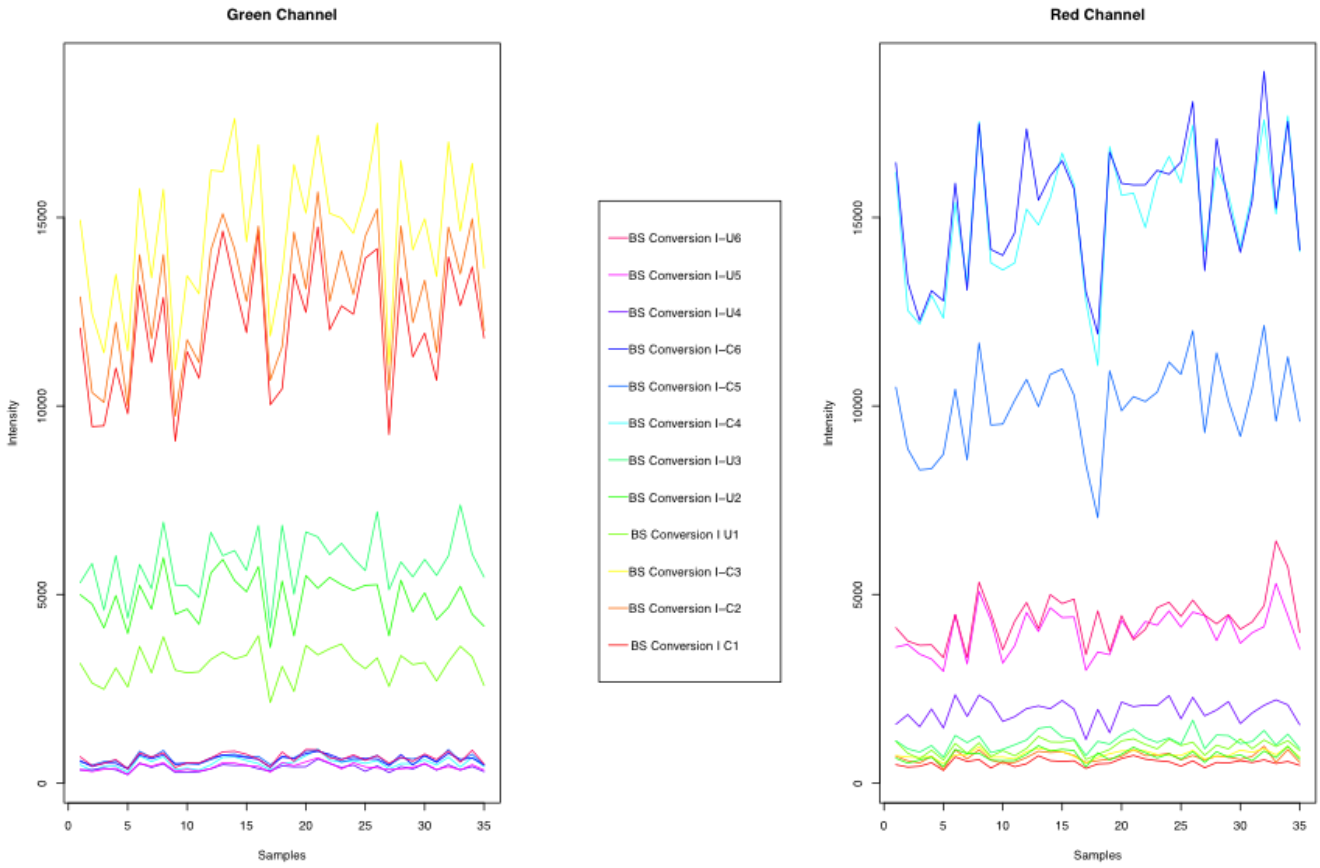


Figure 4n Infinium probe I bisulphite conversion QC plot, represented in both green and red array channels by control probes targeting converted sequence (C1-4) and control probes targeting unconverted sequence (U1-4).

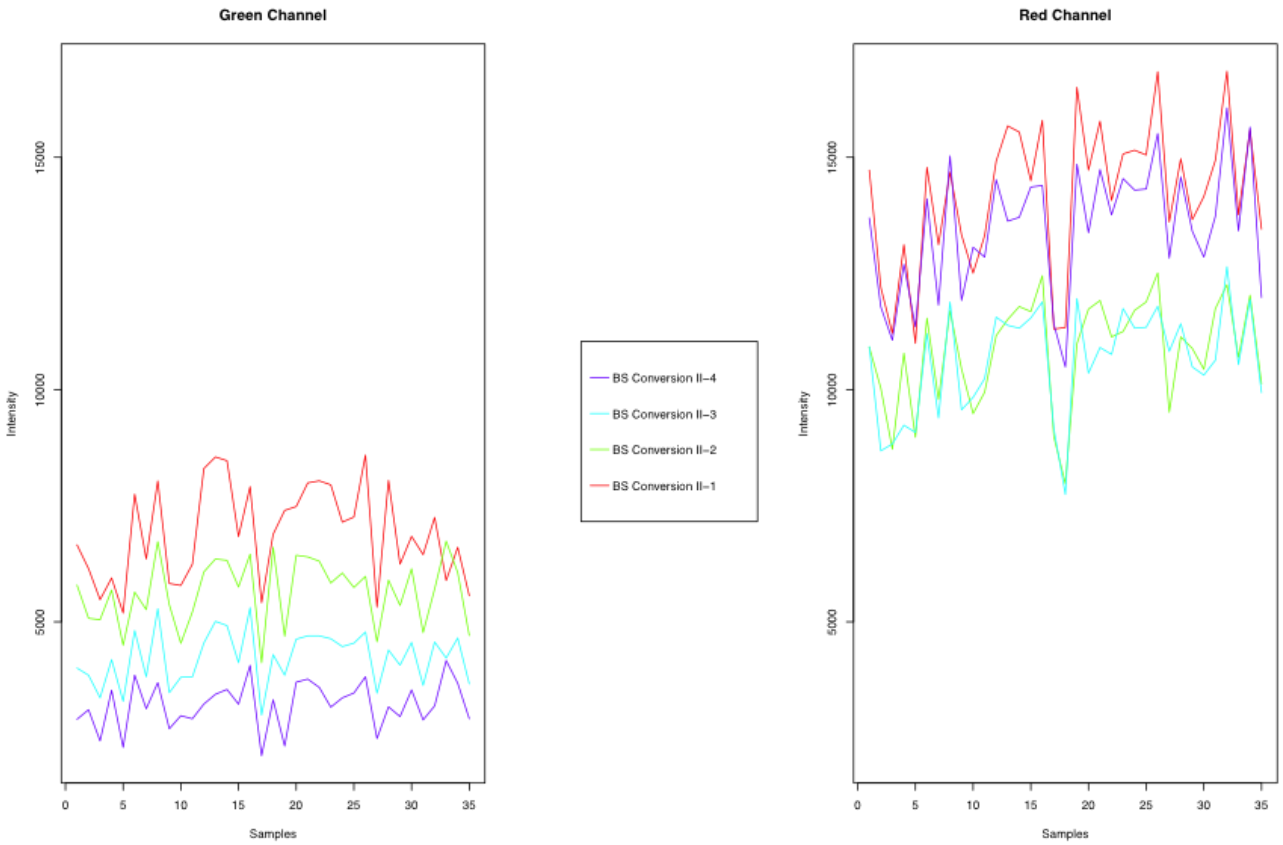


Figure 4o Infinium probe II bisulphite conversion QC plot, represented in both green and red array channels by 4 control probes (II1-4) that hybridise converted and unconverted DNA.

The .idat files generated from raw data represent the beta values following correction for background effects determined by the negative and non-polymorphic control probes. Figure 4p shows a typical bimodal distribution of beta values generated from these .idat files, detected at all array probes, showing that the highest density of probes detect methylation levels at close to 0% and 100%. This is a feature of the array design being targeted to CpG rich and poor genomic regions, but this is also partly representative of methylation across the entire genome. This plot shows uniformity of this distribution across all samples (both cases and controls) and is a useful means of detecting technical problems with individual samples.

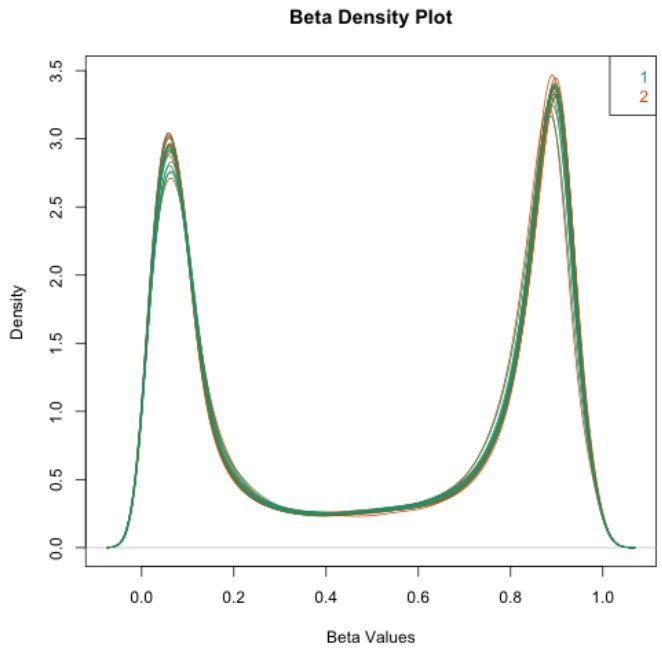


Figure 4p. Density plot of all array samples (cases = 1, controls =2) showing the density of varying beta values across all probes in a typical bimodal distribution

4.7.2 Illumina 450k methylation array data

Data filtering was performed on .idat files to remove X and Y probes and those displaying cross-hybridisation. The final number of probes used for subsequent analysis was 430,975 (out of 485,836). SNP-associated probes were not filtered. A multidimensional scaling plot (MDS) was generated using the beta values for each sample derived from the filtered .idat files (figure 4q). The MDS plot did not show any significant differences that could be attributable to technical bias, such as clustering of samples run on different arrays, nor did it show obvious gross biological differences that would be represented by clustering of cases and controls.

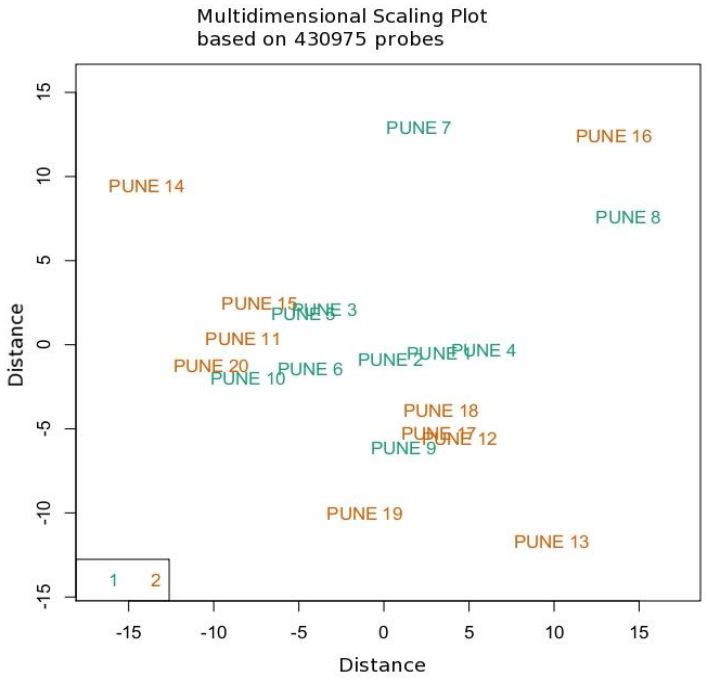


Figure 4q. MDS plot of beta values detected from 430975 probes across all samples (after removal of cross-hybridising probes and those failing to meet the detection p value threshold). Group 1 = cases, Group 2 = control samples.

4.7.3 MVP call

The R-based Limma package was used to identify differential methylation between cases and controls using an F-test applied to transformed beta values (M values). A total of 3963 MVPs were identified using this F-test, with varying statistical significance per MVP (see table 4l). FDR-adjustment of p values (as described in section 2.5.1.6) did not produce any statistically significant q values indicating that the MVPs identified in this dataset could include multiple false discoveries. A sensitivity analysis was performed on the MVPs identified to try and reduce the false positive rate by excluding those driven by just one sample. The sensitivity analysis reduced the total number of MVPs identified to 2045, 52%of the total and, as would be expected, these included the MVPs with stronger p values.

		Hypomethylated	Hypermethylated
MVPs from basic MVP call N = 3963	**Number (% total MVPs)**	2258 (57)	1705 (43)
	Range of Beta value differences	-0.3 to -46%	0.3 to 38%
	P value (unadjusted) range	$1x10^{-5}$ to $1x10^{-2}$	$5x10^{-6}$ to $1x10^{-1}$
	Number of SNP-containing probe (%)	484 (11)	361 (8)
	Genic (%)	1701 (76)	1307 (77)
	Promoter (% of genic MVPs)	465 (27)	434 (25)
	Intergenic (%)	547 (24)	398 (23)
	CpG Island (%)	872 (39)	705 (41)
	CpG Shore or Shelf (%)	706 (31)	495 (29)
	Enhancer (%)	424 (19)	330 (19)
	DNAse hypersensitivity site (%)	348 (15)	206 (12)
MVPs after sensitivity analysis N = 2045	**Number (% total MVPs)**	1175 (57)	870 (43)
	Range of Beta value differences	-0.3 to -46%	0.3 to 38%
	P value (unadjusted) range	$4x10^{-5}$ to $7x10^{-3}$	$5x10^{-6}$ to $7x10^{-3}$
	Number of SNP-containing probe (%)	243 (21)	194 (22)
	Genic (%)	890 (76)	657 (76)
	Promoter (% of genic MVPs)	247 (28)	434 (25)
	Intergenic (%)	285 (24)	213 (24)
	CpG Island (%)	490 (41)	387 (44)
	CpG Shore or Shelf (%)	343 (29)	237 (27)
	Enhancer (%)	241 (21)	169 (19)
	DNase hypersensitivity site (%)	184 (16)	111 (13)

Table 4l. Summary of MVPs from 450k dataset using Illumina annotation.

Subsequent analysis and description of MVPs will now include only those that have passed through the sensitivity analysis. There were more MVPs that were hypomethylated in cases compared to controls, and this may reflect a reduction in available methyl groups for methylation in the context of severe vitamin B12 deficiency. The array design is targeted to CpG islands, promoters and regulatory elements (e.g. enhancers and DNase hypersensitivity sites), and this was reflected by the location of MVPs. Just over 20% of MVPs identified were at probes known to be SNP-containing. At these probes, SNP-CpGs create or abrogate a site for methylation and therefore are identified as an MVP when the genetic polymorphism is different in cases vs. controls. This difference is likely to be a chance effect, and, should information on alleles be available, the study does not have the power to detect genetic association between cases and controls.

4.7.4 Pathway analysis

GO terms were generated using the GOstats add-on to the Limma analysis package.

GO: ID	Pvalue	Odds Ratio	Expected count	Observed count	Size	Term
GO:0032501	0.000001	1.38	356	427	4770	multicellular organismal process
GO:0032502	0.000006	1.37	283	345	3794	developmental process
GO:0009887	0.000018	1.74	49	79	661	organ morphogenesis
GO:0022008	0.000020	1.60	72	107	969	neurogenesis
GO:0030182	0.000029	1.63	63	95	844	neuron differentiation
GO:0048568	0.000037	2.17	20	39	266	embryonic organ development
GO:0051336	0.000047	1.67	53	82	712	regulation of hydrolase activity
GO:0065007	0.000070	1.30	571	628	7645	biological regulation
GO:0048468	0.000096	1.48	91	126	1224	cell development
GO:0048598	0.000152	1.82	31	52	415	embryonic morphogenesis
GO:0048519	0.000157	1.33	198	243	2650	negative regulation of biological process
GO:0035295	0.000165	1.87	27	47	365	tube development
GO:0051179	0.000172	1.29	286	336	3826	localization
GO:0009987	0.000287	1.46	767	801	10570	cellular process
GO:0035148	0.000304	2.82	7	18	98	tube formation
GO:0001843	0.000333	3.53	4	13	59	neural tube closure
GO:2000026	0.000342	1.51	67	94	891	regulation of multicellular organismal development
GO:0045596	0.000407	1.86	24	42	328	negative regulation of cell differentiation
GO:0048704	0.000419	3.24	5	14	68	embryonic skeletal system morphogenesis
GO:0060193	0.000448	2.49	9	21	127	positive regulation of lipase activity

Table 4m. Results of a Gene Ontology analysis, summarising the top 15 pathways overrepresented by 450k MVPs.

The GO analysis performed on the 450k data output shows and differential methylation affecting gene pathways involved in a range of developmental processes, and this could have functional effects via the interaction between DNA methylation and gene expression. The findings of this GO analysis are non-specific and need to be corroborated by a study in larger sample numbers, as well as correlation in other model systems, ideally including gene expression data. The representation of GO terms including neuron development and differentiation is similar to the GO analysis performed in the Matlab experiments (see section 5.3.3.2) and that published previously and their inclusion may represent the normal methylation variation associated with ubiquitously-expressed genes in these pathways, or a role in relevant and common biological processes such as aging (92). It is also interesting to see a pathway of neural tube closure overrepresented in this GO analysis, given the known effects of optimal folate and B12 status on reducing the prevalence of neural tube defects (180).

4.7.5 Medip-seq and 450k array technical validation

A comparison of Medip-seq DMRs and 450k array MVPs was performed by way of technical validation of these results. Medip-seq experiments had been performed on 8 cases and 8 controls and therefore, for this technical validation across experimental platforms, a new MVP call was performed using 450k array data using the same 8 cases vs. controls.

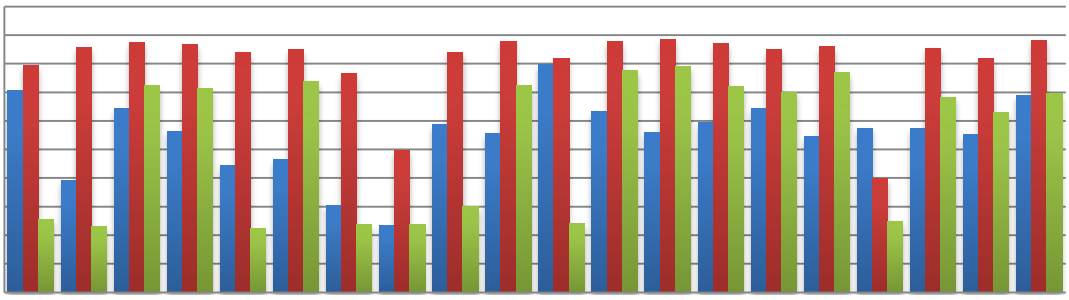
First, the 'top hit' list of 48 Medip-seq DMRs derived from the combined analysis algorithms were compared to 450k array coverage. Out of these 48 regions, only 4 had direct coverage with the 450k array and it was therefore decided that the entire USeq DMR list should be used for this technical validation. There were only 55 450k array probes that overlap the 1800 USeq DMRs called, suggesting that this is a poor means of validating one platform against the other. However, the 55 overlapping probes covered 24 Medip-seq DMRs (i.e. some DMRs had more than one probe mapping within them) suggesting that both methods can identify regions of significant variation. Of the 55 probes that overlapped Medip-seq DMRs, the majority (n = 32) showed differential methylation with the same directionality (i.e. hypomethylation or hypermethylation) between cases vs. controls in both datasets; this finding provides reassurance that the technical means of quantifying methylation are strong. Another 14 probes did not show statistically significant differential methylation, and only 9 showed conflicting directions of differential methylation between platforms.

Co-methylation was observed in the DMRs represented by multiple array probes and provides some reassurance of the biological plausibility of these validated DMRs. Co-methylation, a

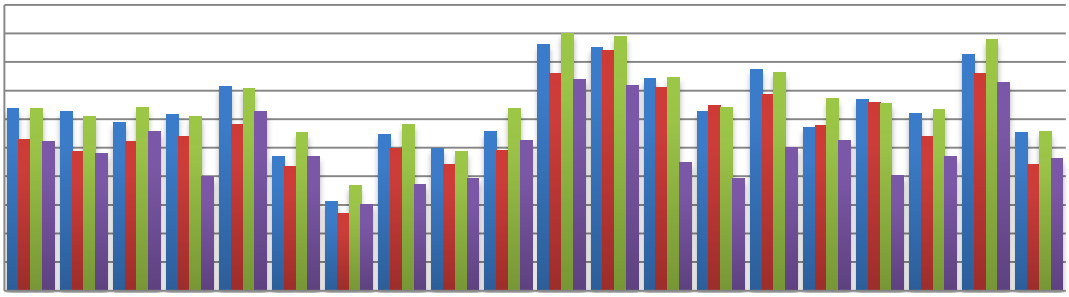
correlation of methylation status of adjacent CpGs that has been described across <1000bp regions of varying CpG density (181).

Of the 55 probes overlapping USeq DMRs, 24 contain annotated SNPs within the probe binding location and a large proportion of these are SNP CpGs. Six of the USeq DMRs that were represented by 3 or more 450k probes (i.e. those that well represent the methylation level across the wider Medip-seq DMR) are shown in graphical format below. Of these DMRs, DMR1 and DMR13 have methylation levels of near 0, 50 or 100% across the probes in all 20 samples, due to each probe being tiled to a SNP-CpG. This genetically driven methylation state is confirmed the presence of SNPs at each of the probes within these DMRs. DMRs 5, 7, and 10 all have at least one SNP-CpG probe within them and the beta values at non-SNP probes within the same DMRs appear to be in *cis*. DMR24 does not have any SNP-associated probes within it and consequently, there is more variation in beta values than would be expected if a SNP or *cis*-effect was present.

DMR1



DMR5



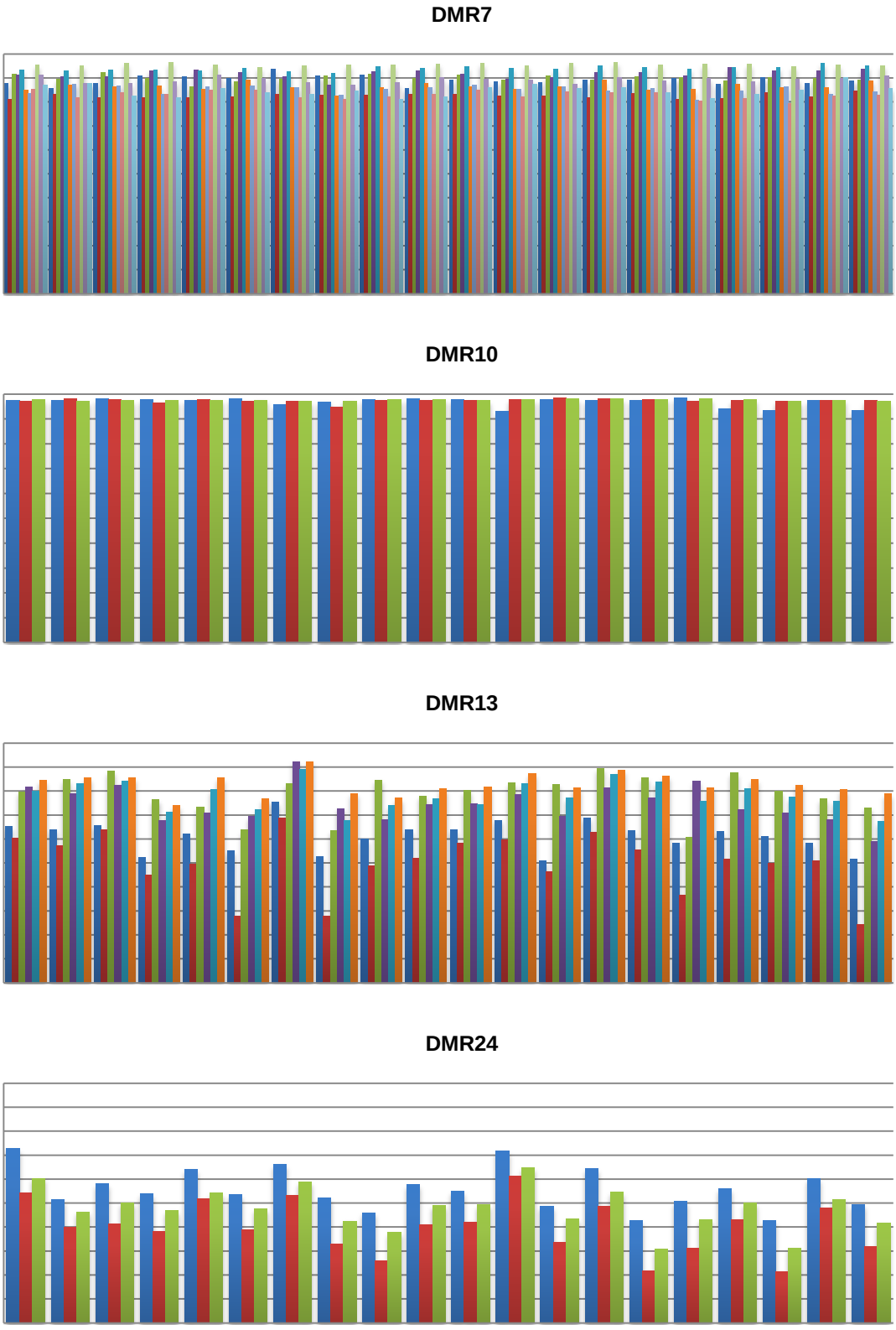


Figure 4r. DMRs represented by USeq and multiple overlapping 450k array probes. Legends show 450k probe names and (*) denotes the presence of a SNP at a 450k array probe.

4.8 Discussion

4.8.1 Discussion points related to the PMNS clinical study

a. Nature of the environmental exposure

Maternal vitamin B12 deficiency with folate repletion was defined as the environmental exposure associated with the adverse childhood phenotype of insulin resistance and obesity and was used to examine environment-epigenetic interactions. One of the strengths of this study was the careful characterisation of the maternal diet using food frequency questionnaires to determine macronutrient intake (i.e. protein, fat, energy and carbohydrates) and other dietary intake e.g. dairy products and green-leafy vegetables. Yajnik and colleagues performed an analysis to identify whether there was an association between any of these variables and the phenotype of insulin resistance, but did not find one. This provides reassurance that macronutrient status does not confound the association with maternal one-carbon status. The study, however, does not detail measurement of other micronutrients that could act as confounders in this association.

b. Timing of the exposure

In PMNS, biochemical and questionnaire data was collected at 18 and 28 weeks gestation to characterise maternal nutritional status in a population for whom pre-pregnancy weight, body mass index and physical activity had previously been characterised (182). Data collected from women at 18 weeks gestation is highly likely to represent pre-pregnancy status; the biochemical measure of red cell folate is unlikely to change during pregnancy, and although serum vitamin B12 levels fall in normal pregnancy, there is not a significant change until after 20 weeks gestation. Other researchers have highlighted the importance of the variable timing of exposure on programming (109). In the sheep model of methyl group deficiency previously described, the exposure was peri-conceptual, a time point that it can be assumed the PMNS mothers were exposed also. One potential difficulty with interpreting the findings of PMNS lies in the exposure lasting throughout pregnancy and possibly also during lactation and postnatally and this concern has only partly been addressed by measuring vitamin B12 levels in 6 year offspring and finding that there was no association of this with phenotypic outcome.

c. Genetic heterogeneity

The genetic differences between participants in PMNS could have an impact on the parental and offspring phenotypes, as well as potential epigenetic differences that are detected. Although the participants come from a small and well-defined rural area with minimal population changes, there will still be considerable heterogeneity between individuals. Genetic differences could control the risk of developing insulin resistance and obesity, but also the nutritional status of the mother. For example, vitamin B12 levels can be affected by the common variant FUT2 and folate status by the common allelic variation at MTHFR (183). In these pilot studies proposed, it was not possible to separate the pure genetic effects on offspring phenotype, nor study the possible genetic-epigenetic interactions at known gene variants, but future studies should be designed to do so. Finally, most of the knowledge about genetic susceptibility to these phenotypes are from studies of populations of European origin and therefore it is important to understand that the genetic variation in participants of PMNS may not be fully characterised or represented on genome browsers and variant databases.

e. Phenotype

PMNS showed some association between birth phenotype (e.g. weight, skinfold thickness and abdominal circumference) and maternal folate concentration. However, this association did not extend to variation of maternal vitamin B12 status (as measured by serum B12 or methylmalonic acid). These findings highlight how PMNS adds considerable new understanding of the phenotypic consequences of early developmental programming and that long-term follow up is crucial to this type of study. The DEXA studies and oral glucose tolerance tests provided evidence of increased fat mass and insulin resistance in the programmed offspring, but these measures do not fall into the diagnostic criteria of either overweight, obesity or type 2 diabetes. They can be considered to be precursors of these conditions, but this subclinical status does not conclusively determine a future phenotype.

4.8.2 Discussion points related to the epigenomic studies

The DNA methylation profiles of 16 offspring born to PMNS were characterised using Medip-seq. This experimental platform generated over 48 million sequencing reads suitable for bioinformatic analysis from DNA collected from 8 cases and 8 controls. After careful quality control processes, normalisation and custom bioinformatic analysis using combined

algorithms, a list of 48 differentially methylated regions was generated showing variable methylation between cases and controls. A significant number of reads was lost through the various QC filters after sequencing, but this was consistent across samples and similar to other Medip-seq datasets generated in the Rakyan lab. The DMRs were located at different genomic regions, and some were located at repeat elements and regions of possible transcriptional regulation. Technical validation using the Illumina HumanMethylation 450k array was attempted using the same samples but array coverage did not overlap many regions of differential methylation identified by Medip-seq. Where there was overlap between 450k probes and Medip-seq DMRs, the 450k beta values supported most of the data generated by Medip-seq in terms of directionality of methylation difference. However, the majority of these validated DMRs were at sites where a genetic polymorphism was regulating the epigenetic state, in most cases by the creation or abrogation of a SNP-CpG. SNP-CpGs display 0, 50 or 100% methylation and are therefore likely to be called as DMRs or MVPs in a study such as this with a small sample size. The allelic variation at these SNP-CpGs is likely to be a random genetic event rather than anything associated with the environmental exposure of phenotypic outcome. Evidence from Chapter 2 suggests that a far larger sample size is required to detect association between risk alleles and methylation status even within a study targeted to known genetic variations. A far larger sample size (probably in the hundreds or thousands) would be required to detect the contribution from unknown SNPs and this is far beyond the remit and economics of this study.

The preponderance of SNP-associated MVPs from the 450k dataset suggests that a combined genetic-epigenetic approach should be taken in future studies to identify the epigenomic basis of programming in this model. Pure epialleles, i.e. stable epigenetic variants independent of genetic state, may be a feature of developmental programming in humans, but the data from this study suggests that a far larger study would also be required to identify these as preliminary evidence suggests that they may be associated with minimal variation in methylation of the order of 2-5% (89). Data from the 450k array shows that the data generated from the 10 vs. 10 MVP call was significantly underpowered to detect such methylation differences between the cases and controls, despite good performance of the array itself with minimal data lost for technical reasons such as cross-hybridisation or poor quality probe binding. The MVPs that were called were statistically too weak for control of multiple hypothesis testing and therefore must include a high false discovery rate. Future studies must incorporate power calculations based on the knowledge generated from this dataset, including an understanding of the background 'noise' of data generated from this array. However, such power calculations are difficult as the amount of DNA methylation

variation that is functionally relevant is not known and is likely to vary according to a complex range of other genetic and epigenetic phenomena. An alternative approach to identifying the epigenetic differences associated with variations in one-carbon metabolism is to understand the consequences of direct exposure, without the transgenerational/fetal programming angle. Additional studies using an adipose cell line exposed to varying B12 and folate environments are currently underway with collaborators and will be discussed in Chapter 8. Such model systems offer the opportunity to study more detailed functional effects, including integrated genomic, epigenomic and transcriptomic platforms. A similar model in humans, using young adults in Pune before and after B12 replacement is being performed by a collaborating group in Southampton (Caroline Fall, Karen Lillycrop and Mark Hanson) and personal communication of their preliminary data suggests that studying individuals pre- and post-intervention is an effective way to increase power to detect differences and uncover relevant biology.

The Medip-seq dataset generated in these experiments has many disadvantages compared to array data. The potential experimental bias from variable fragment sizes and Medip enrichment, are likely to be important factors in obscuring true biological variation in DNA methylation. Careful optimisation of experiments was performed to try and overcome these types of biases, and it is unlikely that they could be improved further. Increased sequencing depth is likely to overcome some of these issues, however the financial cost and bioinformatic workload consequent from this could be prohibitive. Given the findings that many DMRs and MVPs identified in these studies are located at SNP-CpGs or may be associated with *cis* or *trans* effects, the lack of single base resolution from Medip-seq is a significant disadvantage. Quantitative methylation estimates across the Medip-seq bioinformatic windows do not provide significant resolution to identify whether it is a genetic feature driving a methylation differences and studies using targeted BS-pyrosequencing or Sequenom are required to elucidate the underlying architecture of these DMRs.

At the present time, a shortlist of the Medip-seq DMRs from the combined USeq and Thomas Down calling strategy, and super-selected by functional interest and statistical significance, are undergoing technical validation. This validation is being performed using Sequenom by collaborators (Giriraj Chandak et al, CCMB Hyderabad, India) who have additional DNA for validation experiments. In the first instance, the methylation state of these DMRs in the same samples used will be assessed as technical validation of the platform; this will help direct the methodology for future epigenomic studies. Should these DMRs technically validate, replication in a larger sample may be performed and, given the control for false discovery rates in the USeq analysis, it may be that these are true DMRs.

One of the objectives of this study was that it should provide an 'unbiased' view of epigenetic variation, with no *a priori* hypothesis as to where this variation would be occurring (other than that determined by the presence of CpG sites capable of methylation). As discussed, the approach and study design did unintentionally incorporate bias towards genetically driven methylation differences as they create large differences in methylation that are reproducible and widespread across the genome. Since the design and implementation of these experiments, more researchers are using BS-seq as a means of studying DNA methylation across the genome and this now offers the gold standard approach. BS-seq is costly and analytically challenging but provides an epigenomic view of DNA methylation unbiased by genetics and can give a single-base resolution suitable for combined detection of genetic variants such as SNPs.

The functional relevance of DMRs and MVPs identified in these experiments has not been determined. When viewed on standard genome browsers, many of the Medip-seq DMRs overlap regions of possible regulatory function. These sites are represented by data tracks from resources such as ENCODE, and Pol2 binding, nucleosome positioning, the location of DNase hypersensitivity sites and CTCF binding. These findings are of potential interest as it is important to consider the integrative role of DNA methylation with higher-order epigenetic modifications, e.g. histone modifications and chromatin packaging. The identification of co-methylation within some DMRs, often associated with SNP-CpGs or putative cis-acting effects, is important in the possible biological plausibility of these methylation changes. Other important considerations in the interpretation of epigenomic data are how local (*cis*) or long-range effects (*trans*) may exert control of DNA methylation. These important factors will be discussed in more detail in Chapter 8.

Another objective of this study was to demonstrate the feasibility of using whole blood to detect differences in DNA methylation associated with a model of *in utero* programming. Whole blood sampling is a relatively practical, acceptable and cheap means of obtaining biological material for genetic studies in large clinical cohorts and has been the commonest tissue used in genomic studies such as GWAS. The choice of tissue in epigenetic and epigenomic studies is a more complex one and will be discussed in detail in Chapter 8. In the PMNS model presented here, it was felt that whole blood was a suitable tissue to use to generate epigenomic insights as the adverse environment that is associated with the programmed offspring phenotype was present in the periconceptual period. Epigenetic modifications can be induced through early developmental environmental exposure (in

gametogenesis and early embryogenesis) and can affect all 3 developing germ layers. For this reason, it is thought that whole blood may carry the molecular 'memory' of this programming and furthermore, its possible role in the inflammatory and thrombotic pathways suggest it may hold tissue-specific epigenetic changes that relate to the obese and insulin resistant phenotype observed. However, whole blood does contain mixed cell types and therefore differences in cell counts may result in false positive results.

These studies highlighted the importance of sample collection, storage and processing and that the resulting purity of samples is paramount to the data generated from these studies. The sample contamination identified in the first set of biological material processed in this study could have been misinterpreted as an experimental difference had the samples not been processed and analysed individually. Larger studies in the future are going to need to address the sensitivity of epigenomic studies to sample contamination carefully, especially in studies where sample pooling strategies are used.

Finally, it must be considered whether this epigenomic study provides any molecular support of the epidemiological and clinical evidence of fetal programming from alterations in maternal B12/folate status. Technical validation of the Medip-seq DMRs is going to be important to decide whether any of these may have a biological role in this fetal programming. The location of some of these DMRs at regulatory elements, should they validate and replicate, may suggest a functional role in transcriptional regulation that could be far-reaching and affect a number of biological pathways involved in insulin resistance and the development of obesity. It is interesting to note that both Medip-seq and 450k array analysis yielded a greater number of hypomethylated DMRs/MVPs that hypermethylated ones. This finding might be expected given that a low B12 environment would be expected to reduce the production of methyl groups for DNA methylation, therefore resulting in a propensity to hypomethylation. Further studies would also need to determine whether the epigenetic differences identified are primary phenomena or secondary to the phenotype that had already developed in the offspring studied in the case group. This is particularly important as the samples selected in this study were enriched for the programmed and control phenotypes in order to maximise the potential outcome.

Possible confounders in the design of this study, and the wider PMNS study, come from the role of postnatal environmental differences. Although serum vitamin B12 levels were the

same in 6-year old offspring born to case and control mothers, the immediate postnatal conditions of each group of offspring were not characterised. It is conceivable that the neonates born to B12 deficient mothers would remain deficient themselves during lactation and given that the familial environment and dietary behaviours are often similar.

Since performing this study, data from a rat model has recently provided further evidence of the role of one-carbon metabolite deficiencies in the pathogenesis of obesity and diabetes (184). This model exposed female Wistar rats to either folate, B12 or combined folate and B12 deficiency and compared these, in the pre-pregnant state, to controls. After 3 months of dietary restriction and pectin administration (to reduce intestinal absorption of these nutrients), nutritional deficiencies were confirmed biochemically and were said to be equivalent to human dietary deficiency. Some of these female rats were examined phenotypically after 3 months and measures of body weight, body composition (using total body electrical conductivity) and fat mass (dissected fat pad weights) were calculated. There were differences ($p<0.05$) in body weight in the folate deficient and B12 deficient rats, as well as altered body composition (increased fat) and increased fat mass in B12 deficient and folate deficient rats, respectively. Exposure to these dietary deficiencies were continued in 48 rats through pregnancy, weaning and into their male F1 generation and similar phenotypic assessment was performed in offspring until 12 months of age. By 12 months of age, the male F1 generation exposed to folate, B12 and combined deficiencies showed significantly increased ($p<0.01$) body weight, body fat and fat mass compared to controls. In addition, all 3 groups showed higher plasma cholesterol, triglycerides and some features of a pro-inflammatory state, including elevated TNFa, IL-6, IL-1b, and reduced adiponectin in most of the exposed groups. These findings provide convincing evidence of the direct effects of variable one-carbon deficiencies on fat accumulation, and it could be inferred that there is some direct link via adipogenesis. The authors develop this hypothesis by measuring fatty acid synthase and acetyl-CoA-carboxylase activity in the livers of the 12 month-old male offspring, finding increased activity of both of these adipogenic proteins in the livers of deficient animals. To add to this hypothesis, other studies have shown a link between vitamin B12 deficiency and accumulation of liver fat (185). Unfortunately, this model does not provide evidence of the potential 'programming' effect of maternal deficiency as the offspring have been directly exposed to the same dietary conditions as their mothers in their post-natal life. It also does not look at the state of combined maternal B12 deficiency and folate repletion characterised in the PMNS data, described earlier, that conferred the highest risk of adiposity and insulin resistance through programming.

The evidence from these animal models, as well as the difficulties encountered with studying the epigenome of the offspring born to PMNS in this study, highlights the need for careful study design in the future. Future studies should incorporate model systems in order that tissue-specific mechanisms can be identified in molecular studies, and this could include animal models or cell line experiments. Animal models also have the advantage of being able to study long-term phenotypic outcomes in a manageable time window and this is also important in order to correlate epigenetic and expression differences with direct environmental exposure and outcome. The use of inbred animal strains or cell lines also enables the identification of pure epigenetic phenomena, isolated from the complex genetic-epigenetic interactions that are identified in human studies. In human studies, intervention studies may provide the most useful means of differentiating primary and secondary epigenetic modifications, and this will be discussed in more detail in Chapter 8.

Chapter 5: Matlab Famine Study

5.1 Introduction

5.1.1 Background to Matlab Study

As discussed in chapter 1, in utero famine exposure has been characterised as a model of the fetal programming of adult diseases such as type 2 diabetes, hypertension and major affective disorders in the Dutch Winter Hunger and Chinese Great Leap Forward studies. These studies complement the original observation from Barker et al that associated low birth weight with an increased risk of type 2 diabetes in later life. In recent years, researchers have found preliminary evidence for possible epigenetic mechanisms linking the aberrant maternal environmental to the developing fetal genome, showing differences in DNA methylation of candidate genes in adults exposed to famine. However, as yet, this data has not been replicated beyond the Dutch Winter Hunger Study and is limited to the study of specific candidate genes that are assumed to have a role in the programmed disease. Interestingly, although the Dutch Winter Hunger studies have shown that famine-exposed offspring display insulin secretion defects (48), the candidate genes studied in recent epigenetic studies are focused on insulin resistance pathways. The merits of taking an unbiased epigenomic approach to study complex gene-environment interactions have previously been considered and this discovery-based approach will be exploited in this study to generate new insights into the mechanisms of fetal programming.

In this study, an epigenomic study of DNA methylation from offspring born to another famine study will be performed. This famine study is based in Matlab, a rural area 100km from Dhaka in the Chandpur district of the Chittagong division of Bangladesh. The study is run by the International Centre for Diarrhoea Disease Research, Bangladesh (ICDDR,B) and the Principal Investigator of the study is Dewan Alam, Head of the Chronic Non-Communicable Disease Unit.

Dewan Alam and others have described the background population in Matlab in their Health and Demographic Surveillance System (HDSS) of 517 men and women aged 27-50 years. This survey was not designed to study the famine-exposed, but by its cross-sectional nature across this age group will incorporate many individuals who were exposed to famine at some stage in their early life. The results of the survey show that Matlab has many individuals who are underweight (40% of men and 27% of women) but that in addition to this, overweight and obesity are also prevalent, with 7% of men and 14% of women having a BMI>25 kg/m^2. Of the participants studied, 3% had type 2 diabetes and a further 9% had impaired glucose tolerance. Overweight and obesity were usually accompanied by abdominal obesity and this was strongly associated with glucose abnormalities.

The Matlab Famine Study comprises adult cohorts in this same region that were exposed to famine in early life, during the Bangladesh famine that occurred between 1974-1975. The famine occurred 2 years after the Bangladeshi war of Independence and arose due to crop failure, lower purchasing power and grain imports (Alamgir, 1980). The famine was responsible for several million deaths across Bangladesh due to starvation and associated infectious diseases such as cholera. The specific nutritional deficits associated with famine exposure have not been characterised, but the severity of the famine would suggest that it included deficiencies of a variety of macronutrients and micronutrients.

Two cohorts have been created in Matlab, a birth cohort of individuals exposed to famine before, during or soon after birth, and a second cohort of individuals exposed peri-pubertally. These cohorts were formed in 2000-2002 by researchers at ICDDR,B using community-wide recruitment of individuals that had previously been exposed to famine, as per the defined groups above. At the time of sampling, the birth cohort aged 27-32 years and the peri-pubertal exposure cohort aged 38-50 years. The study followed standard consent procedures set out by ICDDR,B and the institute has a long history of successful research and engagement with the local community in Matlab. Individuals within the cohorts were randomly selected individuals to be in a pilot study of the long-term health outcomes, focused on cardiometabolic diseases, of famine exposure. Measurements of blood pressure, body mass index, and waist circumference were performed, as well as 75g oral glucose tolerance tests (WHO standard) and fasting lipid profiles on over 500 individuals. The results of this study (unpublished) are summarised in table 5a.

	Birth cohort		Peri-pubertal cohort	
Number	219		298	
Mean age (years)	29.5±1.3		43.1±4.2*	
Smokers (%)	21.9		25.5	
Owner of cultivatable land (%)	39.7		43.0	
Family size (mean number of children)	5.5±2.5		5.9±2.6	
Education (mean number of years)	5.8±5.1		3.9±4.8*	
	Male	**Female**	**Male**	**Female**
Mean BMI (kg/m2)	19.6±2.5	20.3±3.0	19.9±3.2	21.3±3.7
BMI <18.5 (%)	40.9	33.3	39.7	22.2
BMI ≥25 (%)	4.3	8.7	8.1	17.3
Abdominal obesity (% >90cm in males, >80cm in females)	2.2	14.1	8.8	24.1
Systolic hypertension (% ≥140mmHg)	1.1	0.8	4.4	8.0
Diastolic hypertension (% ≥90mmHg)	0	2.4	5.9	8.0
Mean fasting glucose (mmol/l)	4.8±0.5	4.7±0.5	5.1±1.1	5.2±1.7
Impaired fasting glucose (%)	8.6	2.4	12.5	17.3
Mean 120 mins glucose (mmol/l)	4.8±1.4	6.0±1.7	5.6±2.4	6.6±3.5
Impaired glucose tolerance (%)	3.2	8.7	8.8	11.7
Type 2 Diabetes (%)	0	5.1	3.2	8.1
Metabolic syndrome (%)	2.2	4.0	5.9	14.2*~

Table 5a. Summary of clinical data from participants in the Matlab famine study. * = difference at p <0.05, ~ differences significant between females in each cohort and males vs. females. Metabolic syndrome was diagnosed with NCEP ATP III criteria. ADA criteria used: Impaired fasting glucose ≥5.6 to <7.0 mmol/l, Impaired glucose tolerance 120mins glucose ≥7.8 to <11.1 mmol/l.

These data show that across the entire Matlab cohort, exposed to famine around birth or peri-pubertally, there is a high prevalence of hypertension, impaired fasting glucose, impaired glucose tolerance, type 2 diabetes and metabolic syndrome in adults, in keeping with that described in the HDSS. Results from the famine cohort cannot be compared to the HDSS directly as they are not age-matched, and also because participants in the HDSS may also have been exposed to famine. However, it can be concluded that this rural population does have a high risk of cardiometabolic diseases despite still showing the striking features of a region that continues to be severely impoverished, e.g. low literacy rates, large family sizes and greater than one-third of individuals being underweight. The observed diabetes prevalence in these cohorts falls somewhere between the estimated prevalence of 8.1% and 2.3% in urban- and rural-dwelling Bangladeshis (respectively) (186). The genetic predisposition of the Bangladeshi population towards diabetes is not known but genetic diversity in Bangladesh is thought to be

similar to related ethnic subgroups in India and these are considered, but not yet proved, to have a higher genetic tendency towards type 2 diabetes.

In addition to the data presented from the birth and peri-pubertal cohorts, specific data was collected according to timing of famine exposure in the birth cohort. The participants formed three groups;

Group A: Born before the onset of famine, and therefore exposed to famine in early childhood.

Group B: Born during famine, and therefore exposed to famine in utero. .

Group C: Born after famine, and therefore not directly exposed to famine.

Phenotypic data for each of the three groups was then collated and analysed, by researchers at ICDDR,B (see table 5b). Unfortunately, there is no data on trimester-specific exposure within the 'born during famine' group. It is also important to note that although the 'born after famine' group was not directly exposed to famine, it is likely that they were indirectly exposed via exposure of the parental germline.

	Group A	**Group B**	**Group C**
	Born before famine Exposure in early childhood	**Born during famine** Exposure in utero	**Born after famine** Unexposed
Number	81	68	70
Mean age (years)	30.8±0.35	29.0±0.28	27.9±0.31
Male (%)	42	41	44
Mean weight (kg)	49.7±7.9	47.4±8.6	49.4±6.9
Mean height (cm)	156.4±7.9	155.6±8.1	156.9±8.6
Mean BMI (kg/m^2)	20.3±3.0	19.5±2.8	20.1±2.5
BMI <18.5 (%)	32	48.5	30.0
BMI 18.5 - 24.9	59.3	45.6	64.3
BMI ≥25	8.6	5.9	6.8
Fasting plasma glucose (mmol/l)	4.79±0.53	4.76±0.48	4.67±0.57
Impaired fasting glucose (%)	6.2	5.9	2.9
Mean 120 mins glucose (mmol/l)	5.21±1.38	5.76±1.60	5.58±2.02
Impaired glucose tolerance (%)	3.7	11.8	4.3

Table 5b. Clinical data from Matlab famine study birth cohort, showing phenotypic outcomes in those exposed and unexposed to famine.

Evidence from the Dutch Winter Hunger and Chinese Great Leap Forward studies would suggest that the in utero exposure group would be expected to show that famine exposure in utero is associated with a higher prevalence of diabetes, and specifically, that this may be driven by a problem with insulin secretion, manifest by a raised fasting glucose. In these data, an increased tendency towards impaired glucose tolerance is observed, with an odds ratio of 3.39 in the in utero famine exposure group (group B) compared to groups A and C. Raw data from Bangladeshi has not been accessible or analysed further to identify whether any of the other observed phenotypic differences between exposure groups are statistically significant. It can be seen that there is a higher rate of impaired fasting glucose in both group A and group B, compared to group C, but a statistical test has not been performed to evaluate the strength of these differences any further. The odds ratio of the in utero exposed group having impaired glucose tolerance compared to the other two groups was 3.39 (95% confidence interval 1.02-11.28).

5.1.2 Specific aims of this study

Hypothesis 1: Direct famine exposure in early developmental life may be associated with epigenetic variation. The timing of this direct exposure, in utero or early postnatal life, may influence whether programming occurs.

On the basis of these data presented above, the primary hypothesis of this study is that group B (in utero famine exposure) forms our exposure group, but that group A is a second exposure group and that both may show the epigenetic marks of direct famine exposure in early developmental life. Further, it could be assumed that the absence of direct exposure to famine in Group C makes this the control group with which to compare groups A and B. However, as there is no published data with which to support this hypothesis, the epigenetic studies have been designed to compare all 3 groups as distinct entities and then look for possible commonality across groups. The phenotypic data and previously published work from the Dutch and Chinese studies will be used in support of group B as being the 'programmed' group, but the incomplete data analysis of the former and limited applicability to this study of the latter will not be used to make assumptions.

Hypothesis 2: Three-group analysis of those exposed to famine before, during and after in utero life may identify the timing at which programming occurs, but may also reflect the tissue-specificity of epigenetic signatures.

A secondary hypothesis that will be considered is that exposure to famine during different time windows may result in tissue-specific epigenetic consequences that are not going to be detectable in this study, which uses a single tissue type. If differential methylation is seen only in association with famine exposure in utero, this may be a consequence of the environmental insult having occurred before the 3 embryonic germ layers have started on their differential developmental course and therefore having a detectable effect on all tissues. In contrast, it could be expected that those exposed in later embryogenesis or postnatal life, may not develop pervasive and stable epigenetic marks in all tissues and if they do exist, they may be restricted to tissues with a direct functional role in aetiology such as fat, muscle or liver.

Hypothesis 3: Identification of epigenetic marks associated with famine exposure may characterise secondary phenotypic consequences, rather than the primary programmed event itself.

This same issue has been discussed in Chapter 3, in relation to the Pune study. The 3 group study design in this Matlab cohort may allow some additional insight into primary versus secondary epigenetic events as the strongest phenotypic outcome appears to be in group B but exposure to famine has occurred in group A and B. Therefore, should the same epigenetic signature appear in both group A and B, it could be interpreted that this is phenotype-independent.

5.2 Methods

5.2.1 Matlab Famine Study

DNA samples collected from individuals in the Famine Study were used to investigate the possible long-term effects of programming on diabetes risk. Only the birth cohort, and not the peri-pubertal cohort described above, was used to test the hypotheses set out above.

Three groups of offspring born to the famine study were selected, comprising the following:

Group A: Offspring born before famine exposure, i.e. exposed to famine in early childhood

Group B: Offspring born during famine exposure, i.e. exposure in utero (in whom the highest risk of impaired glucose tolerance was observed)

Group C: Offspring born after famine exposure, i.e. unexposed, but maternal exposure to famine before pregnancy.

5.2.2 Sample collection and selection

Whole blood genomic DNA samples from 120 (40 per group) male and female participants born to the Matlab Famine Study were used (see table 5c) and comprised a random selection of the complete collection of study samples with no selection towards phenotypic outcome or any other variable. These samples had been collected previously, when the participants of this study were aged 27-32 years old, for the phenotypic study described above as well as for future use in genomic experiments. Samples had been stored at the ICDDR,B campus in Matlab in a $^{-}80^{o}$C freezer since collection in 2004. Consent had been taken at the time of collection for genetic studies by the local research team, but re-consenting took place by the same research team in order to use these same samples for epigenetic research, under consent procedures set out and approved by the ICDDR,B ethics board.

Exposure group	Dates of exposure	Age at sampling	Sample number
Group A: Born before famine	April – June 1974	27	40
Group B: Born during famine	Dec '74 – Nov '75	30	40
Group C: Born after famine	April – June 1976	32	40

Table 5c. Details of used in Matlab famine study epigenomic experiments.

The initial design for this experiment included 30 mixed sex samples per experimental group (i.e. total n = 90) that were randomly selected from the entire sample set (n = 120). This sample size was determined by cost constraints at the time. Following this first experiment, extra funds became available to put the remaining samples through the same experiments and therefore a second batch of samples was run as validation.

5.2.3 DNA extraction and preparation

Whole blood samples were transported at 4^{o}C to ICDDR,B, Dhaka. Genomic DNA was extracted by Dr Shamim Iqbal, postdoctoral researcher at ICDDR,B, under the supervision of Sarah Finer, using a Qiagen QiaAmp Maxi kit using the standard protocol. The extracted DNA was split into two aliquots; one was stored at ICDDR,B and the other was transported at room temperature by Shamim Iqbal to the Blizard Institute. On arrival at the Blizard Institute, DNA quality was checked on agarose gels, and quantified using a Qubit system.

5.2.4 Illumina HumanMethylation Array

Genomic DNA samples underwent bisulphite conversion using a starting quantity of 500ng DNA. Bisulphite conversion and qPCR test of conversion efficiency were performed as described in section 2.4.1.1 and 2.4.1.1.

Samples were randomly assigned to 96-well plates with samples from another experiment that was running concurrently. Batch 1 included 90 samples (30 per experimental group) as well as 2 water controls. A second batch included 55 samples, incorporating the 30 samples that had not yet been run on the 450k array as well as repeat samples (technical repeats and inter-batch duplicates) from batch 1 and water controls. No intra-batch sample duplicates were run.

The samples were put through the standard protocol for 450k array hybridisation and processing, and this was performed by UCL Genomics.

5.2.5 Data QC and analysis

5.2.5.1 All samples

Data QC, processing and analysis was performed according to the protocol set out in section 2.5.4. As the experimental samples were from a mixed sex population, X and Y probes were filtered out as well as the standard removal of cross-hybridising probes. The sex of the offspring determined by the data sent to us from ICDDR,B was cross-validated with visualisation of the Y chromosome probes in all samples (a specific pattern of methylation is seen across these probes according to the sex of the DNA source).

Batch 1 samples were analysed according to standard protocols, using MDS plots and then MVP calling. Batch 2 samples went through the same protocols including an MVP call, but the data produced from these samples was used for validation of batch 1 and therefore the data was also analysed as such. The methods used in this validation strategy will be discussed as the results are presented but include identification of overlapping MVPs and concordant directionality of methylation difference between each batch. Batch 2 also contained inter-batch duplicate samples with which an estimate of batch difference in array performance could be made and incorporated into the understanding of the data output.

5.3 Results

5.3.1 DNA quality

Some DNA samples showed degradation of the high molecular weight DNA when visualised on agarose gels (Figure 5a), but this did not appear to affect their performance at any of the quality control checks during the subsequent steps of the experimental protocol. There did not appear to be significant RNA contamination. Qubit concentrations were used to identify the concentration of pure DNA.

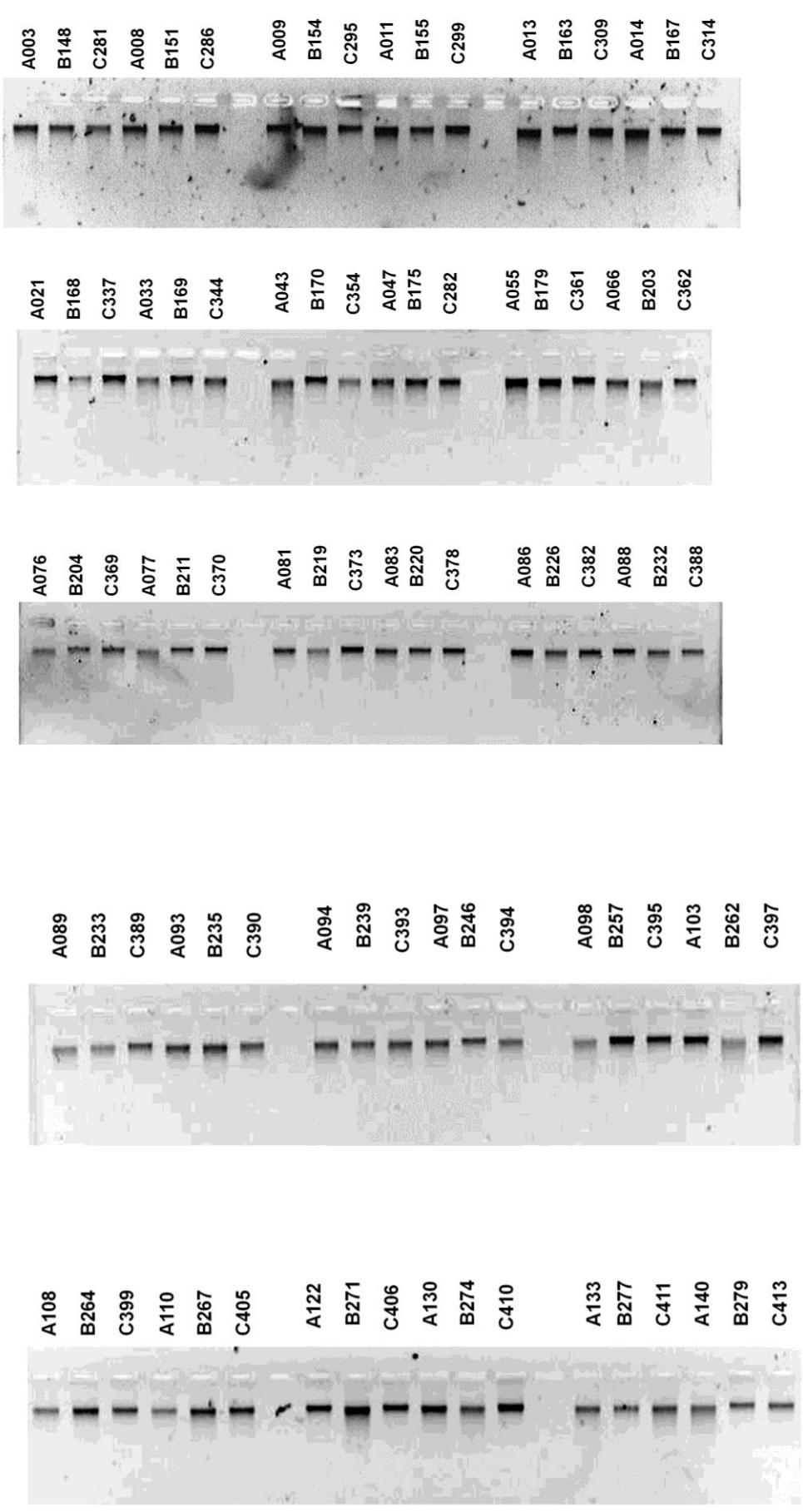


Figure 5a. Agarose gels of DNA samples extracted from Matlab study participants

5.3.2 Illumina 450k methylation array data

5.3.2.1 All samples

A check of samples was performed on the Y chromosome probes to confirm that the sex of the samples was as expected from the participant database. DNA from males shows a predicted bimodal distribution of beta values with a large peak at low CpG densities and second (smaller) peak at high CpG densities (see figure 5b). Female DNA shows an erratic distribution of beta values across the Y probes with lower beta values due to the minimal probe binding in the absence of an X chromosome. The presence of any methylation signal from female DNA at Y probes is thought to reflect cross-hybridisation from sequence homology between the X and Y chromosomes. These cross-hybridising probes are subsequently removed from analysis by the QC filters. This QC process confirmed the expected sample gender in all, and this was considered to be a reassurance that samples had not been swapped or mislabelled.

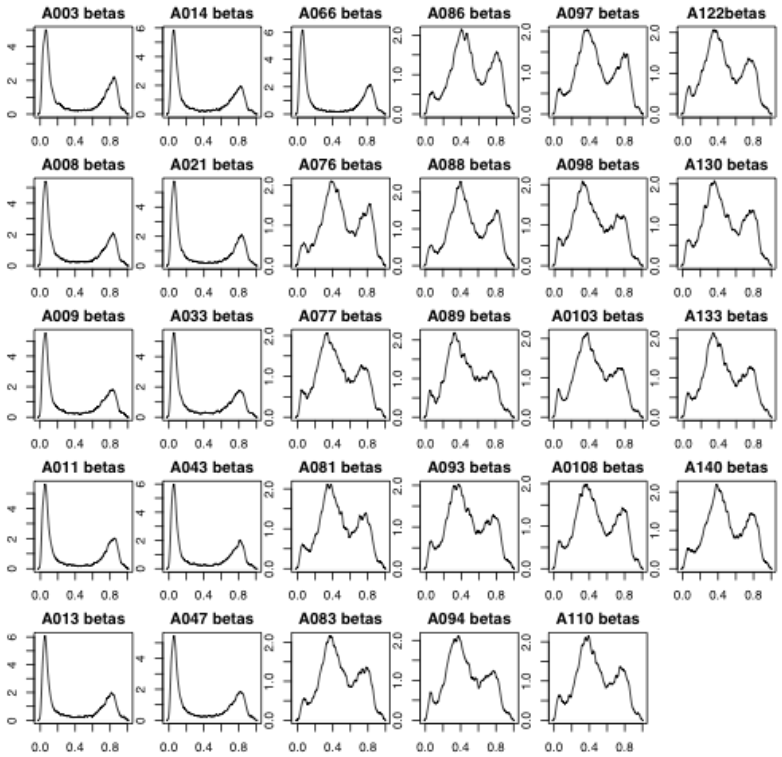


Figure 5b. Beta values from Y chromosome probes in a selection of samples from batch 1. Y axis = beta value; X axis = CpG density. This data shows clear and expected distinction between male DNA (e.g. A003) and female DNA (A097).

5.3.2.2 Batch 1 samples

Standard QC checks of each step of the experimental protocol (from bisulphite conversion to all aspects of probe hybridisation and staining) confirmed that this had been performed successfully (data not shown). Subsequently, the data was normalised and standard data filters were applied to remove X and Y probes as well as those that cross-hybridised.

An MDS plot was generated for the data from batch 1 (figure 5c) and this identified 12 samples that had been hybridised to one complete array clustered away from the other samples. The defined clustering of the randomised group of samples hybridised to this array suggested that this was a technical problem and one that was associated with an array or hybridisation problem rather than an issue with anything upstream, such as the bisulphite conversion. The detailed QC plots for each step of the experimental pathway do not highlight where in the array processing this problem lay, but it was decided that the data indicated a chip failure and it was therefore excluded from subsequent analyses. As a result, all 12 samples that had been hybridised to this array were removed from further analysis, leaving 78 samples (29 from group A, 25 from group B and 24 from group C) to be analysed in this experiment. These samples were repeated in the second batch of arrays.

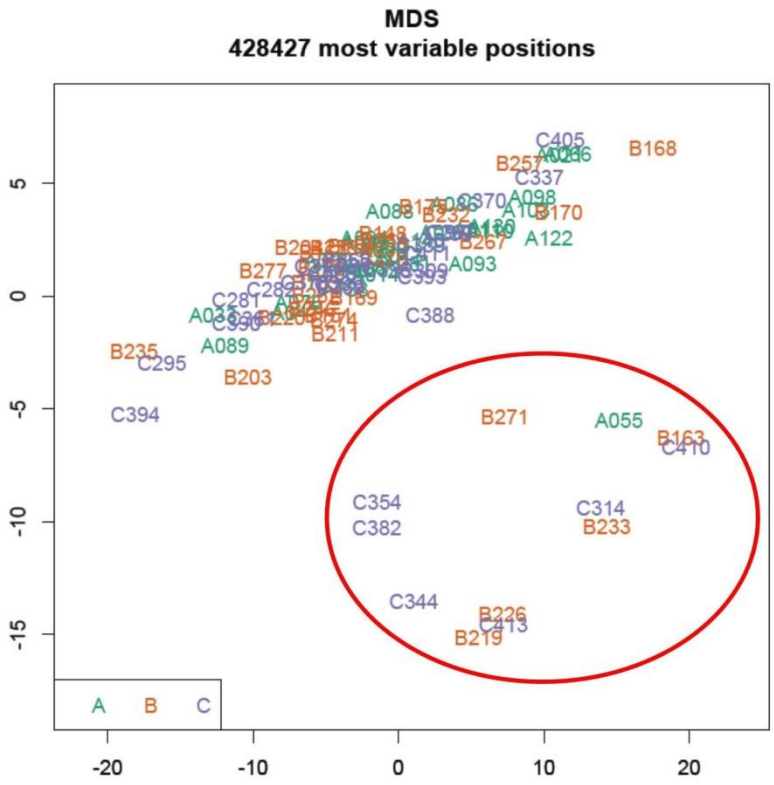


Figure 5c. MDS plot of 90 Matlab samples from batch 1 (after normalisation). This shows plot identifies 12 outlying samples (circled in red) that were run on a single chip.

A summary of the 78 samples selected for batch 1 analysis, after exclusion of the chip failure samples, is presented below (table 5d). The data shows an imbalance of males and females across the sample groups due to the random sampling and exclusion of some samples for technical reasons. The results of a Chi-squared test suggest that there is no overall difference at a statistically significant level across groups however, on observation of the differences, it does appear to be important, especially in Group C.

Exposure group	**Number of samples**	**Males**	**Females**
Group A:	29	11	18
Group B:	25	13	12
Group C:	24	7	17
Total	**78**	**31**	**37**

Table 5d. Sample numbers used for analysis of batch 1. Chi-Squared 2.7, df 2, p = 0.255.

5.3.2.3 Batch 2 samples

There were no concerns with the performance of the second batch of samples as they passed through the experimental steps and QCs. Data from all 55 samples was normalised (independently of batch 1) and was used for downstream analysis. The sample numbers are presented in table 5e.

Exposure group	**Number of samples**	**Males**	**Females**
Group A:	14	9	5
Group B:	19	7	12
Group C:	22	8	14
Total	**55**	**24**	**31**

Table 5e. Sample numbers used for analysis of batch 2. Chi-Squared 3.3, df 2, p = 0.20

5.3.2.4 Batch differences

Samples from batch 1 and batch 2 compared using pre- and post-normalised data to examine overall data quality and make comparisons between them. An MDS plot of pre-normalised data (after X, Y and cross-hybridising probes and chip failures were removed) showed a clear distinction between batch 1 and 2, however, each batch clustered well (figure 5d).

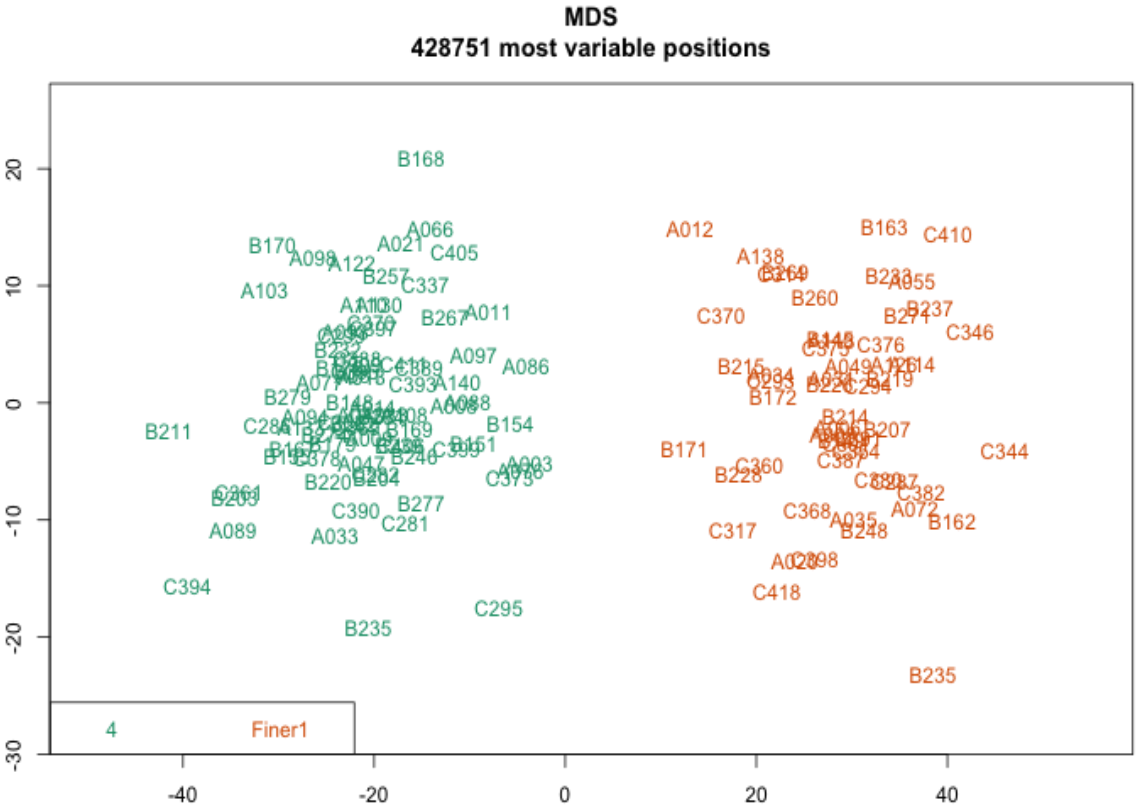


Figure 5d. MDS plot of batch 1 samples (green) and batch 2 samples (orange). The data is presented after X, Y and cross-hybridising probe filters have been applied, chip failures have been removed, but pre-normalisation of each dataset has not yet taken place.

An MDS plot of the two datasets post-normalisation again shows distinct clustering between batch 1 and 2 but the differences are small and the clusters appear tighter (figure 5e).

Figure 5e. MDS plot of batch 1 samples (green) and batch 2 samples (orange) using normalised data (and filters applied as per figure 5d)

5.3.2.5 Controlling for batch effects

With any array experiment, batch effects can introduce variation between batches due to the inevitable technical variability between experiments, despite standardised experimental protocols and the use of robotics. The careful quality control checks that were performed in all experiments ensures that the batch differences are not due to a specific technical problem that could introduce significant bias into the datasets. The QC plots from both batch 1 and 2, were analysed carefully and problem samples from batch 1 were removed and re-run with batch 2, along with inter-batch duplicates. Duplicate samples allowed careful examination of the data QCs to ensure that the variation between them was within what would be expected from batch differences.

Batch effects may be controlled through normalisation if the differences that they produce are linear rather than random. Controlling batch effects that are non-linear, or random with

normalisation introduces an artificial assumption that the data varies in the same way. If such data is normalised, a bias will be introduced that may mask true differences and/or reduce the control of false positives.

Batch effects were seen between batch 1 and 2, and these are easily visible in the MDS plots (figures 5d and 5e). The nature of the batch effects, including whether the differences were linear or random, was examined further with an additional plot (figure 5f). In this plot, one pair of sample duplicates from both batches are taken and the probes showing <1% methylation difference between the duplicates selected (X = 1st batch; Y = 2nd batch). The beta values at these probes are then identified in 3 other sample duplicates (marked in blue, green and red) as a way of visualising the noise between the samples across the two batches. The noise is not linear, making it difficult to correct. There is some linearity, but only a minimal amount and this is unlikely to have much effect on the ability to validate MVPs across the arrays.

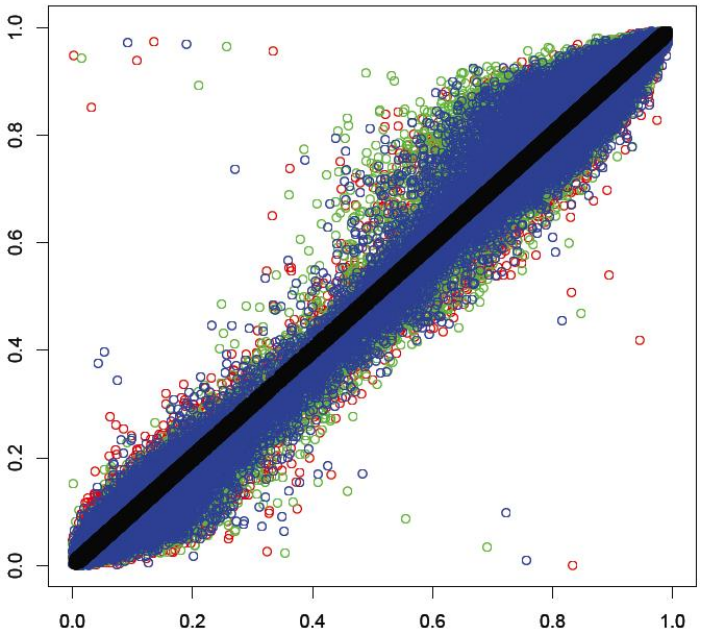


Figure 5f. XY plot examining the differences between batch 1 (X axis) and batch 2 (Y axis) using the sample duplicates run across both batches. The beta values of three sample duplicates (green red and blue) at probes identified by a 4th sample duplicate where there are <1% difference in beta values.

As a result of this plot showing non-linear inter-batch differences, no batch corrections were performed across both datasets as this would have introduced bias. Both datasets had been normalised individually, and the benefits are clearly seen between figures 5d and 5e, but no further normalisation was thought appropriate given the potential risks of over-normalisation.

5.3.3 Differential methylation

5.3.3.1 Three-way MVP call

A three-way MVP call was performed using an F-test on the 78 samples from batch 1 that had passed through the QC procedures. Of the 12 samples removed due to the chip failure, there were 29 group A, 25 group B and 24 group C samples suitable for analysis. The 3-way test (an F-test) identified 4506 MVPs with p values ranging from $4x10^{-6}$ to $1x10^{-2}$. Methylation differences within these MVP calls ranged from -20 to +20%. A sensitivity analysis using LOOCV reduced this list to 3441 MVPs, with p values between $8x10^{-6}$ and $1x10^{-2}$.

As described in the objectives of this study, the hypothesis was that the direct exposure groups (A and B) would show methylation difference with group C (born after famine) but that no assumptions would be made prior to discovery-based analysis of all 3 groups and the differences between them. After this 3-way MVP call, no post-hoc test was performed to identify what groups were driving the biggest difference, instead two-way (pairwise) calls were performed to localise which groups were driving the most of the methylation difference. This will be discussed in the subsequent section.

5.3.3.2 Two-way MVP calls

Two-way MVP calls were performed between groups AB, AC and BC in batch 1 to determine which groups were driving methylation differences, and in batch 2 for the purposes of validation. These MVP calls used the same procedure as for 5.3.3.1 except that they employed a t-test for the two-way comparison rather than an F-test.

The two-way calls were performed on batch 1 samples and batch 2 samples, independently of each other. The number of MVPs and their p-value ranges are described in table 5f. The two-way calls in the experimental dataset (batch 1) identifies that the largest number of MVPs occurs between groups A and C, i.e. those offspring born before and after famine. The comparison of A and B, offspring born before and during famine, produces the second highest number of MVPs. Comparison of groups B and C yielded the fewest MVPs.

The list of MVPs generated from batch 2 was not used in their own right as an independent dataset to perform further analyses, e.g. GO analysis. The purpose of the batch 2 calls was to validate batch 1 MVPs, and this validation will be discussed in the next section.

	Batch 1		**Batch 2**		**Batch 1 and 2**
	MVPs (n)	P value range	MVPs (n)	P value range	**Validated MVPs (n)**
AB	3672	$3x10^{-6}$ to $1x10^{-2}$	1467	$2x10^{-4}$ to $1x10^{-2}$	**2763**
AC	4039	$1x10^{-5}$ to $1x10^{-2}$	2335	$1x10^{-6}$ to $1x10^{-2}$	**2856**
BC	2237	$2x10^{-6}$ to $1x10^{-2}$	1983	$2x10^{-4}$ to $1x10^{-2}$	**1525**

Table 5f. MVPs identified in pairwise calls in Batch 1 and Batch 2, after LOOCV, with p value ranges. The MVPs highlighted in yellow indicate those identified in the first call (batch 1) that show similar directionality of methylation difference in batch 2, i.e. the validated MVPs.

5.3.3 Validation

The 55 samples that were processed as batch 2 were used as a validation set to see if they could add strength to the 3-way MVP call performed using batch 1. It is usual to have a smaller sample set as the experimental group and use more samples for validation as this enables less false negatives to be excluded from analysis. In this experiment, it was not clear that there would be funds to process all the samples at the outset and therefore the maximum number of samples was processed in the first round (n = 90, with 78 suitable for analysis). Subsequently, more funds became available and the remaining samples (n = 55, including repeats from batch 1) were processed and analysed.

5.3.3.1 Validation of Batch 1 MVPs

First, the MVPs identified in the 3-way call from batch 1 were compared to the MVPs called from batch 2. Due to the small sample size in batch 2, there was a negligible number of MVPs produced from a 3-way MVP call. The MVPs identified from the two-way calls in batch 1 and 2 also failed to overlap significantly. The lack of power to detect true MVPs in batch 2 was thought to be the reason that this approach to validation was unsuccessful, rather than and inappropriate rejection of the null hypothesis. An additional contributory factor in this inability to validate the MVPs directly against each other was the methylation variation produced by technical differences, or batch effects, between the batch 1 and 2.

In view of this lack of power and the batch effects, it was decided that a different approach would be taken to validate the MVPs from the experimental group (batch 1). The alternative approach used identified whether batch 1 MVPs were validated by a similar directionality of methylation difference across group comparisons in batch 2. This method maximised the potential from the two datasets that included an experimental group that had sufficient power to detect MVPs and a validation set that was underpowered. In taking this approach, it was seen that a high proportion of the MVPs identified from the experimental group were validated by the same directionality of methylation in the batch 2 dataset, provides support of the technical and biological validity of the MVPs identified.

In addition to the comparison of batch 1 MVP calls with directionality in batch 2, linear regression modelling was performed per pairwise MVP call (i.e. AB; BC; AC). Regression plots were drawn using the direction of change of beta values in the MVPs from batch 1 (X axis) and the data from the same probes in the same pairwise group comparisons from batch 2 (Y axis). In doing this, it was possible to examine the strength of the association and between the two datasets using the output of linear regression modelling, including the intercept of the model and the correlation coefficient (R^2).

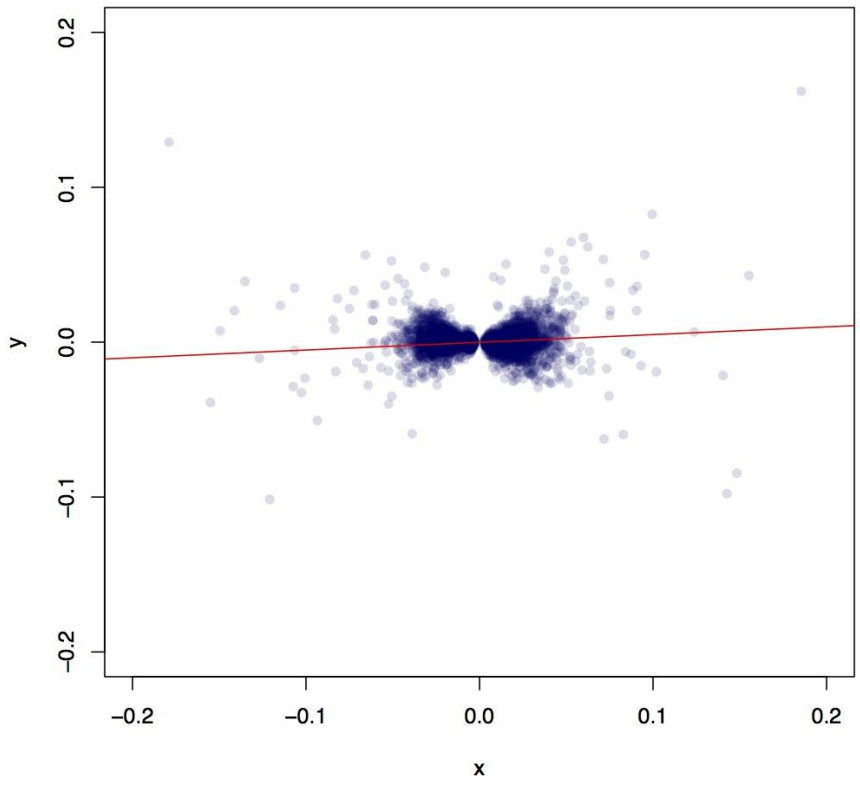


Figure 5g. Validation of Groups A vs. B MVPs. X axis shows the direction of change in batch 1 beta values and the Y axis shows direction of change in beta values from batch 2. Intercept = $^{-}7x10^{-7}$, y = $5x10^{-2}$

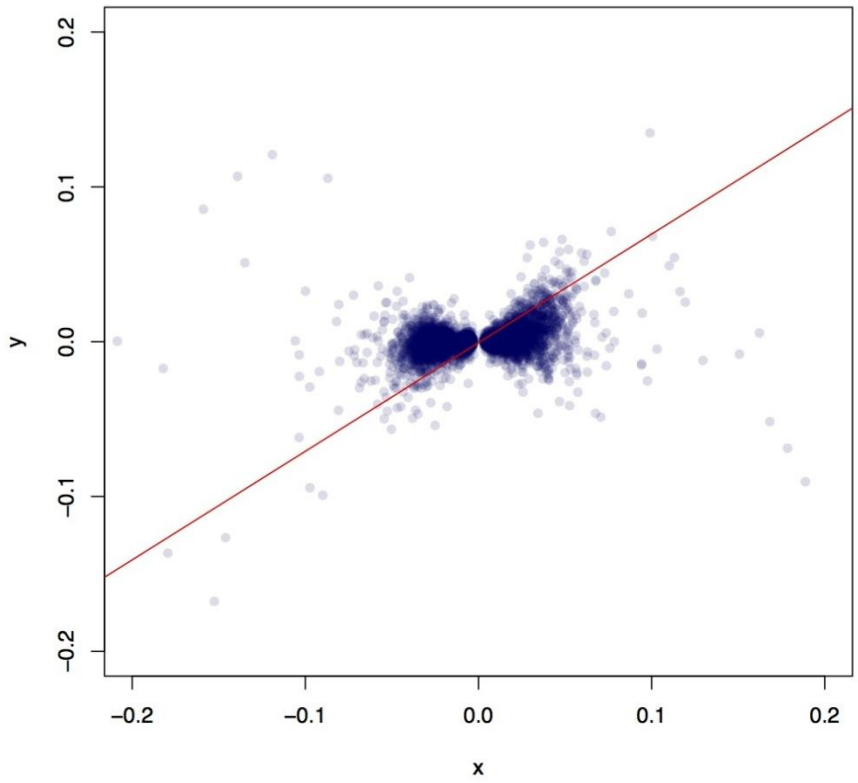


Figure 5h. Validation of Groups A vs. C MVPs. X axis shows the direction of change in batch 1 beta values and the Y axis shows direction of change in beta values from batch 2. Intercept = $^{-}7x10^{-4}$, y = 0.7

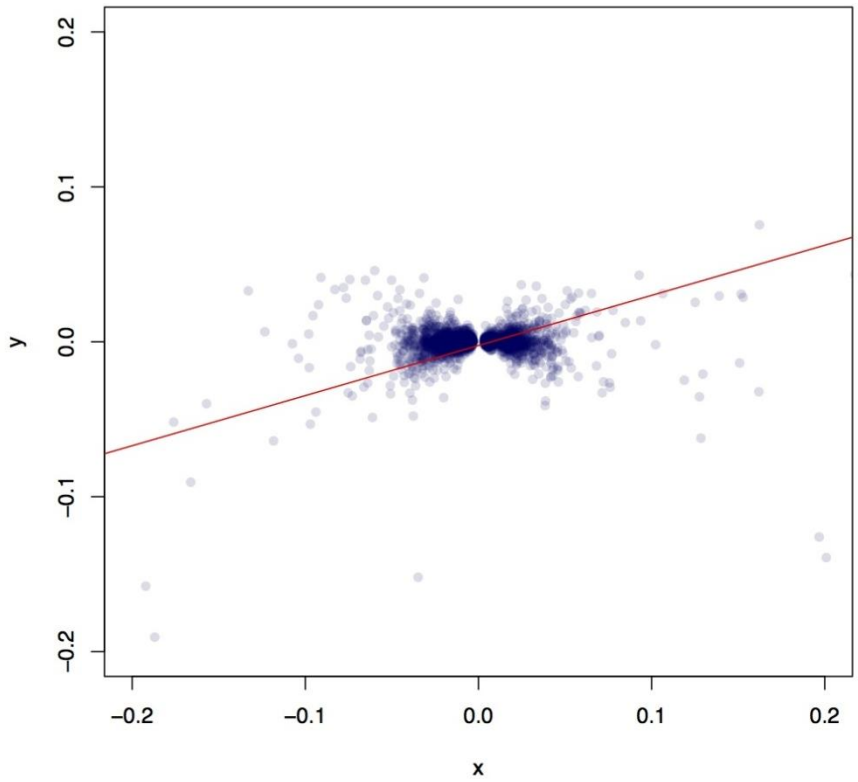


Figure 5i. Validation of Groups B vs. C MVPs. X axis shows the direction of change in batch 1 beta values and the Y axis shows the direction of change in beta values from batch 2. Intercept = $^{-}2x10^{-3}$, y = 0.32

These plots (figures 5g, 5h and 5i) show that the pairwise group A vs. group C comparison validates best, as shown by the high correlation (R^2 = 0.7) between the directionality of methylation variation in batch 1 MVPs and the direction of methylation difference when comparing the change in methylation variation between A and C at the same probes (figure 5h). This was the same comparison that yielded the highest number of MVPs and it is likely, therefore, that these are technically and biologically sound. Group B and C comparisons also validated, although the weaker association (R^2 0.32) probably reflects the lower number of MVPs called (figure 5h). The group A and B comparison did not validate, as shown by the absence of a clearly positive correlation between the two datasets (figure 5g).

The strength of association between AC and BC calls across both batches provides evidence that the methylation differences in these two calls is biologically sound as they have validated across different datasets that incorporate different samples. The BC validation has a lower R^2 value than AC, but this could reflect the fact that there are fewer data points to compare. The presence of reproducible MVPs across the datasets also provides some reassurance that the results of these experiments technically validate.

Identification of a large number of MVPs in AB batch comparisons that do not correlate in the linear regression model suggests that these are not true differences and are likely to represent what could be expected to occur by chance.

The validation of MVPs was further examined by identification of which overlapped the pair-wise analyses. The following Venn diagram (figure 5j) shows the numbers of MVPs from the batch 1 that have similar directionality in batch 2, i.e. the validated MVPs. Understanding that the AB MVPs do not validate, the 1454 MVPs that are identified in this comparison alone can be considered to be that which is expected by chance. In contrast, the AC (n=2122) and BC (n=958) MVPs are those that are likely to represent the true epigenetic differences in association with the famine exposure during pregnancy and in childhood, respectively. The 113 validated MVPs that overlap the AC and BC calls may therefore reflect an epigenetic signature that is present in famine-exposed individuals, irrespective of the timing of the insult. Finally, the 166 AC and 82 BC validated MVPs that overlap are likely to be the background 'noise' or false positives that have arisen from these pair-wise comparisons as these are present in the non-validated AB group-wise comparison.

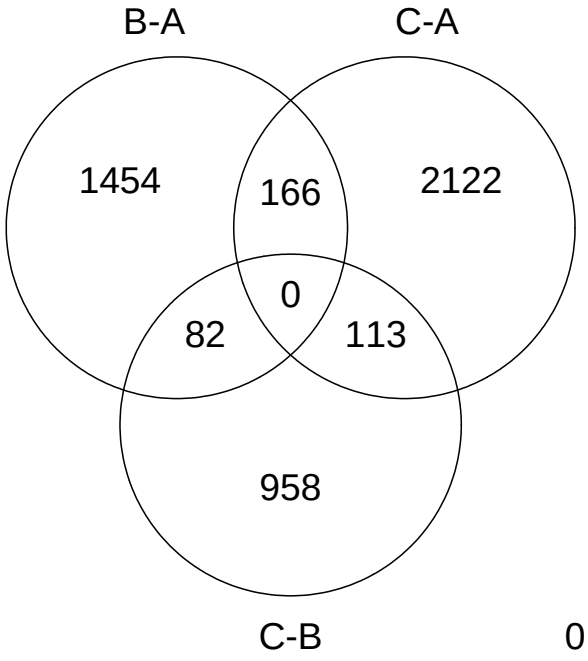


Figure 5j. Venn diagram showing validated Matlab MVPs from pairwise analyses.

5.3.3.2 Pathway analysis of validated MVPs

Gene ontology analysis was performed on the MVPs called from the A vs. C and B vs C analyses using the GOstats add-on to the Limma analysis package. The pathways over-represented in the MVP calls are summarised as GO terms in table 5g and 5h.

GO: ID	P value	Odds ratio	Expected count	Count	Term
GO:0071363	0.00000514	2.40	18	39	cellular response to growth factor stimulus
GO:0007173	0.00001105	2.76	12	28	epidermal growth factor receptor signaling pathway
GO:0010523	0.00002980	Inf	0	4	negative regulation of calcium ion transport into cytosol
GO:0043271	0.00004908	5.34	3	11	negative regulation of ion transport
GO:0030154	0.00005879	1.37	173	219	cell differentiation
GO:0007155	0.00010903	1.57	64	93	cell adhesion
GO:0042060	0.00012292	1.69	43	68	wound healing

GO:0009987	0.00017445	1.46	801	837	cellular process
GO:0009725	0.00028019	1.60	49	74	response to hormone stimulus
GO:0048011	0.00032918	2.11	16	31	nerve growth factor receptor signaling pathway
GO:0007399	0.00038618	1.39	115	149	nervous system development
GO:0048856	0.00039149	1.28	251	297	anatomical structure development
GO:0014051	0.00039588	25.13	0	4	gamma-aminobutyric acid secretion
GO:0010628	0.00039821	1.49	70	98	positive regulation of gene expression
GO:0001503	0.00043552	1.99	19	34	ossification
GO:0031644	0.00061165	2.14	14	27	regulation of neurological system process
GO:0071495	0.00062999	1.65	38	58	cellular response to endogenous stimulus
GO:0070887	0.00065755	1.39	100	131	cellular response to chemical stimulus
GO:0000381	0.00067820	7.55	1	6	regulation of alternative nuclear mRNA splicing, via spliceosome
GO:0033089	0.00086952	16.75	1	4	positive regulation of T cell differentiation in thymus

Table 5g. GO analysis of C vs. A validated MVPs.

GO: ID	P value	Odds ratio	Expected count	Count	Term
GO:0032990	0.00000024	2.44	22	48	cell part morphogenesis
GO:0048667	0.00000028	2.56	19	43	cell morphogenesis involved in neuron differentiation
GO:0048812	0.00000092	2.46	19	42	neuron projection morphogenesis
GO:0022008	0.00000135	2.02	36	65	neurogenesis
GO:0030030	0.00000338	2.08	29	55	cell projection organization
GO:0048468	0.00000996	1.87	39	67	cell development
GO:0000902	0.00001010	2.01	29	54	cell morphogenesis
GO:0048731	0.00002467	1.53	106	144	system development
GO:0048646	0.00003612	2.07	23	43	anatomical structure formation involved in morphogenesis
GO:0009887	0.00004974	2.05	22	42	organ morphogenesis
GO:0007411	0.00006910	2.45	12	27	axon guidance

GO:0007268	0.00008687	2.03	21	40	synaptic transmission
GO:0050885	0.00009400	6.51	1	8	neuromuscular process controlling balance
GO:0035637	0.00011184	1.96	24	43	multicellular organismal signaling
GO:0051960	0.00016227	2.28	13	28	regulation of nervous system development
GO:0033152	0.00017088	82.39	0	3	immunoglobulin V(D)J recombination
GO:0030182	0.00022028	1.97	21	38	neuron differentiation
GO:0007420	0.00026558	2.14	15	30	brain development
GO:0002009	0.00035023	2.36	10	23	morphogenesis of an epithelium
GO:0023051	0.00037237	1.54	61	87	regulation of signaling

Table 5h. GO analysis of A vs. B validated MVPs.

These GO analyses provide a useful insight into the possible biological processes affected by the methylation differences. The representation of growth factor signalling pathways in the AC analysis highlights a plausible functional regulatory pathway involved in the development of the programmed phenotype. The representation of GO terms including nervous system development, neuron differentiation and neurogenesis is similar to the GO analysis performed in the Pune experiments (see section 4.7.4) and that published previously and their inclusion may represent the normal methylation variation associated with ubiquitously-expressed genes in these pathways, or a role in relevant and common biological processes such as aging (92).

5.3.4 Cross validation of MVPs with related datasets

It could be considered that specific genomic regions are particularly susceptible to influence by an aberrations in the in utero environment, e.g. under-nutrition, and therefore that there might be some overlap between regions of methylation variation in studies following a similar methodology, such as the Pune study (chapter 4) and published work on samples from the Dutch Winter Hunger and Gambian season of birth studies. In addition to this, identification of SNP-dependent methylation at regions characterised in Chapter 3 may provide further technical and biological strength to the findings in this study.

5.3.4.1 FTO

The 450k array does not cover the FTO SNP rs8050136 identified as a CpG-SNP in Chapter 3 and there are no probes tiled to regions within its LD block. Therefore it was not possible to identify possible commonality across these two experiments.

5.3.4.2 Pune

There was minimal overlapping coverage of Medip-seq DMRs and 450k array probes, and therefore these datasets were not compared. A direct comparison of 450k data from the Pune study and Matlab studies has not yet been performed but will look for commonality in MVPs identified in both studies and whether this is more than would be expected by chance.

5.3.4.3 Dutch Winter Hunger and Gambian studies

The regions of differential methylation studied in the Dutch Winter Hunger papers were compared to the datasets generated from the Matlab experiments. The two experiments are not directly comparable, as the Dutch study did not use a genome-wide approach, instead taking a selection of candidate genes for DNA methylation studies. Table 5h shows a summary of the findings of this comparison. The *IGF2, IL-10, MEG3* and *GNASAS* loci studied in the Dutch studies are well represented by the 450k array.

MVPs were identified in *IGF2, GNASAS* and *MEG3*, and these MVPs validated between batch 1 and 2 analysis. The *GNASAS* and *MEG3* loci that overlap show hypermethylation in the Matlab exposed groups (groups A and B) compared to the unexposed (group C) and the direction of methylation difference is the same as in the Dutch study. The *MEG3* locus studied in the Dutch study is within 200bp of the 450k probe where the MVP lies (both are located within the gene promoter). The *GNASAS* locus in the Dutch study and 450k MVP are not co-located, but both reside within the gene body rather than the promoter. An MVP is located in the *IGF2* locus, but its direction of methylation difference is opposite to that of the Dutch study locus and this is likely to be explained by their different location within the gene (gene body vs. promoter, respectively).

The Gambian study (80) used a different experimental approach that started off genome-wide but honed down its MVPs into those that were present across tissue types and unaffected by genetic or structural variation, and thus has described stable epialleles. The two epialleles from the Gambian Study that show corresponding methylation variation in Matlab are hypermethylated in exposed groups relative to the unexposed. The genomic coordinates of

the Gambian epialleles are not published and therefore it has not been possible to identify whether these have a common location within the gene to the MVPs in the Matlab study.

Cohort	Gene	Matlab MVP identified	B-val difference in exposed vs control	Direction of B-val difference	Genomic location
Dutch Winter Hunger (exposure = conceived during famine)	***IGF2***	Yes	Hypomethylation	Opposite	Promoter (Dutch); gene body (Matlab)
	IL10	No			
	LEP	No			
	ABCA1	No			
	GNASAS	Yes	Hypermethylation	Same	Gene body (Dutch and Matlab)
	MEG3	Yes	Hypermethylation	Same	Promoter (Dutch and Matlab)
Gambian Study (exposure = conceived during rainy season)	***BOLA3***	No			
	EXD3	Yes	Hypermethylation	Same	Unknown
	PAX8	No			
	SLITRK1	No			
	ZFYVE28	Yes	Hypermethylation	Same	Unknown

Table 5h. Overlap between differentially methylated regions in Gambian and Dutch Winter Hunger Studies with Matlab.

The identification of commonality between the regions of differential methylation identified in the Dutch Winter Hunger and Gambian Study and the Matlab study provide reassuring external validity of the results.

5.4 Discussion

5.4.1. Discussion of findings

In this study, the long-term consequences of famine exposure during early development on DNA methylation patterns have been assessed. The Matlab famine study has established that in utero famine exposure can increase the risk of impaired glucose tolerance in the exposed offspring in early adulthood. These phenotypic findings are consistent with the Dutch Winter Hunger and Chinese Great Leap Forward studies and add further weight to the hypothesis that the developing embryo and fetus can be 'programmed' towards cardiometabolic disease by insults in the maternal environment. Preliminary insights from the study of candidate genes

associated with exposure to the Dutch Winter Hunger and Gambian studies suggest that long-standing epigenetic change from such insults can be stable and detectable from the study of whole blood DNA methylation patterns. The Dutch study suggests, but does not prove, that these epigenetic modifications are associated with the programmed cardiometabolic phenotype in the adults. Interestingly, the Dutch studies show that the strongest phenotypic outcome is associated with late gestational exposure but the epigenetic changes are associated with early gestational exposure. This finding could be due to the fact that the epigenetic marks were studied in whole blood, a tissue that may only carry the signature of a developmental insult that has occurred before the division of the embryological germ layers. The Dutch Winter Hunger epigenetic studies studied 122 case-control pairs and have the additional strength from having sibling controls to reduce the amount of genetic heterogeneity between cases and controls. However, the subjects studied were considerably older than in the Matlab study and therefore the DNA methylation patterns in the Dutch offspring may have become more heterogeneous and affected by ageing-associated epigenetic variation.

The study presented in this chapter builds on the preliminary evidence from the targeted Dutch and Gambian epigenetic studies and presents the first genome-wide view of programmed DNA methylation changes associated with early developmental famine exposure. The Matlab study offers additional perspective over the Dutch study by its 3 exposure groups that include pre-natal, in utero and postnatal exposure, as well as its targeting of a non-mobile community-based population in a region with a high diabetes risk in young adults.

The findings of this study will now be discussed in relation to the experimental hypotheses set out in section 5.1.

The first hypothesis suggested that direct famine exposure in early developmental life would be associated with epigenetic variation, and that the timing of this direct exposure, in utero or early postnatal life, may influence whether programming occurs. This study has identified changes in DNA methylation associated with direct famine exposure by comparing the sample 3 groups in the Matlab study. Although it was considered likely from the outset that direct exposure to famine in groups A and/or B would contrast with group C offspring who were not directly exposed, the initial analyses did not make this presumption. The processed data generated was first examined in a three-way analysis (using an F-test) to identify the most significant methylation differences that occurred between any of the 3 sample groups. This produced a list of 3441 MVPs within the experimental dataset of 78 samples that passed a

sensitivity analysis. These MVPs did not hold up to p value adjustment for genome-wide significance and this reflects a lack of power in the study to differentiate between true differences and false discoveries.

Following the 3-way MVP call, pair-wise group MVP calls were performed to understand which of the sample group comparisons were driving the most variation, and with which to perform validation of the findings with additional samples. These analytical steps address the second experimental hypothesis that suggests that the comparison of the 3 sample groups may identify the developmental time-points that are susceptible to programming. The pair-wise MVP calls were performed between AB, AC and BC groups and generated a much higher total number of MVPs in batch 1 (9948) compared to that generated by the 3-way call (3441). The comparison of groups B and C produced the least MVPs, and this is likely to be because of the smaller number of samples in these groups, compared to group A, in batch 1. There were fewer MVPs produced in all 3 pair-wise comparisons in batch 2 due to the smaller number of samples. The most important finding of the pair-wise analyses and comparisons is that the batch 1 MVPs were validated (by similar directionality of methylation difference) in batch 2 in AC and BC comparisons. This finding is significant, and as the batches contain samples from different individuals, it suggests biological validation of the findings, and therefore support that direct exposure to famine in utero and in postnatal life induces DNA methylation differences and that these are not seen in individuals born after, and therefore not directly exposed to, famine.

It is interesting to discuss the identification of DNA methylation differences in offspring exposed to famine postnatally. The notion that postnatal life, and in particular nutrition, can programme adult disease is well founded, but the identification of these marks in whole blood is perhaps surprising. Whole blood is not considered to be a tissue directly relevant to the pathogenesis of type 2 diabetes, unlike liver, muscle and pancreas and therefore do the MVPs identified have a direct role in the development of type 2 diabetes and the phenotype observed in this study? It is plausible to think that these MVPs have a separate role to the phenotype studied, and especially given that they are identified in the postnatally exposed group who do not appear to have a programmed phenotype. However, the GO analysis suggests that the epigenetic variants have a role in processes that could affect relevant homeostatic mechanisms for cardiometabolic disease, such as altered growth factor signalling. The second question that the use of whole blood raises in this study is whether the absence of a peri-conceptual insult (e.g. in group A) is further reason for whole blood not to show the memory of the environmental insult. It might be expected that a peri-conceptual insult would result in methylation disturbances during epigenetic reprogramming that could then be

transmitted to all 3 germ layers that have not yet diverged resulting in detectable marks in all tissues including whole blood. This argument is commonly used in to justify the need for relevant tissues to understand the aetiological and pathogenic role of epigenetic variation. However, whilst this theory can be easily applied to studies of cancer, where genetic mutations and aberrant epigenetic marks may accumulate within tissues, it is more difficult to understand in the context of environmentally determined epigenetic change. If an organism is exposed to a significant environmental insult, such as famine, it is not unreasonable to think that the epigenetic modifications could be induced in any tissues exposed to the insult. This is principle is exemplified by the ability to detect ageing-specific MVPs whole blood (92), which could themselves be considered as a mark of chronic environmental exposure, albeit less well-defined. Further work will inform whether tissue specificity is important as important in identifying epigenetic marks induced by the external environment of an organism as it is in disease models affecting specific tissues. Functional studies and pathway prediction of the AC versus BC MVPs may inform this question by highlighting, for example, aetiological pathways with an early developmental role in the BC comparison. It will important to consider these challenging questions in future study design, perhaps through the incorporation of different tissue types (where possible) or by sorting whole blood into cellular subtypes with relevance to the disease of interest. Trimester specific information about in utero exposure is also vital to understand the relationship between epigenetic processes and development.

The final hypothesis and objective of this study was to address whether famine exposure-associated epigenetic marks characterise the primary programmed event or are secondary to the phenotypic outcome of it. The identification of MVPs in offspring exposed in utero and postnatally, both susceptible to the cardiometabolic phenotype may indicate that the epigenetic events could be secondary events. However, the phenotypic data from Matlab suggests that only the in utero exposed display the phenotype of impaired glucose tolerance, and therefore the MVPs that are common to groups A and B should not be expected to be purely secondary. But the small sample number and limited analysis of the phenotypic data does not provide conclusive proof of the absence of a diabetic phenotype in Group A, and therefore it is difficult to draw a definite conclusion on this. Larger studies using this 3-group approach to study programming may have the benefit of being able to differentiate primary from secondary epigenetic events, and a better understanding of the epigenetic signature of the studied phenotypes will also facilitate this understanding.

5.4.2 Limitations of this study

There are several limitations of this study that need to be taken into account when interpreting its findings. Importantly, the MVPs identified did not hold up to multiple hypothesis testing correction and therefore they must be interpreted with caution due to the likelihood of false discoveries. In characterising the overlap between the non-validating AB group comparison, it is possible to generate an estimate of false discovery, but ultimately a larger sample size to validate the batch 1 MVPs is necessary. A second, important limitation of this study is the imbalance of males and females in each exposure group. Although X and Y chromosome probe data has been filtered out, there may still be sex-specificity in methylation elsewhere in the genome and it is possible that this may be driving some of the MVPs. This limitation could be overcome by performing an MVP call to compare male and female samples at autosomal probes to see if there are any MVPs common to those generated in the exposure group analysis.

The concern of all of the studies of epigenetic variation in fetal programming is that of their inability to prove causality over association. This study improves on others in its 3 group design, and the evidence of epigenetic variation in the early childhood exposure group who do not have the same adverse phenotypic outcome that is evident in the exposed in utero group suggests that the epigenetic marks may be causal, or at least pre-date the onset of phenotype. Further analysis of the phenotypic data from this study is going to be important to expand on this observation, and future studies should incorporate longitudinal sampling so that epigenetic marks can be studied before, during and after the onset of the programmed phenotype. Additional insights could also be gained from studying offspring in Matlab that have been part of the MINIMat randomised intervention study of micronutrient supplementation (187) and elucidating whether epigenetic changes are modifiable through nutritional supplementation.

The genetic tendency towards the cardiometabolic phenotype in this population is assumed from background data showing a high prevalence of type 2 diabetes and related conditions. There is a paucity of data proving that South Asians have a higher genetic risk of type 2 diabetes, although it is assumed that this is the case from population-wide observations of disease prevalence. This is an area that needs more research, and epigenomic studies need to incorporate an understanding of the association of epigenetic variants with genetic risk variants. This study does not incorporate any genetic insights, and is underpowered to do so in its own right. The Dutch Winter Hunger study group have begun to examine these interactions in their work (188) and show weak, but significant, interactions between SNPs and DNA methylation within the *IGF2/H19* locus. These kind of studies will require much larger sample

numbers to elucidate this interactions in more detail and allele-specific studies of epigenetic modifications will be important to characterise the precise molecular changes. Future work within this cohort could however include study design, such as that in Chapter 3, where risk haplotypes are studied in order to generate an understanding of cis control over DNA methylation.

Genetic variation is unlikely to cause bias between the three exposure groups in this study. However, an understanding of the complex interacting factors that might predispose to a diabetic phenotype in offspring with a combined environmental insult are important. Future studies might need to focus on whether certain parent-child genetic factors, such as polymorphisms in the glucokinase gene, could influence not only the genetic risk of disease within families, but also whether they may hold a particular susceptibility to abnormal fetal development in the context of environmental insult. It is also worth considering the evolutionary impact of famine exposure and whether there may be any differences in the ability of individuals to conceive during famine. It could be hypothesised that there is a selection pressure towards the 'fittest' individuals to reproduce during famine and that there is a genetic bias in the exposure groups in this kind of famine study. Again, these are much wider questions that are not going to be addressed by a study of this size, but are important considerations in the interpretation of these studies in their wider context.

The final important limitation of this study is how the 'control' group is interpreted in the context of famine exposure. The control group (group C) consists of individuals born to Matlab families 6 months after famine exposure, and therefore there was no direct exposure to these offspring. This does not, however, exclude indirect exposure to famine via the parental germline and the possible transgenerational transmission of epigenetic marks. This is a contentious area in epigenetic research at the moment, with much focus being put on the notion of 'epigenetic inheritance' and whether it is a commonplace occurrence in humans. Animal studies do show that epigenetic marks, including those which have been induced through the environment, may be transmitted across generations, but it remains unclear whether this always represents stable inheritance and whether these findings are applicable to human models (reviewed in (81)). Whilst research into this area is ongoing, it is important to consider the possibility that marks of parental exposure to famine may be transmitted to the offspring in group C and that this could be a limitation in the interpretation of these results. It is not possible to exclude these potential transgenerational influences in the Matlab study design as all individuals come from the same rural cohort that was universally exposed to famine. The exclusion of transgenerational epigenetic variants of famine exposure from this

kind of study would require the inclusion of non-famine exposed families, and in this context it would require a different population to be studied, itself bringing in a range of different biases.

5.4.3 Potential environmental confounders

The Matlab study did not characterise the exact nature of the famine exposure at the time and our data relies on post-hoc interpretation of the published studies describing the famine and direct experience within the community in Matlab and ICDDR,B. It is important to consider what other adverse environmental factors co-existed with famine exposure to interpret the results of this study. At the time of famine exposure, it can be assumed that malnutrition was associated with a high prevalence of diarrhoeal and other infective disease, already known to be common in Matlab. Reduced physical activity may also have been associated with famine, in parents or offspring directly exposed. The postnatal life of offspring born during famine may also have held lasting nutritional deficits, repeated infections as well as socioeconomic deprivation in families particularly affected by famine. It is also important to consider whether 'famine exposed' is an appropriate definition that applies equally and universally to all families living in Matlab at that time, or whether there was differences between and within households in the amount of nutrition available. It will not be possible to characterise any of these possible confounders retrospectively, and therefore it is important to consider their potential influence on the conclusions drawn form the study.

Another environmental influence that may confound the results of this study that has been characterised within Matlab is that of arsenic exposure via contaminated well water. This exposure has been documented and studied in Matlab, by ICDDR,B researchers, and evidence exists that maternal exposure to arsenic in pregnancy can influence thymic development in offspring and is independent of micronutrient supplementation (189). These observations are important to understand the complexities of environmental exposures and phenotypic outcomes, and interestingly may provide insight into new mechanisms to explain epigenetic variation. Research suggests that genetic polymorphisms of one-carbon metabolism influence arsenic metabolism (190) and that folic acid and vitamin B12 may reduce the oxidative stress induced by arsenic exposure (191). A recent study of mice exposed to arsenic in utero identified a range of DNA methylation differences when coupled with a high folate diet (192). Future plans in the Matlab study include the collation of arsenic exposure data and identification of any association with the programmed phenotype or epigenotype.

5.4.4 Applicability of this study to other fetal programming studies

The DNA methylation variation in this study in part overlaps with that identified in other studies, including the Dutch Winter Hunger and Gambian Studies that have a similar focus on nutritional insults in early developmental life. These findings provide support of the technological validity of these experiments, but also some reassurance of their biological validity too. Unfortunately, the approach that these two studies took to identify epigenetic variation was not genome-wide and therefore it is difficult to compare the results of these studies directly. The question that further research will need to answer is whether there are regions of the genome that are particularly susceptible to environmental insults in development, and if so, whether these are timing-specific and/or specific to particular nutritional insults.

Given that fetal programming of cardiometabolic disease is known to occur via specific micronutrient deficiencies, e.g. B12 deficiency in PMNS, it is also important to consider whether the famine exposure in Matlab could be due to the same factors or is indeed reflecting the much wider nutritional insult of famine. In the absence of specific characterisation of micronutrient status during famine exposure, it will not be possible to assess this further but must be held in consideration when drawing conclusions across these studies.

Chapter 6. Mouse Model of Gestational Diabetes

6.1 Introduction

Exposure to gestational diabetes has been described as a mode of fetal programming towards an offspring phenotype of type 2 diabetes and obesity in Chapter 1. Epidemiological studies provide some evidence of this model, and carefully designed animal models of a hyperglycaemic environment in late pregnancy provide important data in support of these observations (36). The means by which gestational diabetes may induce epigenetic change is poorly understood, however studies of histone modifications in a hyperglycaemic environment provide an important insight into possible mechanisms (104). The developing embryo is exposed to maternal hyperglycaemia via maternal-placental-fetal transfer, and itself may become hyperinsulinaemic, resulting in accelerated growth. Whilst these direct effects of gestational diabetic pregnancy are well described, the long-term outcome of these exposures is poorly understood.

A mouse model of gestational diabetes can provide useful insight into these mechanisms, offering the ability to study tissue-specific effects as well as long-term phenotypic outcome. An inbred mouse strain also offers an invaluable insight into pure epigenetic phenomena, independent of genetic heterogeneity that is evident in human studies and that can obscure the maternal-child transmission of risk.

This study has been designed to identify the potential mechanisms of fetal programming in the context of gestational diabetes, and uses a mouse model to replicate this disorder as much as possible. Heterozygote female mice of the BKS.Cg-m+/+Lepr^{db}/J strain (referred to in the simpler nomenclature, db/+), emulate human gestational diabetes in their leptin resistance and spontaneous glucose intolerance (due to insulin resistance) during pregnancy. This model has been well studied in pregnancy, but only one study so far has looked at the potential for these in utero effects to programme offspring (37). Six-month old wild-type (WT) offspring of db/+ mothers were found to have raised fasting plasma glucose and hyperinsulinaemia (female mice only) and were fatter than WT offspring of non-diabetic (+/+) mothers. These researchers studied the livers of these offspring and found decreased insulin-stimulated *Akt* phosphorylation and therefore impaired insulin signaling and hypothesise that this may reduce insulin-dependent suppression of hepatic glucose production. However, this study is limited in its phenotypic studies of these mice, and does not propose any mechanisms by which these insulin-signaling changes are induced. Epigenetic mechanisms are likely to play a role in mediating these effects, and by comparing WT offspring from both GDM and WT mothers, it is possible to eliminate the potential for genetic causes. Hyperglycaemia itself may affect the

epigenetic state of important genes, e.g. *NFkB*, and therefore their expression (104) (105). Therefore it can be seen that the glycaemic environment (i.e. from the mother) of a developing fetus could modulate the developing epigenotype and produce lifelong epigenetic aberrations. Altered feeding patterns and changes in leptin signaling may also be associated with changes in DNA methylation in the melanocortin pathway, an important regulator of body weight. The placenta may also play a role in altered growth patterns in the fetus and the studies have identified altered gene expression of a range of growth factor genes that could be mediated by epigenetic processes and reflect long-term aberrations that could predispose to insulin resistance and obesity (193). The discrepancy between maternal and neonatal levels of growth factors in diabetic pregnancies may highlight epigenetic dysregulation of gene expression in neonates in response to the maternal environment.

By using this mouse model of gestational diabetes, it will be possible to test the following hypotheses:

a) Does maternal gestational diabetes programme a diabetic and obese phenotype in offspring through non-genetic means?

b) Do changes in DNA methylation underlie this model of fetal programming?

c) Do these epigenetic changes occur in a tissue- or sex-specific manner, and can they be associated with gene expression and phenotypic outcome?

6.2 Methods

6.2.1 Mouse model

Home office approval for the study protocol was obtained, under appropriate personal and project licences.

6.2.2 Mice

Mice were purchased from Charles River at 6 weeks of age. A C57BL/6 background strain was common to all mice, and female mice were from the BKS.Cg-m+/+Lepr^{db}/J strain. The mice used are described in Table 6a.

Strain	Genotype	Number	Sex
C57BL/6		12	Males
BKS.Cg-m+/+Lepr^{db}/J	db/+ (heterozygotes)	25	Females
BKS.Cg-m+/+Lepr^{db}/J	+/+ (wildtypes)	24	Females

Table 6a. Number, sex and genotypes of mice used in mouse model

All mice were housed in the Biological Services Unit (BSU), Barts and the London School of Medicine and Dentistry, in individually ventilated microisolator cages (IVCs), with *ad libitum* access to food (standard chow) and water, 12-hour light/dark cycles. On arrival to the BSU, mice were housed in groups of 6, separated by sex and genotype.

6.2.3 Breeding model

Female heterozygotes (db/+) of the BKS.Cg-m+/+Lepr^{db}/J strain were used as a model of 'gestational diabetes', based on their pregnancy-specific hyperglycaemia previously identified by several researchers e.g. Friedman et al (37). Female wildtypes (+/+) from the same strain were used as controls. Dams were age-matched and in their first pregnancies. Female mice were bred with male C57BL/6 animals according to the following model (Figure 6a):

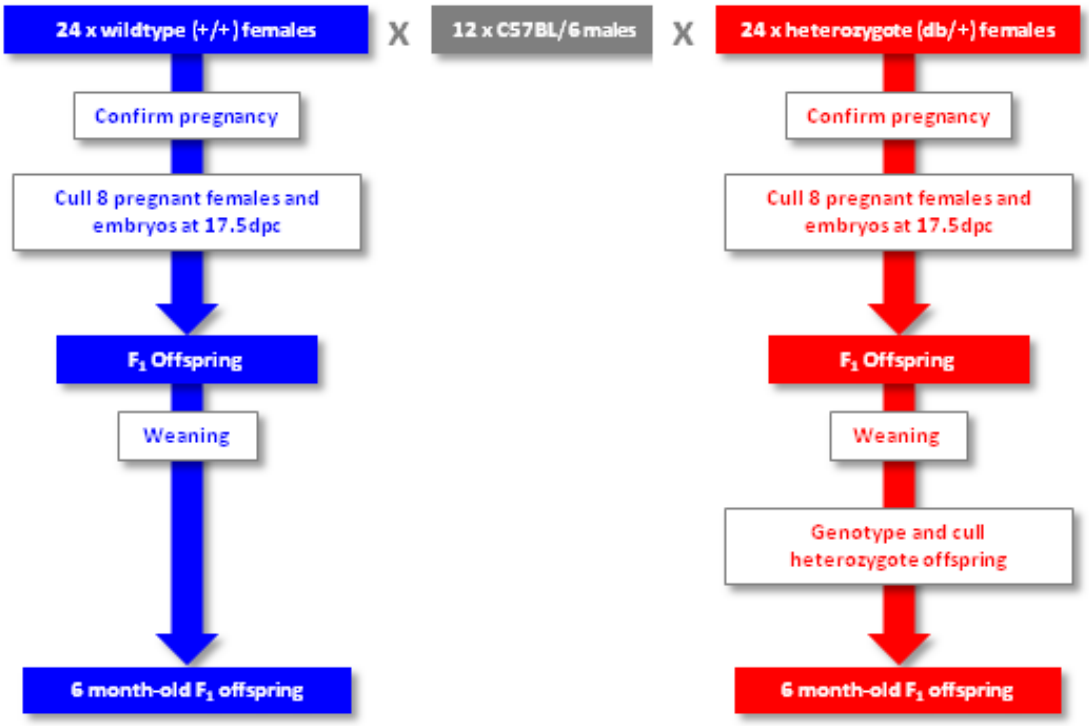


Figure 6a. Outline of mouse breeding model used in GDM experiments. Wildtype (control) animals and their offspring are shown in blue, and GDM animals and their offspring in red.

6.2.4 Phenotypic tests in pregnancy

Mice were paired overnight and conception confirmed by the presence of a vaginal mucus plug. Males were removed from females' cages after overnight pairings, and only re-paired after an interval of 5 days to ensure the precise timing of conception.

All female mice underwent intraperitoneal glucose tolerance testing (IPGTT) at 17.5 days post-coitum (dpc) to determine the glycaemic response to a weight-specific intraperitoneal glucose load. Mice were fasted overnight (from 6pm) and IPGTT was performed 14 hours later. A tail vein bleed was performed, the first drop of blood being used to test the glucose level using a glucometer. Another 20µl of blood was collected into a PCR tube and sat at room temperature for 1-2 hours to clot. Intraperitoneal glucose (2g/kg, from a 100mg/ml solution) was administered. Blood samples were taken 30 and 60 minutes after the administration of glucose. Clotted blood samples were spun at 4^{o}, 2000g for 15 minutes, and the serum layer was pipetted off into 20µl aliquots and frozen immediately at -80^{o}.

6.2.5 Cull and dissection of 17.5dpc females and embryos

After IPGTT, 8 db/+ and 8 +/+ females were culled in a CO_2 chamber. Maternal cardiac puncture was performed peri-mortem and blood was serum collected using the approach described above. Embryos were dissected from the uterus and, after isolation from their embryonic sacs, placenta and umbilical cord, were weighed. Subsequently, the liver, head, remaining body and placenta of each embryo were dissected and stored in individual Eppendorf tubes and frozen (within 15 minutes of cull) in liquid nitrogen. Female liver, heart, gastrocnemius muscle and abdominal fat were dissected. All serum and tissue samples were snap frozen in liquid nitrogen and transferred to -80^{o} freezers.

The remaining pregnant female mice were returned to their IVC cages, fed ad libitum, and continued their pregnancy. On delivery of their pups, mothers were put into individual cages with their litters and allowed to wean normally. Litter sizes were not standardised.

6.2.6 Cull and dissection of remaining F_0 generation

After completion of breeding and weaning of all litters, all male and female F_0 mice were culled and dissected using the same procedures described in 2.3.4.

6.2.7 Genotyping and cull by genotype post-weaning of offspring

DNA extraction was performed from embryonic mouse livers using the phenol:chloroform protocol described in Chapter 2. Offspring from heterozygote mothers were genotyped in order to maintain only wild-type offspring in each group for later study of pure intra-uterine differences, rather than genetic differences.

6.2.7.1 Standard PCR-restriction digest genotyping protocol

A standard protocol from the Jackson laboratory suggests genotyping using a PCR-restriction digest protocol. The mutant *Lepr* allele contains a G to T transversion, creating a donor splice site that induces alternative splicing and a 106nt insertion into the transcript. The suggested PCR protocol produces 135bp product from WT allele, whereas the
G to T transversion in the mutant mice introduces an Rsa1 restriction digest site (GTA) that is visible as bands of 27bp and 108bp after Rsa1 treatment. However, this reaction requires the reverse primer to incorporate the base substitution to provide produce a stable PCR amplification product. However, this primer design is unstable and suboptimal with the

formation of primer dimers and low annealing temperature. NetPrimer software calculates that these primers have a 69/100 rating, Tm 61^{o}, GC 33% with multiple primer-dimers and hairpins formed between the two primers. This protocol was abandoned due to its unreliability.

6.2.7.2 Tetra-ARMS PCR protocol

The Tetra-ARMS approach (194) has been developed for SNP genotyping, and uses two primer pairs to amplify the 2 different alleles of a SNP in a single PCR reaction. Two allele-specific amplicons were generated using two pairs of primers, targeted to the G and A alleles conferred by the target SNP, designed using specific software for this technique (http://cedar.genetics.soton.ac.uk/public_html/primer1.html). Mismatch between the allele and one of the two primer pairs (at its 3' end) confers allele specificity when the PCR is performed, and a second mismatch (*) is introduced at position -2 of the 3' end to enhance allelic specificity. The two outer primers are positioned at different distances from the target allele to allow for discrimination between the amplicons (and therefore determination of the allele) when the PCR product is run on a gel.

The primers used are longer than usual PCR primers to enhance the stability of the annealing to both target and non-target alleles. This ensures that differences in allele specificity are due to differences in extension rate rather than differences in hybridisation efficiency. The diagram below (Figure 6b) outlines the primer design around the target region in the *Lepr* gene:

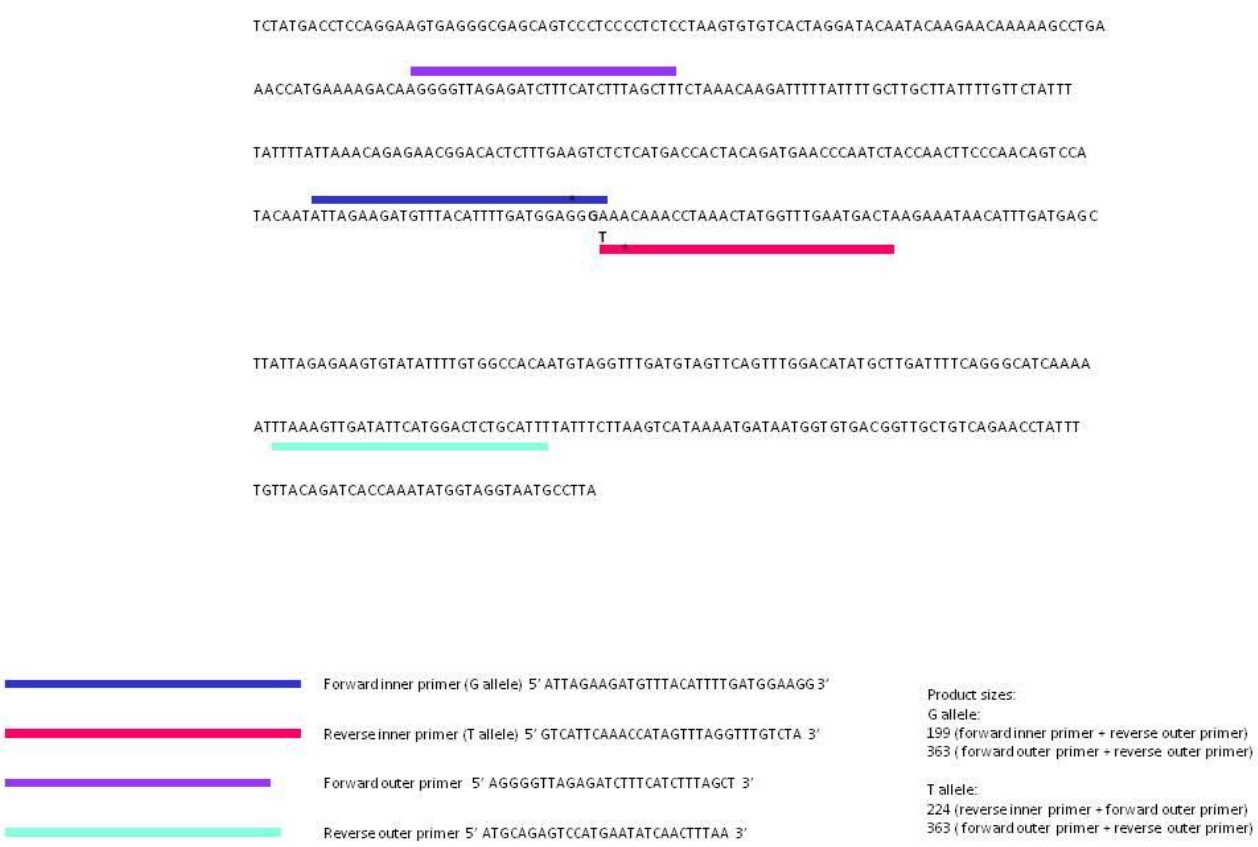


Figure 6b. Diagram of tetraARMS PCR designed to detect *Lepr* genotype

6.2.8 Ageing of F_0 offspring

All remaining offspring were maintained until 6 months of age in cages containing 1-7 animals from mixed litters. Ear marking was used to identify individual mice. Litter sizes were not standardised.

6.2.9 Phenotypic testing of aged offspring

All mouse offspring for whom there was maternal phenotype data underwent phenotypic testing using a three-hour IPGTTs following a 14 hour fast. The data presented shows the mean of the values in all mice from a single litter. The procedure for performing these tests and serum collection was the same as that described in 6.2.4 except that blood glucose measurements were performed at 0, 15, 30, 60, 90, 120 and 180 minutes. This change in

protocol was on the basis that the IPGTT results from pregnant mice were highly variable and therefore the incorporation of more measurements may improve the ability to detect differences. Advice was also sought from a colleague (R. Batterham, UCL) with extensive experience of mouse models of metabolic disease as to determining the adequacy of the intraperitoneal injection technique during IPGTTs. It was suggested that a 30 minute glucose <15mmol/l was likely to result from an inadequate or misplaced injection of glucose and that IPGTT results should be excluded if this criterion was not fulfilled. The fasting glucose (i.e. pre-injection) results from those IPGTTs deemed inadequate were still included in analysis.

6.2.10 Cull of aged offspring

After IPGTT, mice were culled and dissected. Four mice were provided with standard chow for 2 hours prior to cull, in order to later investigate the potential effects of fasting and stress on these mice in the future. Culling and dissection was performed as previously described, but in addition, gonadal and retroperitoneal fat pads were dissected out and weighed, and for a selection of mice, kidney and brain were also harvested.

6.2.11 Selection of offspring for epigenetic study

Embryonic mouse livers were used for epigenetic studies, comparing the WT offspring of WT and HT mothers (n = 6 vs. 6). The selection criteria to define the offspring for these experiments will be discussed in the relevant results section. Liver was selected as the tissue to study as it is readily dissected out of a 17.5dpc embryo under direct vision and it has an important function in glucose and fat metabolism. The study of liver in embryonic offspring also allows direct comparison with liver tissue in aged animals.

6.2.12 DNA extraction

As discussed above in relation to genotyping, DNA extraction from the embryonic mouse livers was performed using the phenol:chloroform protocol, including an RNase digest step. This approach was used to maximise the DNA yield from the small amount of liver tissue. DNA was extracted from all embryonic mouse livers and samples for Medip-seq analysis were selected.

6.2.13 Medip-seq library preparation (multiplexed)

Medip-seq library preparation was performed using the standard protocol set out in Chapter 2. A multiplexed approach was used, using mouse-specific adapter sequences and primers, and pooling of all samples into a single library.

6.2.14 Illumina GAIIx sequencing

Medip-seq libraries were processed by UCL genomics across a 7 lanes of a single flowcell, leaving one lane for a PhIX control.

6.2.15 Sequence read processing

This was performed using the same protocol as described in section 4.3.3.

6.2.16 Bioinformatic analysis

Analysis of Medip-seq data was performed using the same technique set out in 4.5.6. The Thomas Down DMR caller was used to generate a long list of DMRs, first in a group-wise (case vs. control) analysis, and subsequently in a pair-wise analysis using individual case-control sample pairs. Following this DMR call, USeq was applied to the processed and normalised data in a new DMR call, and those DMRs that overlapped all 3 approaches were defined as the 'top hits'.

6.3 Results

6.3.1 Phenotypic tests in pregnancy

Insulin and leptin resistance of mothers, and related factors such as body weight and glucose tolerance, were recorded during pregnancy in order to characterise maternal factors that could lead to 'programming' of the developing fetuses.

Heterozygote mothers (db/+) were heavier than wildtype mothers before and during pregnancy, and they did not undergo excess weight gain as a result of pregnancy (corrected for litter size) (see table 6b, figures 6c and 6d). The liver weights of GDM (db/+) mothers was significantly higher than in WT mothers, in those culled at 17.5dpc which would be consistent with excess insulin resistance (Figure 6e).

The potential for glucose tolerance testing to have induced a stress response and glucocorticoid production in these animals should be considered in the light of evidence to suggest that this might be a cause of programming itself (195). However, glucose tolerance testing was performed very late in gestation (17.5dpc) and in both study groups, and therefore the potential for bias is likely to be small.

	Wildtype		**Heterozygotes**		**unpaired t test**	
	n	mean +/-SEM	n	mean +/-SEM	95% confidence interval	p value
Body weight (pre-pregnancy)	23	19.6 +/- 0.30	25	24.9 +/- 0.44	4.1-6.3	<0.0001
Body weight (17.5dpc)	19	34.6 +/- 0.71	24	40.8 +/- 0.51	4.2-7.7	<0.0001
Weight gain during pregnancy	19	15.1 +/- 0.72	24	15.8 +/- 0.52	1.0-2.5	0.38
Weight gain during pregnancy (per embryo)	19	2.4 +/- 0.34	24	2.3 +/- 0.13	0.5-0.9	0.59
Litter size	19	6.8 +/- 0.5	24	7.5 +/- 0.38	0.6-1.9	0.32
Liver weight at 17.5dpc	10	1.54 +/-0.06	8	1.78 +/- 0.04	0.1-0.4	0.02

Table 6b. Maternal and litter phenotypic characteristics.

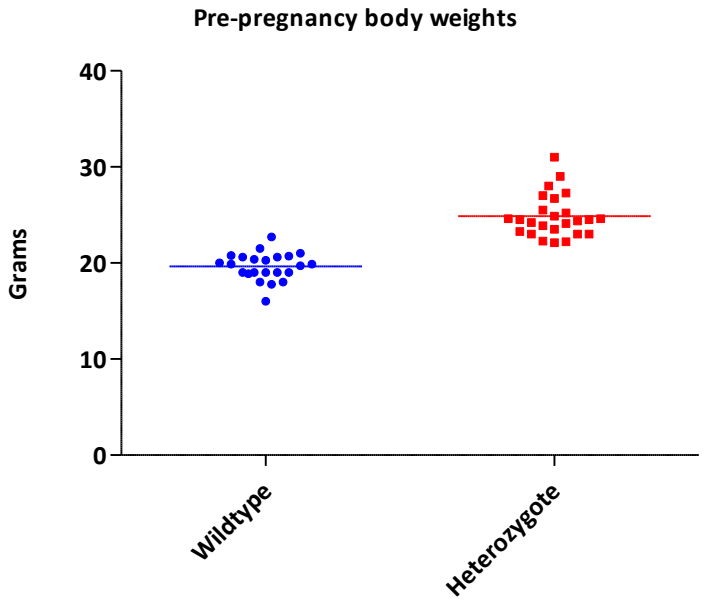


Figure 6c. Pre-pregnancy body weights of females. Differences in means statistically significant ($p<0.0001$)

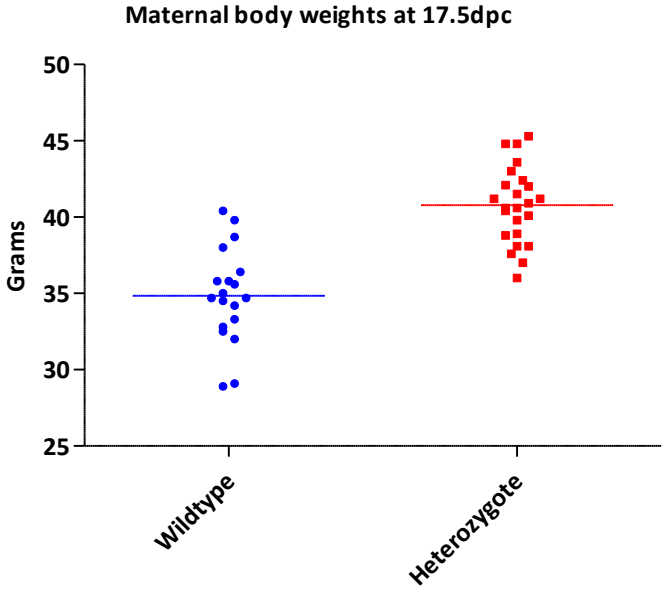


Figure 6d. Maternal body weights at 17.5dpc. Differences in means statistically significant ($p<0.0001$)

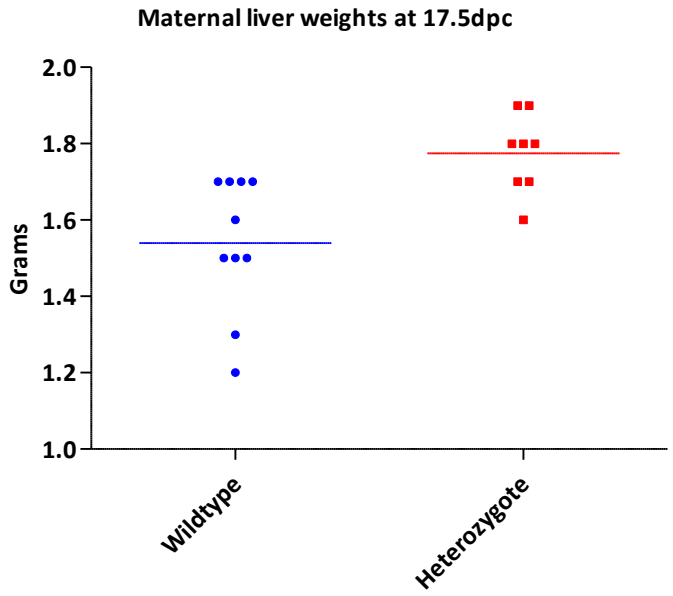


Figure 6e. Maternal liver weights at 17.5dpc. Differences in means statistically significant (p=0.02)

A one-hour glucose tolerance test was performed on the majority of pregnant mice at 17.5dpc (with the exception of 4 WT females due to inaccurate timing of their conception). The glucose tolerance tests identified higher basal (fasting) and post-challenge glucose levels in the heterozygote mothers (table 6c, figure 6f), confirming this model as one of relative maternal hyperglycaemia as expected from the literature.

Glucose (mmol/l) post-glucose challenge	**Wildtype**		**Heterozygotes**		**unpaired t test**	
	n	mean +/- SEM	n	mean +/-SEM	95% confidence interval	**p val**
0 minutes	19	5.8 +/- 0.27	24	6.6 +/- 0.19	0.1-1.5	**0.02**
30 minutes	19	11.3 +/- 0.66	24	14.5 +/- 0.65	1.4-5.1	**0.001**
60 minutes	19	9.8 +/- 0.68	24	12.8 +/- 0.91	0.6-5.4	**0.02**
Area under the curve (mean)	19	572 +/- 28.7	24	726 +/- 31.7	65.2-242.3	**0.001**

Table 6c. Results from glucose tolerance tests of female mice at 17.5dpc

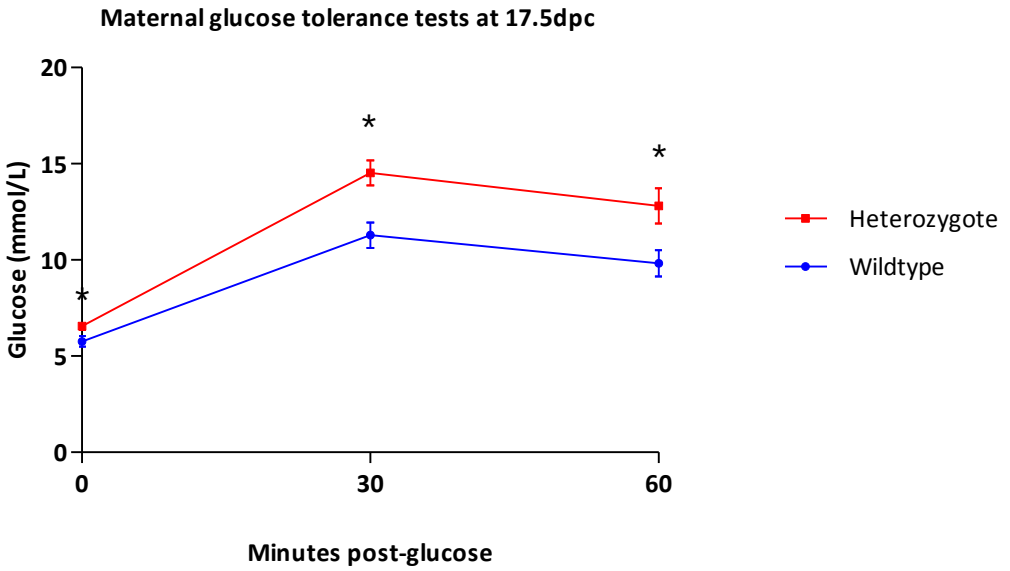


Figure 6f. Glucose tolerance test results of 17.5dpc mice, showing mean maternal glucose with SEM and statistically significant differences ($p<0.05$) are represented by *.

6.3.2 Cull and dissection of 17.5dpc females and embryos

Embryos of heterozygote mothers were genotyped as only WT offspring will be studied. There was an expected 1:1 proportion of WT (n=60) to HT (n=61) offspring.

6.3.3 Cull and dissection of remaining F_0 generation

Mice from the F_0 generation (i.e. parents) were culled at weaning and their tissues dissected, snap frozen and stored for later analysis.

6.3.4 Genotyping and cull by genotype post-weaning of offspring

Offspring of GDM (heterozygote) mothers were genotyped at weaning using tetra-ARMS PCR from ear snips and 61 heterozygous offspring were culled, leaving 60 WT offspring from HT mothers, in addition to 69 WT offspring from WT mothers. Sex genotyping was also performed on WT offspring and identified 34 male and 25 female offspring from WT mothers as well as 27 male and 33 female offspring to HT mothers.

6.3.5 Ageing of F_0 offspring

Mice were kept in mixed cages in standard conditions for 6 months.

6.3.6 Phenotypic testing and cull of aged offspring

In male offspring to HT mothers, increased body weight and liver weight were observed, relative to offspring of WT mothers (table 6d; figures 6j and 6m). These body weight differences appear to be due to an increase in fat mass within the retroperitoneal fat pad. Offspring of HT mothers were more glucose intolerant than controls, as determined by a higher AUC of the 3-hour glucose tolerance test, and this effect was driven by differences from 60 minutes onwards, indicating the role of glucose disposal mechanisms (Figure 6g and 6i).

Following IPGTT, all animals were culled and tissues were dissected out and stored, as described in the methods section. Female offspring showed the same body and liver weight differences (table 6e; figures 6j and 6m), and more consistent differences in both gonadal and retroperitoneal fat pads (table 6e, figures 6k, 6l). Female offspring of HT mothers were also more glucose intolerant, as judged by the AUC of the glucose tolerance test, but the differences were driven by changes in the early part of the curve, including the fasting value (Figures 6h and 6i). This observation supports the findings of Yamashita et al (37).

IPGTTs from 5 offspring of WT mothers and 3 offspring of HT mothers were excluded due to inadequacy of the tests, judged by a 30-minute glucose of <15mmol/l. These IPGTTs were performed on day 1 and 2 of experiments and were likely to reflect a lack of practice in performing these tests. Exclusion of these results may have weakened the ability to detect phenotypic differences between offspring as more results were excluded from WT offspring and these would have strengthened the finding that these offspring had lower glucose levels.

In addition, three IPGTTs (all from offspring of wildtype mothers) were also terminated early due to intramuscular injection of glucose at the time of testing. This was immediately apparent at the time of testing due to difficulty in performing the test and a misplaced injection site, with the tip of the needle passing through the retroperitoneal space and hitting a bony structures around the posterior aspect of the pelvis or vertebral column.

Beyond these excluded data, the variability of glucose measurements may, in part, be due to the stress of the animals whilst undergoing glucose tolerance testing. The stress response, mediated by catecholamine and glucocorticoid production, could cause hyperglycaemia in these animals via gluconeogenesis. However, the glucose tolerance testing was performed in a controlled fashion with optimal environmental conditions, and any bias due to these effects would be expected to be common to both groups and therefore should not obscure true differences.

	WT offspring of wildtype mothers		WT offspring of heterozygote mothers		unpaired t test	
	n	mean +/-SEM	n	mean +/-SEM	95% CI	p value
Body weight		35.9 +/- 0.67		38.4 +/- 1.08	0.0-4.9	0.05
Liver weight		1.28 +/- 0.03		1.51 +/- 0.06	0.1-0.3	0.0006
Gonadal fat pad weight		1.55 +/- 0.11		1.81 +/- 0.14	0.1-0.6	0.15
Retroperitoneal fat pad weight		0.58 +/- 0.04		0.73 +/- 0.05	0.0-0.3	0.04
Liver weight (as % of body weight)	34	3.57 +/- 0.01	27	3.93 +/- 0.09	0.2-0.6	0.0008
Gonadal fat pad weight (as % of body weight)		4.18 +/- 0.24		4.56 +/- 0.26	0.3-1.1	0.28
Retroperitoneal fat pad weight (as % of body weight)		1.57 +/- 0.10		1.86 +/- 0.11	0.0-0.6	0.06
Glucose (mmol/l) post-glucose challenge						
0 minutes	34	7.2 +/- 0.33	27	7.6 +/- 0.28	0.4-1.4	0.27
15 minutes	21	21.7 +/- 0.75	21	21.9 +/- 0.84	2.1-2.5	0.86
30 minutes	33	22.8 +/- 0.83	26	24.4 +/- 0.93	0.9-4.0	0.21
60 minutes	33	21.5 +/- 1.12	26	25.1 +/- 1.16	0.4-6.9	0.028
120 minutes	33	12.2 +/- 0.86	26	17.1 +/- 1.60	1.7-4.9	0.006
180 minutes	33	8.1 +/- 0.36	26	9.9 +/- 0.79	0.2-3.5	0.028
AUC (mean)	33	2777 +/- 130	26	3365 +/- 178	157-1019	0.008

Table 6d. Phenotypic characteristics of all male offspring at 6 months' of age

	Offspring of wildtype mothers		Offspring of heterozygote mothers		unpaired t test	
	n	mean +/-SEM	n	mean +/- SEM	95% CI	p value
Body weight		28.3 +/- 0.91		32.7 +/- 0.65	2.1-6.5	0.0002
Liver weight		1.08 +/- 0.03		1.21 +/- 0.02	0.0-0.1	0.0001
Gonadal fat pad weight		1.23 +/- 0.14		1.96 +/- 0.23	0.3-1.1	0.0003
Retroperitoneal fat pad weight		0.58 +/- 0.08		0.96 +/- 0.07	0.2-0.6	0.0006
Liver weight (as % of body weight)	25	3.83 +/- 0.10	33	3.74 +/- 0.06	0.1-0.3	0.37
Gonadal fat pad weight (as % of body weight)		4.05 +/- 0.37		5.83 +/- 0.29	0.8-2.7	0.0003
Retroperitoneal fat pad weight (as % of body weight)		1.89 +/- 0.22		2.86 +/- 0.16	0.4-1.5	0.0006
Glucose (mmol/l) post-glucose challenge						
0 minutes	25	5.8 +/- 0.25	32	6.6 +/- 0.20	0.2-1.4	0.016
15 minutes	13	19.1 +/- 0.72	26	21.8 +/- 0.70	0.4-5.0	0.02
30 minutes	18	21.0 +/- 0.88	30	22.8 +/- 0.69	0.4-4.1	0.1
60 minutes	18	16.1 +/- 1.07	30	18.6 +/- 1.10	0.9-5.7	0.15
120 minutes	18	8.3 +/- 0.74	30	8.8 +/- 0.70	1.6-2.7	0.6
180 minutes	18	6.2+/- 0.46	30	6.5 +/- 0.37	1.0-1.4	0.69
AUC (mean)	18	2190 +/- 100	30	2431 +/- 89	38-521	0.09

Table 6e. Phenotypic characteristics of all female offspring at 6 months of age

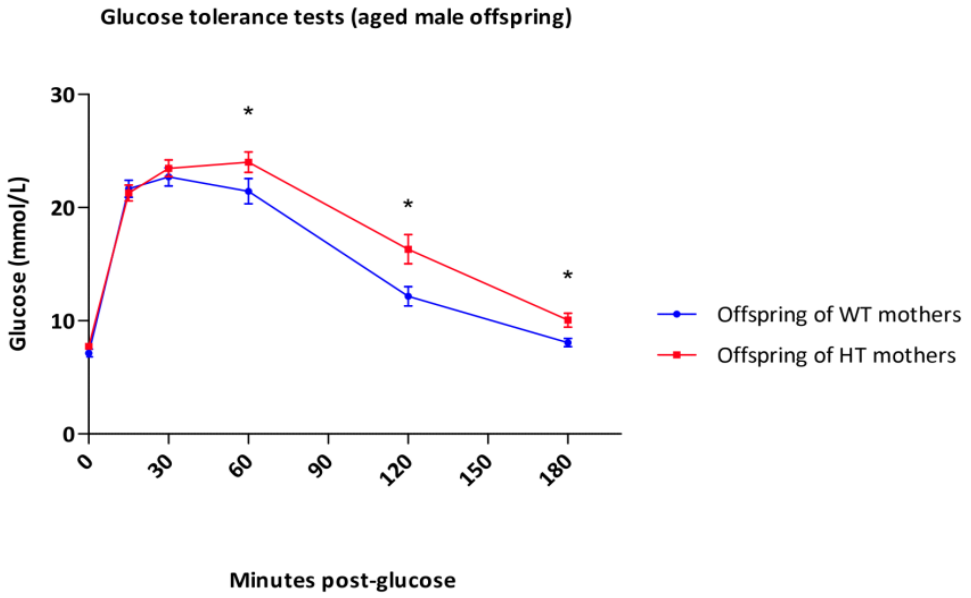


Figure 6g. Glucose tolerance test results of aged male mice. Mean maternal glucose with SEM and statistically significant differences ($p<0.05$) are represented by *.

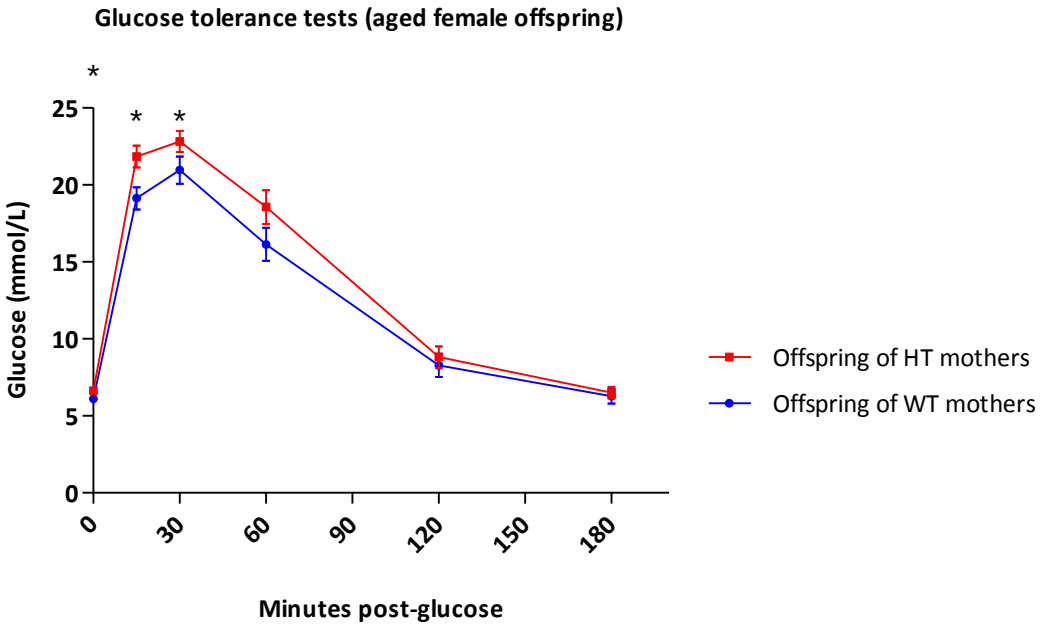


Figure 6h. Glucose tolerance test results of aged female mice. Mean maternal glucose with SEM and statistically significant differences ($p<0.05$) are represented by *.

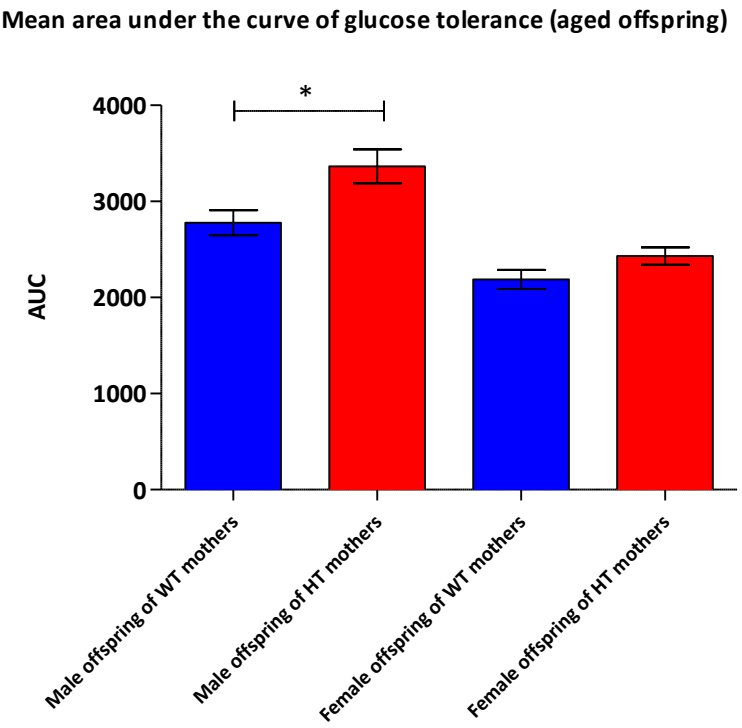


Figure 6i. AUC analysis of glucose tolerance tests with SEM and statistically significant differences ($p<0.05$) are represented by *.

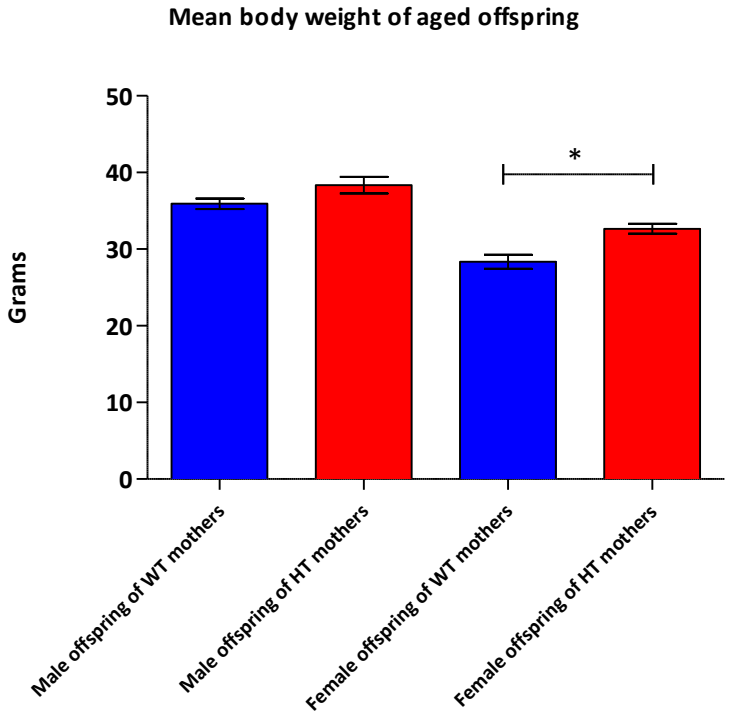


Figure 6j. Mean body weight of aged mice with SEM and statistically significant differences ($p<0.05$) are represented by *.

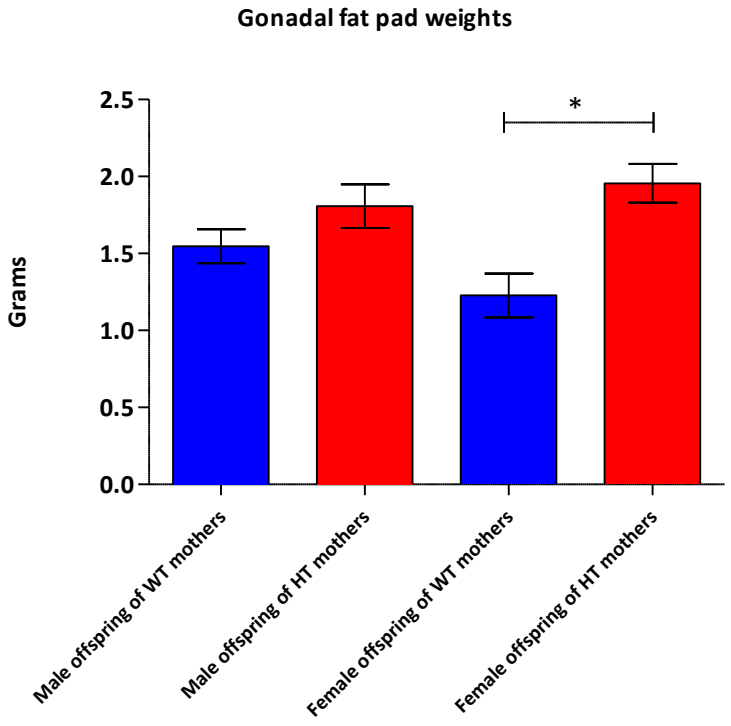


Figure 6k. Mean gonadal fat pad weights with SEM and statistically significant differences ($p<0.05$) are represented by *.

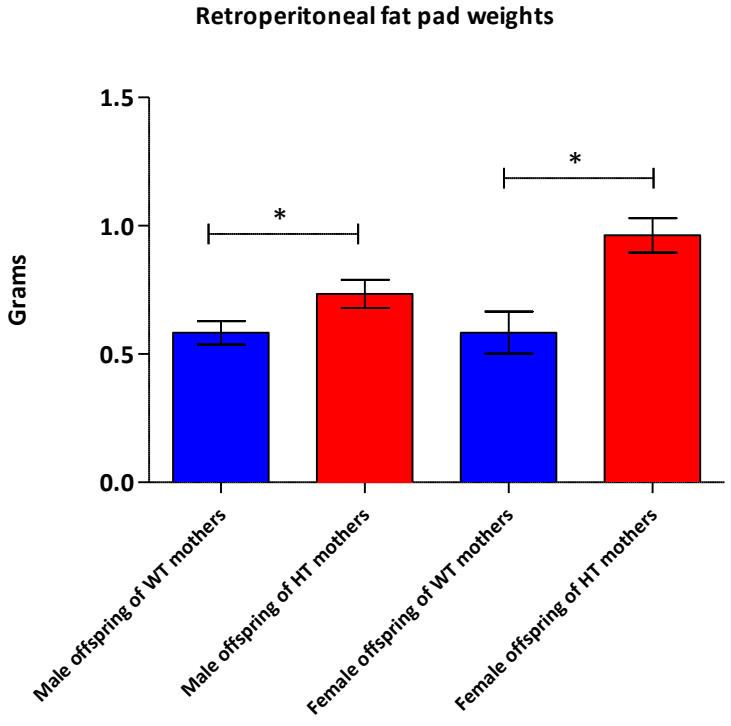


Figure 6l. Mean retroperitoneal fat pad weights with SEM and statistically significant differences ($p<0.05$) are represented by *.

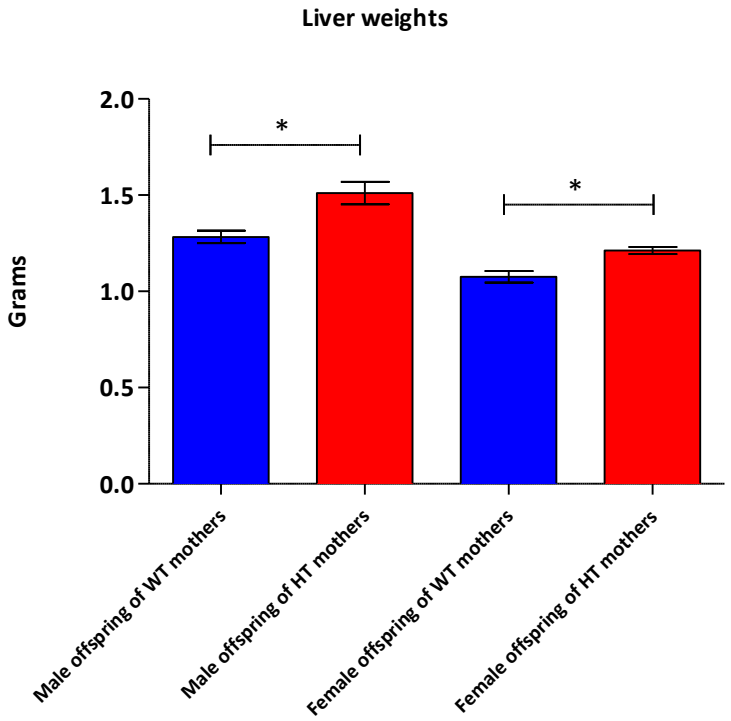


Figure 6m. Mean liver weights with SEM and statistically significant differences ($p<0.05$) are represented by *.

6.3.7 Selection of offspring for Medip-seq experiments

Medip-seq was performed on liver tissue from 17.5dpc female embryos. It was decided that female offspring should be studied, not male, due to the stronger programmed phenotype evident from measurements of glucose and fat mass. Six embryonic livers from offspring of WT mothers in the highest quartile of glucose tolerance and six from offspring of GDM mothers in the lowest quartile of glucose tolerance (i.e. most glucose intolerant) were selected for the initial study and identification of DMRs. No more than one offspring per mother was selected and that nearest the first of a litter was selected. Prior to library preparation, all samples were checked again for the correct *Lepr* and sex genotype and these were confirmed.

6.3.8 Library preparation

All QC steps confirmed successful library preparation. All 12 multiplexed samples were pooled after individual library preparation to allow gel extraction of a single, and therefore size-matched, band.

6.3.9 Illumina GAIIx sequencing

The standard sequencing protocols were followed and met all routine QC checks in the process. Raw sequence read counts (see table 6f) seemed consistent with successful sequencing and their quality was interpreted in more detail during the bioinformatic workflow.

			Read length statistics (pre-processing)						Read processing						
	Tag	**Sample**	**Min**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**	**Raw**	**Mapped**	**Q10**	**Paired**	**100-160bp**	**Pairwise Normalisation**	**% Passed**
Cases	ATCACG	1	0	119	132	130.1	144	254	18,647,664	17,174,068	12,419,776	11,989,304	10,883,590	10,882,418	58.4
	CGATGT	2	1	119	132	130.1	144	254	24,453,337	22,970,169	17,056,904	16,715,048	14,939,453	13,270,571	54.3
	TTAGGC	3	0	121	133	131.8	145	251	28,462,599	26,025,053	15,231,579	14,850,734	13,611,688	12,466,250	43.8
	TGACCA	4	0	121	133	132.2	145	252	23,041,735	21,233,900	14,522,487	14,178,550	13,002,142	12,991,657	56.4
	ACAGTG	5	0	120	132	131	144	252	23,100,200	21,168,472	12,534,229	12,243,588	11,164,502	10,370,423	44.9
	GCCAAT	6	0	120	132	131.4	144	252	24,326,299	22,071,608	13,035,306	12,554,988	11,484,304	10,442,957	42.9
	Mean cases		0.2	120.0	132.3	131.1	144.3	252.5	23,671,972	21,773,878	14,133,380	13,755,369	12,514,280	11,737,379	50.1
Controls	CAGATC	7	0	119	131	129.8	144	252	23,636,623	21,867,400	14,961,929	14,692,898	13,270,571	13,270,571	56.1
	ACTTGA	8	1	121	133	132.3	145	261	24,726,878	26,025,053	14,515,250	14,210,120	13,044,899	12,991,657	52.5
	GATCAG	9	0	119	132	130.1	144	255	22,116,100	20,458,933	14,021,964	13,775,088	12,470,014	12,466,250	56.4
	TAGCTT	10	0	121	133	132.2	145	251	21,481,662	19,733,557	11,923,476	11,629,054	10,442,957	10,442,957	48.6
	GGCTAC	11	1	121	133	132.2	145	252	23,882,272	19,606,799	13,924,632	13,633,844	12,463,697	10,882,418	45.6
	CTTGTA	12	0	122	134	132.8	145	252	20,296,846	18,602,289	11,590,292	11,330,660	10,404,136	10,370,423	51.1
	Mean controls		0.3	120.5	132.7	131.6	144.7	253.8	22,690,064	21,049,005	13,489,591	13,211,944	12,016,046	11,737,379	51.7
	Total		**3**	**1,443**	**1,590**	**1,576**	**1,734**	**3,038**	**278,172,215**	**256,937,301**	**165,737,824**	**161,803,876**	**147,181,953**	**140,848,552**	**50.9**

Table 6f. Medip-seq sequencing read counts through before, during and after read processing.

6.3.10 Bioinformatic analysis

6.3.10.1 Sequence reads pre-processing

The standard sequence read processing steps were performed and showed predictable and comparable losses at each filtering step (table 6f). As all samples were from female mice, there was no filtering of the X and Y chromosome reads that was performed in the Pune experiments. After mapping, pairing, quality filters were applied, as well as removal of PCR duplicates, there were similar rates of attrition across samples with a mean of 51%% of mapped reads suitable for analysis. This was considerably less than when similar processing was performed in the Pune experiments (see 4.6.3.1), however the reduction was systematic across all samples and therefore was thought likely to be a difference in sequencing rather than anything during library preparation that could have introduced bias between individual samples.

Calibration plots were drawn using case-control pairs and it was found that Medip enrichment across all case samples vs. control samples appeared equal (see figure 6n) and this was reassuring for the purposes of group-wise comparisons. Individual Case-control pairs were also matched according to Medip enrichment (as described in 4.6.3.1) for pair-wise DMR calling using the Thomas Down algorithm.

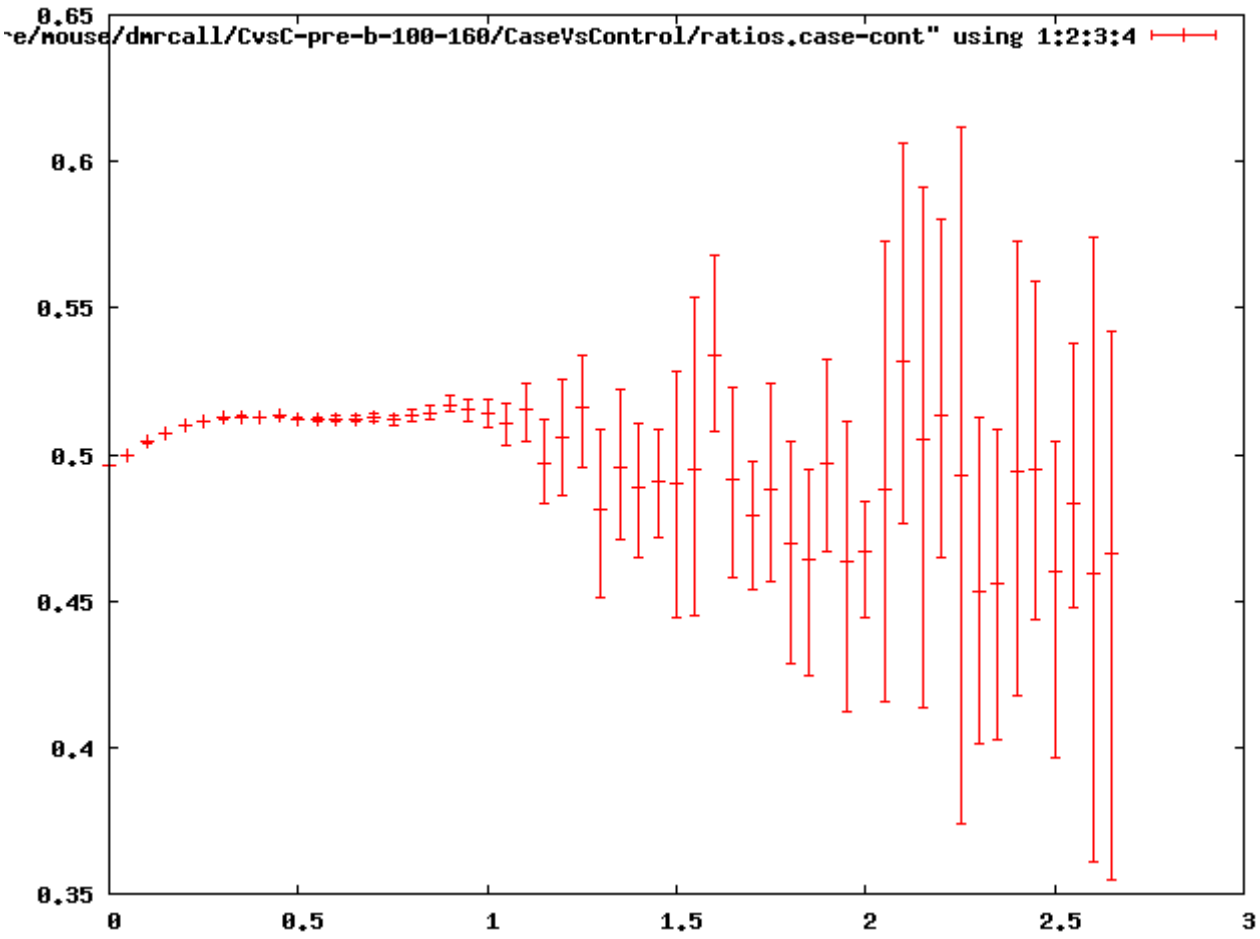


Figure 6n. Calibration plot of one case-control pair to assess Medip-seq enrichment efficiency

6.3.10.2 DMR calling

DMR calling was performed using fewer sequence reads per sample, compared to in the Pune experiments in which there was a minimum of 15 million reads per sample. It was decided that a second flowcell would not be run due to cost constraints and the fact that differences in methylation should be more readily detectable in an isogenic mouse model.

Chris Mathews performed DMR calling using a combination of bioinformatic algorithms, the Thomas Down custom DMR callers and USeq (see table 6g), combining all methods to generate a 'top hit' DMR list

	DMR caller	Method	Size Normalisation	Enrichment matched	Test used	Number of DMRs
Single method	Thomas Down CaseControl	Pairwise Caller (pooled case and control groups)	Yes	No	T-test	**34,993**
	Thomas Down Pairwise	Pairwise Caller (individual case/control samples paired using enrichment plots). DMRs common to all pairs.	Yes	Yes	T-test	**15,218**
	Useq	Useq	No	No	Negative binomial distribution	**92**
Combined methods	Final Combination	Thomas Down CaseControl + Pairwise + Useq	n/a	n/a	n/a	**12**

Table 6g. DMR counts using different bioinformatic algorithms

Despite the fewer reads per sample in this experiment, the Thomas Down calls yielded similar numbers of DMRs to Pune experiment 2. This is surprising given that these samples come from an isogenic mouse model and therefore there are no SNP-dependent methylation changes that are highly variable and therefore prone to being called as DMRs. However, USeq generated significantly fewer DMRs compared to the Pune experiment, and the element of false discovery control using this method, as well as its probabilistic approach, are well-applied to a dataset such as this that comprises biological replicates within each experimental group. Given the knowledge that inbred mouse strains may have inter-individual genetic differences due to the presences of structural variation, the USeq analysis was performed with varied set window sizes. There was no disproportionate change in the number of DMRs identified according to the window size being increased or decreased. Varied window sizes were applied to X chromosome reads (for ease of computation) and whilst 7 DMRs were called using 500bp windows, there were 48 identified with 250bp windows and no DMRs identified if 2000bp windows were applied. These data make sense, as a smaller window size would be expected to identify DMRs that are less consistent over a genomic region, and a further assumption could be made that these DMRs would be less likely to be significant. An increased window size would be expected to reduce the number of DMRs called (as in this example) unless a different form of variation, e.g. structural variation, was driving the difference. On the basis of this analysis and interpretation, it was decided that the 500bp window was optimal for further analysis.

6.3.10.3 DMR characteristics

The DMRs identified by the combined analysis are described in Table 6h and 6i. There were 5 DMRs hypermethylated and 7 hypomethylated in cases vs. controls. Half of the DMRs were located in genes and all of these were intronic. Four of the intergenic DMRs were at, or within 2kb of, regions of the mouse genome with possible regulatory function determined by UCSC data. The putative regulatory features denoted by UCSC tracks from Encode and ORegAnno included regions of enrichment for p300 (a transcriptional activator), RNA polymerase II, (gene transcription), DnaseI hypersensitivity (denote active gene expression) and CTCF binding (gene insulation) in the samples studied by these databases. The majority of DMRs are located in repeat-rich regions, something that was also found in Medip-seq DMRs in the Pune study, the relevance of which will be discussed later.

The genes at which DMRs are located seem to have functional relevance to the mouse model and programmed phenotype in aged offspring. DMR4 is located within the intronic sequence of *Asxl2*, a gene that has recently been described as having a role in adipogenesis in a human pre-adipocyte cell line. Overexpression of the gene was shown to promote adipogenesis and this appeared to be mediated through modulation of *PPARα* and *PPARg*. Interestingly, this DMR is hypermethylated in cases compared to controls, and the location of this DMR in a gene body, in which hypermethylation may be expected to increase gene expression, an observation that would fit with the presumed increased adipogenesis in cases. DMR3, located within an intron of *Npc1l1*, could also have an important role in promoting adipogenesis and dysglycaemia through its expression in the liver and role in enhancing intestinal and biliary cholesterol absorption (196). Again, the expected result of the hypermethylation seen in cases at the *Npc1l1* intron would fit with overexpression of the gene and the end phenotype of increased fat mass and hyperglycaemia.

DMR	Cases vs. controls	Genomic location			Useq (data from 'best windows')			Thomas Down			Genomic features			
		Chrom	Start	Stop	#Window	FDR	Diff*	Case reads	Control reads	Reads difference	Repeat	CpG island	Gene	Genic region
1	**Hypermethylated DMRs**	chr1	192505673	192506174	1	14.3	1.3	138	65	73	no	yes	**no**	intergenic
2		chr4	148931475	148932010	3	11.8	1.8	74	31	43	yes	no	***Ctnnbip1***	intron 4
3		chr11	6119659	6120284	4	12.0	1.8	58	27	31	yes	no	***Npc1l1***	intron 10
4		chr12	3481927	3482454	2	10.5	2.1	47	18	29	yes	no	***Asxl2***	intron 5
5		chr16	24056029	24056530	1	10.9	1.6	89	53	36	no	no	**no**	Intergenic
6	**Hypomethylated DMRs**	chr4	48066275	48066775	1	8.7	-2.0	6	35	29	no	near	***Nr4a3***	intron 1
7		chr4	86370153	86370836	3	9.3	-1.3	49	113	64	yes	no	**no**	Intergenic
8		chr6	135113039	135113532	1	10.0	-1.9	21	50	29	no	no	***Hebp1***	intron 1
9		chr8	45770616	45771095	1	7.8	-2.0	8	35	27	no	no	**no**	Intergenic
10		chr15	48311014	48311576	3	16.4	-2.1	9	46	37	yes	no	***Csmd3***	Intron
11		chr16	37635843	37636340	1	10.9	-2.1	9	29	20	yes	no	**no**	Intergenic
12		chr17	81577021	81577634	4	13.2	-1.3	36	95	59	yes	no	**no**	intergenic

Table 6h. Genomic locations and characteristics of DMR tophits. * = Log2((sumT+1)/(sumC+1))

Cases vs. controls	DMR	Gene	Genic region	UCSC genomic features	Gene function
Hypermethylated DMRs	**1**	no	intergenic	Nil	
	2	*Ctnnbip1* (ICAT)	intron 4	Repeat-rich region. 10kb downstream from CTCF binding site, 10kb upstream from 3' UTR.	Inhibits beta-catenin and *TCF4* (a T cell transcription factor). Beta-catenin-interacting protein. Widely expressed, including in liver.
	3	*Npc1l1*	intron 10	Dense repeat sequence	Niemann-Pick disease, type 1. Critical for intestinal and biliary cholesterol absorption. A target for ezetimibe (J Lipid Res, 2012). Expressed in liver. Regulated by *PPARα* in humans.
	4	*Asxl2*	intron 5	Repeat-rich	Promotes fat cell differentiation, via PPARs (Park, JBC, 2011). Also a regulator of bone mineral density and osteoclastogenesis (Farber, Plos Gen, 2011). In drosophila, it is a trithorax with a dual role in transcriptional repression and activation
	5	no	intergenic	adjacent (upstream) to ORegAnno and TFBS, overlying microsatellite repeat	
Hypomethylated DMRs	**6**	*Nr4a3*	intron 1	Adjacent to 5'UTR, minimal repeat	Nuclear receptor subfamily 4A3 - in humans, member of steroid-thyroid hormone-retinoid receptor subfamily. Multiple other metabolic functions including glucose and lipid metabolism (Pearen, Mol Endocrinol 2010).
	7	no	Intergenic ?promoter	Adjacent to TFBS with p300, RNA pol2 and DNaseI peak. 15kb upstream of Dennd4c 5' UTR	Dennd4c is a guanine nucleotide exchange factor required for *GLUT4* translocation in adipocytes (Sano, JBC, 2011)
	8	*Hebp1*	intron 1	500bp downstream of ORegAnno, 4kb upstream of 5' UTR	Heme binding protein 1. A transcription factor expressed in the liver.
	9	no	intergenic	2kb upstream of ORegAnno. Repeat rich	
	10	*Csmd3*	intron	Moderate p300. Repeat rich	CUB and Sushi multiple domains 3 – possible role in juvenile myoclonic epilepsy and autism
	11	no	Intergenic	Adjacent to CTCF (500bp downstream) and p300 (500bp upstream) sites, 4 kb downstream of Hgd 3' UTR. Repeat rich.	Hgd is required for tyrosine catabolism. Human mutations cause alkaptonuria
	12	no	Intergenic	2kb downstream of p300 peak. Repeat rich with microsatellite repeats. Slc8a1 15kb downstream	

Table 6i Genomic details of DMR tophits

6.4 Discussion

6.4.1 Mouse model

The mouse model used in these experiments exploited a genetic tendency towards diabetes in pregnancy in *Lepr* heterozygote females, comparing these to wildtypes. By breeding heterozygote ('GDM') and wildtype ('control') mothers, with standard C57BL/6 males, it was possible to study the effects of maternal diabetes in pregnancy on offspring. By selecting and studying only the wildtype offspring from both 'GDM' and 'control' mothers, it was possible to isolate the effects of the maternal environment and exclude any genetic tendency towards diabetes in the offspring. This model has been used before to study fetal programming in response to maternal GDM, and showed raised fasting glucose and fat mass in 24-week old female offspring born to GDM mothers. In this study, which used larger numbers, increased body weight, fat mass and liver weights were seen in offspring (male and female) and glucose intolerance was noted in male and female offspring, indicated by the differences at some (but not all) glucose measurements during 3 hour glucose tolerance tests. There were some differences between male and female offspring in terms of their programmed phenotype, and the increase fat mass was more evident in females than males, however, the data does support the presence of a programmed phenotype in both and the variation may result from variability of testing and/or sample size.

The study by Yamashita et al showed that the maternal GDM phenotype was associated with, and perhaps caused by, overfeeding. Furthermore, by using taking a pair-feeding approach to some of the pregnant mice, it was possible to reduce the programmed phenotypic outcome in their offspring. This study has also identified that the programmed offspring (at 6 months of age) displayed hepatic insulin resistance, as shown by decreased hepatic *Akt*-phosphorylation and increased hepatic glucose-6-phosphatase activity. These findings were sufficient evidence to support the use of this model in the study of GDM-associated fetal programming, however they did not attempt to explain why this programming may have occurred, nor what was occurring on a genome-wide level.

This study therefore adds to this evidence in support of fetal programming via the maternal GDM environment, and the study of DNA methylation in the livers of embryonic offspring born to GDM and control mothers has produced insights into the underlying mechanisms by which it occurs. By studying the epigenetic consequences of maternal GDM exposure in 17.5dpc embryos, it is possible to conclude that this is a programmed event occurring in utero, rather than during post-natal life. The only caveat to this is that the *Lepr* heterozygote mice were

slightly heavier to their wildtype counterparts pre-pregnancy, and therefore it is conceivable that the programmed event could have occurred prior to pregnancy through the effects of the different genotype on the developing germ-line.

It is not possible to define the specific environmental factor that has induced programming, nor whether it results from a combination of factors. The maternal environment in this experiment was known to vary according to glucose tolerance and fat mass and it can be assumed from other studies using this model that the GDM mothers were also hyperinsulinaemic, insulin and leptin resistant, and that they had different feeding behaviours. Within this model, it would not be possible to identify which of these factors are important, other than perhaps by pair-feeding. It might be possible to use other models, such as that described by Gauguier et al to identify whether hyperglycaemia alone is the programmable influence. However, as the *Lepr* mouse model of GDM is analogous to the range of aberrant metabolic conditions of human gestational diabetes, it is perhaps unimportant to study this further. Should it be necessary to study this further, cell line experiments could also be designed to expose cells from specific tissues to altered environmental conditions, such as glucotoxicity.

Additional areas of study using this mouse model would be to identify whether the programmed outcome was dependent on litter size, sex of offspring, or male characteristics. However, within this study, the emphasis was not put on these possible influences and the size of the model was probably too small to dissect out these fine details.

6.4.2 Medip-seq

The Medip-seq library preparation experiments appeared to be technically-sound, however, post-sequencing quality control filtering resulted in a loss of approximately 50% of sequenced reads. Although some loss of reads is expected, there was a greater attrition than in the Pune experiments. It is not clear as to why there were greater losses and as the other quality control checks, were satisfactory and even across all samples, it is hoped that the filtered data is of high quality.

6.4.3 DMRs

The identification of differential methylation in the Medip-seq dataset has used two different bioinformatic algorithms and selected a 'shortlist' of DMRs according to those where there is

overlap. Further discussion of the limitations of this approach will be included in Chapter 8. In contrast to the Pune study, in which 44 DMRs were identified, only 12 DMRs were found in this study. This was an unexpected finding as it was thought that the absence of genetic variation between these mice would facilitate the detection of pure epigenetic difference. However, the 12 samples from this study were sequenced over a single flowcell (in contrast to the two flowcells in the Pune experiment) and therefore it is likely that it was more difficult to detect differences due to the smaller dataset. Interestingly, there were similar numbers of DMRs called by the Thomas Down algorithms across both studies, and it was not until USeq was used that the detection of DMRs fell. This is likely to be because the Useq algorithm is more stringent and has filtered out those DMRs that were being driven by difference originating from a single sample.

The DMRs that have been identified in this experiment are located within regions of the genome with a putative functional role towards the programmed phenotype. The majority of DMRs are genic, and of these, 100% are intronic and overlie genes that have a function in adipogenesis and glucose/insulin metabolism. Furthermore, the genes identified appear to be expressed in mouse liver, again supporting these being 'true' DMRs rather than false discovery. However, this study is only preliminary and does not exclude the possibility of the DMRs being chance events, and further experiments will need to be performed to address this important question. In the first instance, technical validation of these DMRs will be performed using targeted assays of DNA methylation (BS-pyrosequencing) and in a larger set of embryonic mouse tissues. These targeted studies will also interrogate the methylation state at *Akt* to see whether there is varied methylation that might associate with the altered expression in programmed offspring livers described in Yamashita et al.

The location of the DMRs at putative regulatory features gives some suggestion that, if true, they could have important functional effects on gene activity and higher order chromosome functions. Again, this question will require further study with experiments focused on transcriptional regulation and gene expression rather than DNA methylation.

Another finding from this whole genome interrogation of DNA methylation is that the majority of DMRs are located at repeat-rich regions. This was also a feature of Pune Medip-seq DMRs and the wider context of this finding will be discussed in more detail in Chapter 8. However, the relevance of this finding to mouse experiments is significant in the light of recent understanding of the genomic architecture of inbred mice. Whilst it is assumed that inbred mice colonies are isogenic and derived from well-characterised strains with limited haplotype diversity, studies of the commonly used laboratory mouse strains have shown unexpected diversity and uncharacterised structural variation within strains (197). Whilst it is unlikely that

the mice used in this experiment, which were all purchased from one animal laboratory, vary significantly from each other, it is important when comparing genetic data from them to reference databases and genome browsers.

The final point of consideration in regard to the DMRs identified in this study are that they are only from the female offspring born to this model. It will be important to replicate the findings of this experiment in male offspring to see if there are any sex-specific differences.

6.4.4 Other limitations

There are other, wider limitations to this study that must be discussed in the light of the findings of differential methylation in programmed offspring. The means by which mouse tissues were dissected and stored for DNA extraction did not allow for histological analysis of cell types within the tissues. The programmed phenotype in offspring could have influenced the tissue directly and resulted in a different proportion of cell types within the liver. The liver is composed of several different cell types, including hepatocytes, endothelial cells and Kupffer cells and each is likely to have its own specific DNA methylation profile. Therefore, differential methylation may in fact represent a difference in cell types studied, rather than a true epigenetic difference consequent on programming.

There are many studies in the field of fetal programming that make an association between maternal stress, glucocorticoid levels in utero, and an adverse phenotypic outcome (e.g. (198)). Whilst this area of research remains contentious, it is important to consider these findings in the context of this animal experiment. The environment of the laboratory mouse in these experiments could be considered as stressful, in particular during the time of glucose tolerance testing, tail snipping for genotyping and during culling. During these procedures, the mice were visibly stressed and, in the case of the glucose tolerance tests and the long fast prior to them, this stress lasted many hours. Theoretically, the same techniques/stresses were applied to all mice and therefore should not bias the results that compare the GDM vs. control-exposed offspring, however stresses are by nature unpredictable and it is not possible to fully control for this in mouse experiments. Furthermore, there are always possible means by which cases and controls could be differently exposed, e.g. by heavier animals being more difficult to pick up, and administer intraperitoneal injections.

The final potential limitation of epigenetic studies in isogenic mouse models should be discussed in the light of the issue of sample contamination in the Pune study. Sample processing and library preparation, especially when cases and controls are processed randomly

to avoid systematic bias, is prone to sample mix-ups and potential contamination. With outbred mouse or human studies, it is possible to detect sample contamination during genetic analysis, however it is not possible to when isogenic mice are being studied. Whilst this is not thought to have been a problem within this study due to very cautious laboratory processing, it is an important factor to consider when analysing datasets.

6.4.5 Next steps

As discussed above, the next steps will be to validate the 12 DMRs identified in this experiment in the larger sample set of embryonic mouse livers, to include male and females and across the whole spectrum of maternal glucose tolerance. This will provide an important technical validation of the Medip-seq DMRs, but may also produce some insight into quantitative differences according to maternal glucose status and sex-specificity. The ability to design appropriate primers for this analysis may be hindered by the amount of repeat sequences in these regions, but it is hoped that it will be possible to achieve sufficient coverage of some of the CpG sites within the 500bp Medip-seq windows.

In this sample set, the possibility of inter-individual structural variation occurring must be considered, especially due to the observation that there are known microsatellite repeat sequences within some of the DMR regions, as determined by UCSC. Should funds and DNA quantity permit, an Agilent CGH array will be used to assay for copy number variation and other structural variants within these samples.

The next approach to understanding the mechanisms underlying fetal programming in these samples will be to perform gene expression studies in the embryonic livers as well as liver and adipose tissue from aged offspring. The aim of these studies will be to determine the functional effects of the programming event in the two most relevant tissues to the phenotypic outcome. The Illumina WG6 mouse expression array will be used to give genome-wide coverage and generate an expression profile across 24 case and 24 control offspring in all tissues. Liver tissue expression profiles will be compared between embryonic and 6-month old (aged) offspring so that developmental changes associated with the programming can be assessed, as well as to identify associations with phenotypic outcome. Adipose tissue is hoped to provide a more homogenous tissue than liver with which to identify tissue-specific differences, and individual fat pads (gonadal) will be used in this analysis. The application of expression analysis has been chosen over the study of more DNA methylation profiles or other epigenetic modifications with the understanding that expression differences between animals exposed and unexposed to GDM should be a good means of targeting further regions for

epigenetic studies. The uncertainty over the validity and interpretation of Medip-seq has suggested that this approach may yield a more hypothesis-driven approach whilst still having the advantage of a genome-wide and unbiased approach. It will also be possible to design additional targeted studies (or genome-wide studies if funds permit) in the other tissues that have been stored from the mice in this experiment, e.g. placenta, muscle, brain.

Following these expression studies, an integrated approach to data analysis will be taken, using Ingenuity software. The aim is to be able to associate gene expression differences with DNA methylation profiles from Medip-seq analysis. Ingenuity will facilitate this understanding with its ability to draw together interacting genomic influences on epigenetic and functional levels. It will be important to consider the influence of both *cis* and *trans* effects and their role in controlling gene expression. The regions of methylation difference identified in this Medip-seq experiment suggest that they may have a role in transcription factor binding, a finding that would mean that an epigenetic variant could have wide-ranging effects across the genome and its function. It may be necessary to design studies that identify specific transcription factors or their binding sites, e.g. ChIP-chip, ChIP-seq or newer techniques that identify sequence-specific TF DNA binding to putative TFBS using dsDNA microarray or SELEX-seq (199). By taking this integrative approach, it is hoped that regulatory networks can be identified and understood in the context of this model of fetal programming.

Finally, a crucial downstream application of this study is to integrate its findings with those of human studies of gestational diabetes (see Chapter 7). This mouse experiment has been used as a model system with which to identify the molecular mechanisms of fetal programming from gestational diabetes that would be difficult to elucidate in a human model. The particular advantages of this animal model over a human one is the ability to study tissue-specific epigenetic variation, long-term outcomes in 'aged' animals, and to study epigenetic events that are independent of genetic variation. The following chapter will discuss a human model of gestational diabetes and put into context the complexity of the environmental differences in such a study, where multiple confounding variables may co-exist with gestational diabetes due to converging risk factors.

Chapter 7. Human Model of Gestational Diabetes

7. 1 Introduction

7.1.1 Gestational diabetes as a model of fetal programming

Understanding gestational diabetes as a model of fetal programming has an increasing importance in the context of the rising prevalence of the disorder in global populations. The neonatal complications of diabetes in pregnancy (including pre-existing type 1 and 2 diabetes), such as macrosomia, have been recognised since Pedersen's original 1954 study of women with severe, untreated hyperglycaemia in pregnancy (200). Over the last decade, large clinical trials of gestational diabetes, have raised awareness of the adverse outcomes associated with relatively mild degrees of maternal hyperglycaemia, resulting in a significant change in clinical practice in both diagnosis and treatment of the condition (201).

During these years, the role of diabetes in pregnancy in fetal programming of future chronic diseases has also been elucidated, but is isolated to a few specific clinical models. These models, such as the Pima Indians and Mysore Parthenon Study described in Chapter 1, are important at defining the risk of programming but may have unique characteristics that do not allow wider contextualisation of their findings. For example, the Mysore population has been shown to have a high prevalence of maternal vitamin B12 deficiency and is predictive of gestational diabetes, a finding that has not been replicated outside of India (202). The Pima Indian population has provided invaluable insights into fetal programming via gestational diabetes and obesity, but recent studies suggest that they have unique genetic susceptibility underlying these diseases that may differs to other populations, thereby limiting how easily they can be compared (203). Evidence from animal models does support the notion that gestational diabetes can induce fetal programming, but these models are sparse. The most convincing evidence for the ability of maternal hyperglycaemia to induce a programmed offspring phenotype of hyperglycaemia and hyperinsulinaemia comes from Gauguier et al. (36). The mouse model presented in Chapter 6, confirms the findings of Yamashita et al. that maternal gestational diabetes programmes offspring towards diabetes (37). Together with the innovative studies identifying the increased propensity towards type 2 diabetes in the offspring of mothers with type 1 diabetes, these models provide convincing evidence of the longstanding metabolic changes induced in offspring exposed to maternal hyperglycaemia, independent of the potential confounding genetic risks associated with human gestational diabetes. The Gauguier model also suggests that the key environmental influence on programming in these models is hyperglycaemia rather than the other pathophysiological features associated with gestational diabetes. In contrast, Lindsay et al. show that maternal

hyperleptinaemia in association with obesity may also have programmable effects on offspring (39); and important consideration as maternal diabetes and obesity commonly co-exist.

With these limitations of the existing evidence base, the aim of this study is, first, to characterise maternal environmental factors that may vary in a pregnancy population and confound or associate with gestational diabetes exposure. In particular, this study seeks to characterise the micronutrient status of women with and without gestational diabetes to identify prevalent deficiencies and evaluate their associations. The second aim of this study is to investigate whether there are any epigenetic differences detectable in cord blood and placenta in offspring born to mothers with and without gestational diabetes, and whether these may be signatures associated with fetal programming.

7.1.2 Gestational diabetes and micronutrient deficiency

There is a body of evidence that suggests certain specific micronutrient deficiencies are more common in women with gestational diabetes. Several studies, including a systematic review and meta-analysis of 2146 participants (of whom 433 had gestational diabetes) and a post-hoc analysis of HAPO participants, report an association of vitamin D deficiency (as measured by 25(OH)D) and gestational diabetes (e,g (204)). However, conflicting data exists and the studies reporting an association do not adequately control for the multiple confounders that could underlie their finding, e.g. ethnicity, socioeconomic status, health behaviours and uptake of nutritional supplements recommended in pregnancy. For both, evidence of a true association needs to be obtained from larger cross-sectional studies with adequate control for confounders, and to identify causation, an RCT looking at whether vitamin D supplementation reduces the risk of gestational diabetes should be performed.

As highlighted above, there is a reported epidemiological association between the functioning of the one-carbon cycle (see figure 7a) and gestational diabetes in the Mysore Parthenon study. This study identified an association of B12 deficiency with gestational diabetes incidence (8.7 vs. 4.6% of women had GDM in those with and without B12 deficiency, OR = 2.1, p = 0.02) although the association seemed partly mediated through maternal BMI (202). This study, performed in south India where a lacto-vegetarian diet is common, may not relate to other global populations. However, a range of studies find an association of maternal homocysteine levels with gestational diabetes and glucose intolerance in pregnancy (205), and this appears to have a variable relationships with B12 and folate, although the studies performed have not made adequate control for trimester-specific changes in their metabolism nor the need for specific assays in pregnancy that are unaffected by physiological changes or

use of nutritional supplements. None of these studies of micronutrient deficiencies and gestational diabetes explain why these associations exist, and further work to understand whether there is a biochemical or metabolic cause for this, or whether they are due to a common, e.g. socio-economic or behavioural, confounder.

The studies described above use standard clinical assays of serum cobalamin (B12) and folate that are insufficient to examine the functioning of the one-carbon cycle effectively in pregnancy (reviewed in (206)). The standard assay of cobalamin measures total circulating serum B12 and, in pregnancy, levels fall after the first trimester of pregnancy due to haemodilution making it difficult to interpret these accurately in the diagnosis of deficiency. Holo-transcobalamin is emerging as a useful marker of B12 status, and represents the fraction of cobalamin that is bound to the transcobalamin II molecule and is ready for tissue uptake. This test is not routinely available in clinical laboratories but is thought to offer a useful biomarker of B12 status in pregnancy. Methylmalonic acid (MMA) is an intracellular metabolite that is converted to succinyl-CoA for use in the Krebs cycle and requires B12 as a cofactor for its conversion. Multiple studies now show that MMA is a sensitive functional marker of cobalamin deficiency at a tissue level and this is now increasingly used to diagnose true B12 deficiency in routine clinical practice, but some uncertainty lies over whether MMA can diagnose deficiency in pregnancy as its concentration rises in the third trimester of pregnancy. These concerns are summarised by Murphy et al. (207) and addressed in their study that characterises B12 status from pre-pregnancy, through gestation, and to fetal (cord blood) levels, using cobalamin, MMA and holoTC. They identified that preconception MMA was significantly associated with MMA in pregnancy (at 8, 20 and 32 weeks) and in cord blood, and also had a significant inverse relationship with maternal B12 and holoTC during pregnancy and with cord blood holoTC. Using regression modelling, the authors also found an association between low pre-pregnancy and a rising maternal MMA at the end of gestation and suggest that MMA is a useful biomarker of B12 status in pregnancy. Unfortunately, this study did not include assays of folate status, or other intermediates of the one-carbon cycle such as homocysteine, both of which may be useful to determine the relevant influences on fetal programming.

Folate assays also have to be chosen carefully in pregnancy as the routine serum folate assay is affected by recent ingestion of folate, and therefore true deficiency may be masked in women who are taking folic acid supplements (as is routine in early pregnancy) and have taken a dose prior to blood sampling. The red cell folate assay is the optimal measure of folate status as it offers a marker of availability over the 90-120 day lifecycle of an erythrocyte.

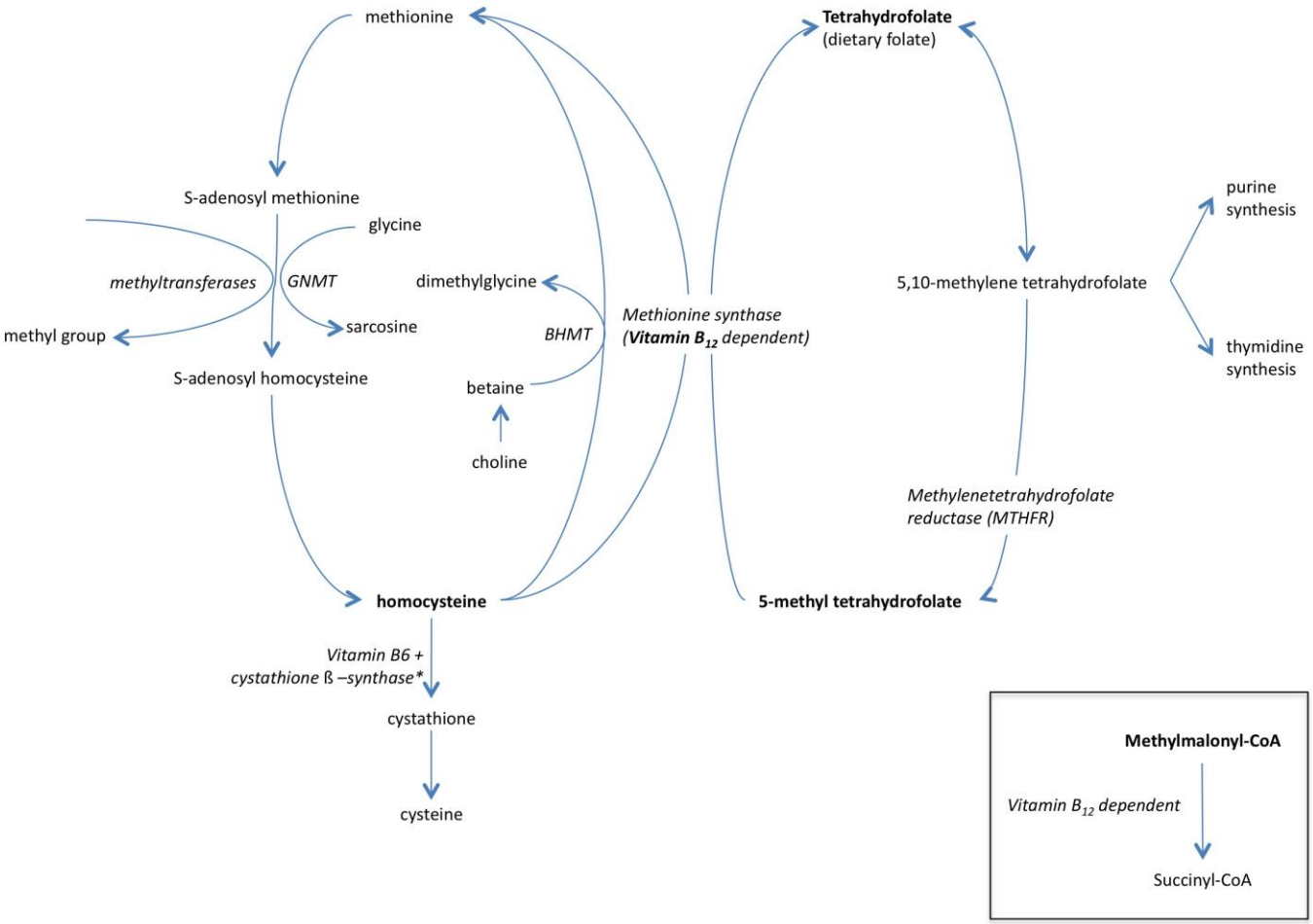


Figure 7a. The one-carbon cycle. Metabolites assayed in these studies are highlighted in bold. GNMT = glycine N-methyltransferase; BHMT = betaine homocysteine S-methyltransferase

7.1.3 Generation of pilot epigenetic data

Complementary to the aim of this study to characterise maternal micronutrient status in gestational diabetes is the aim to generate an initial insight into the epigenetic mechanisms underlying programming via human gestational diabetes. This study, whilst small, is hoped to elucidate how specific variations in the maternal environment, with respect to diabetes status and micronutrient deficiency, may play a role in the developmental origins of disease, and also to act as a test of feasibility for future hypothesis-driven studies. It was decided that gestational diabetes would be used as a model, rather than type 2 or type 1 diabetes, partly as a pragmatic decision (there are greater numbers of women with GDM pregnancy) but also so that there was a direct correlate with the mouse model in order to maximise their combined potential.

It was hoped that epigenetic modifications associated with GDM exposure may represent early markers of programming that could be followed up with longitudinal sampling and in larger cohorts. The study was also designed to obtain some relevant tissue samples, i.e. placenta, and umbilical cord, to fetal growth and nutrient delivery, in addition to fetal cord blood.

To date, there is no published evidence of epigenetic changes associated with this model of programming, and the putative link between maternal hyperglycaemia and DNA methylation is less clear than other models, e.g. programming mediated by one-carbon metabolism. However, there is a wealth of clinical and epidemiological data in support of this model of programming, as summarised above and in the introduction. In addition, the work of El-Osta et al, summarised in Chapter 1, suggests that glucose can have lasting effects on gene expression via epigenetic modifications at genes (e.g. NFkappaB) that regulate transcriptional activity and a possible role in metabolic complications. Data showing the epigenetic regulation of nutrient transfer and glucose metabolism genes in placenta according to glucose exposure has also been discussed in Chapter 1 and lends further support to this study in its attempt to elucidate whether there are epigenetic signatures of maternal glycaemia in the cord blood and placenta of offspring.

7.1.4 Specific study aims

The role of gestational diabetes and possible interacting nutritional influences on the neonatal epigenome will be studied with the following specific aims in mind:

1. Identification of differential methylation in the offspring of mothers of South Asian origin with and without gestational diabetes
2. Detailed characterisation of the maternal pregnancy environment with respect to metabolic and micronutrient status.
3. Collection of cord blood and placenta to identify tissue-specific DNA methylation in the offspring studied.
4. Development of standardised sample collection protocols for future epigenetic studies in larger clinical cohorts powered to detect associations with longitudinal offspring phenotype.
5. A human correlate of the mouse model of gestational diabetes in Chapter 6.
6. Combined analysis with related studies, including growth-restricted pregnancy, and the models of fetal programming described in chapters 4 and 5.

7.2 Methods

7.2.1 Local population – background demographics

The study was performed in the East London borough of Tower Hamlets. Data from the 2001 Census (available through towerhamlets.gov.uk) describes that this borough has a young population, with 59 percent of its population falling between the ages of 15 and 44 years of age, compared to the national average of 42%. Tower Hamlets is ethnically diverse, and its second largest ethnic group (after White British) are Bangladeshi, comprising 30% of the total population. In 2008/9 there were 68 births per 1000 women aged 15-44 in Tower Hamlets, compared to 62 per 1000 in England. During this same time period, of the 4255 births within the borough, 45% were to women of Bangladeshi origin.

Gestational diabetes is estimated to affect 5% of pregnancies in the UK (NICE guidance). However, there is a lack of published data on GDM prevalence, in part due to the variation in screening practices and diagnostic cut-offs. The current clinical guidelines at the Royal London Hospital use a hybrid of NICE and IADPSG diagnostic criteria for the diagnosis of gestational diabetes (see Chapter 1). In an unselected audit of pregnancies (n=433) at the Royal London Hospital in 2010, 9% (n=39) of women had pregnancies affected by gestational diabetes, diagnosed by their own criteria. Of these 39 women with gestational diabetes, 92% were of Asian (Bangladeshi) origin. By comparison, estimated prevalence of GDM in a rural Bangladeshi population, using the less stringent WHO criteria, was 8.2% (208).

7.2.2 Recruitment

Pregnant women were recruited from the Royal London Hospital antenatal clinics, and where possible, male partners were also recruited. Women were identified and invited to join the study after routine screening for gestational diabetes mellitus (GDM) was performed at 28 weeks' gestation. The screening test used at this time, the 75g 2-hour oral glucose tolerance test (OGTT) is offered to all women of Asian origin at the Royal London Hospital due to their high risk of GDM and the glucose cutoffs are ≥5.8 and 7.8mmol/l fasting and 2 hours post-challenge. Women with and without a diagnosis of GDM made at this time were selected to join the study in equal number. Alison Fiddler, Senior Midwife performed most recruitment to the study, under supervision from Sarah Finer, involving verbal discussion of the study and the use of written information sheets and leaflets given to all potential participants. The study

was performed under standard Good Clinical Practice guidelines, with REC approval. The inclusion and exclusion criteria for the study were as follows:

Inclusion criteria:

Women with and without GDM

Women of Asian origin

Exclusion criteria:

Multiple pregnancy

Pre-existing type 2 diabetes (diagnosed pre-pregnancy, or on the basis of an abnormal 16 week OGTT)

Type 1 diabetes

Women and their partners who gave verbal consent to join the study were invited to one study visit, run by Sarah Finer, with assistance from Francess Fode and Margaret Mbah-Ebako, Research Nurses from the North-East London Diabetes Research Network. The study visit comprised the following:

1. Additional explanation of the study and requirements of participants
2. Obtaining formal written consent from each adult. Pregnant women were also asked to give consent on behalf of their offspring to join the study as well as Human Tissue Act consent for the use of stored tissue samples.
3. Medical, social and dietary history taking.
4. Brief clinical examination
5. Venepuncture

7.2.3 Clinical data collection

7.2.3.1 Parental study visits

Maternal study visits were performed in the 3rd trimester of pregnancy, from 28 weeks' gestation onwards. The timing of paternal study visits was not fixed and was generally performed at the same time as the maternal study visit, or to coincide with other routine antenatal care appointments.

All participants were asked to attend study visits in the fasting state, and appointments were all performed during the hours of 8am and 11am each morning.

Structured questionnaires were used to take medical, social and dietary histories and were based on standardised questionnaires validated in the Bangladip study (MRC funded study of diabetes prevention in Bangladeshis, PI: Graham Hitman). Brief clinical examinations of participants were performed and included measurement of height, weight, abdominal circumference (men only) and blood pressure.

Additional clinical data collection was performed post-partum and prior to her discharge from hospital. At this time, details of the pregnancy (from study visit to delivery) were recorded from the hospital notes, such as treatment of gestational diabetes, administration of steroids, onset of pre-eclampsia and use of any other medications during pregnancy. Routine clinical data on the labour, mode of delivery and postnatal complications were also recorded from the hospital notes. Approximate total cumulative doses of metformin and insulin during pregnancy were calculated from recommended prescriptions recorded in hospital notes.

7.2.3.2 Offspring

Prior to discharge from hospital, brief neonatal anthropometry was measured on all babies, including measurements of head and abdominal circumference, length, birth weight and ponderal index. A customised birth weight centile calculator, available at www.gestation.net, was used to generate birth weight centiles, adjusted for maternal height and weight, ethnic origin, parity and sex of the offspring. This algorithm is based on data collected from 96,830 births from a database in the West Midlands and used in routine clinical use in some UK hospitals and in some research studies (209).

Follow-up of offspring beyond the neonatal period was not performed, however mothers were asked to give consent to be re-contacted in the future should this become necessary.

All biological samples obtained were stored using anonymised codes with no direct link to personal information, other than in a password-protected NHS computer database.

7.2.4 Nutritional, hormonal and biochemical assays

Maternal blood samples taken at the study visit were processed and analysed in 3 clinical laboratories in London, all with Clinical Pathology Accreditation. Laboratories were chosen due to convenience, expertise in specific assays and cost. The tests performed on study visit samples are summarised in table 7a.

Blood test	Sample type	Method	Normal range	Units	Lab
Serum glucose (as part of oral glucose tolerance test)	Fluoride oxalate	Hexokinase assay (Roche)	Dependent on context	mmol/l	Barts Health
Insulin (fasting)	Serum	Electrochemilluminescent sandwich immunoassay (Roche)	2.6 - 24.9	mIU/L	Barts Health
Fructosamine (corrected for serum protein)	Serum	Nitroblue tetrazolium (Roche)	<300	umol/L	Barts Health
Serum B12	Serum	Electrochemilluminescent competitive immunoassay (Roche)	191-900	ng/L	Barts Health
Holotranscobalamin	Serum	Immunoassay	<25 - Result suggests vitamin B_{12} deficiency	pmol/L	St Thomas' Hospital
			25-50 - Vitamin B_{12} deficiency cannot be excluded. Measure MMA		
			51-165 - Vitamin B_{12} replete		
			>165 - Seen in early cobalamin treatment		
Methylmalonic acid	Serum	Gas chromatography-mass spectrometry	≤280 (<65 yrs)	nmol/L	St Thomas' Hospital
			≤360 (>65 yrs)		
Serum folate	Serum	Electrochemilluminescent competitive immunoassay (Roche)	3.8-20.0	ug/L	Barts Health
Red cell folate	EDTA whole blood	Competitive binding paramagnetic particle assay (Beckman Coulter)	160-640	ug/L	Barts Health
5-methyltetrahydrofolate	Plasma	HPLC (Variant II)	7.6-42.0	nmol/L	St Thomas' Hospital
Homocysteine	Plasma	HPLC (Variant II)	<10 (in pregnancy)	umol/L	Homerton
Vitamin D (25(OH)D)	Serum	LC-MS/MS (Acquity Quattro Premier XE)	<30 deficiency	nmol/L	Barts Health
			30 – 79 insufficiency		
Vitamin A	Serum	HPLC (Variant II)	1.05 - 2.45	umol/L	Barts Health
Vitamin E	Serum	HPLC (Variant II)	11.6 - 46.4	umol/L	Barts Health

Table 7a. Summary of blood tests performed on maternal samples.

7.2.5 Sample collection for molecular studies

7.2.5.1 Parental samples

Whole blood was collected for DNA in a 5ml EDTA tubes during venepuncture at study visits. Samples were divided into 1ml aliquots and frozen at -80^{o}C within 2 hours of collection.

7.2.5.2 Offspring samples (at delivery)

At delivery, fetal samples were taken from the umbilical cord and placenta. 15-20ml of cord blood was obtained from the umbilical cord arteries and vein using a needle and syringe as soon as possible after delivery, using a standard operating protocol. EDTA and Tempus tubes were filled with cord blood for DNA and RNA samples, respectively. EDTA samples were divided into 1ml aliquots and frozen -80^{o}C, along with the Tempus tube, within 12 hours of collection.

Placenta samples (approximately 2 x 2cm) were taken adjacent to the insertion of the umbilical cord on the fetal surface of the placenta. Umbilical cord samples (also 2 x 2 cm) were taken from the end of the cord near to its placental insertion. Tissue samples were rinsed in sterile PBS and then placed in 5ml tubes containing RNA-later solution and stored at 4^{o}C. The time between delivery and sampling was documented for each collection. SF took the majority of samples, and the midwifery and obstetric staff of the Labour Ward took the rest. After 24-48 hours, samples were dissected into 4 pieces and divided into two cryotubes before being placed at -80^{o}C for long-term storage.

There was careful documentation of the time taken between placenta delivery and sample collection to ensure that external factors, such as tissue hypoxia or blood coagulation, can be considered in sample QC and data interpretation.

7.3 Results

7.3.1 Samples collected

Recruitment to the study yielded 80 maternal samples, of which 41 were paired with paternal samples (see figure 7b). Samples from 65 of the 80 women recruited were obtained at delivery, with 15 lost due to the women delivering elsewhere or the research team being unaware of their labour. 3 offspring samples were excluded from analysis due to significant fetal problems, (prolonged neonatal hypoxia during delivery and fetal abnormality due to a chromosomal abnormality).

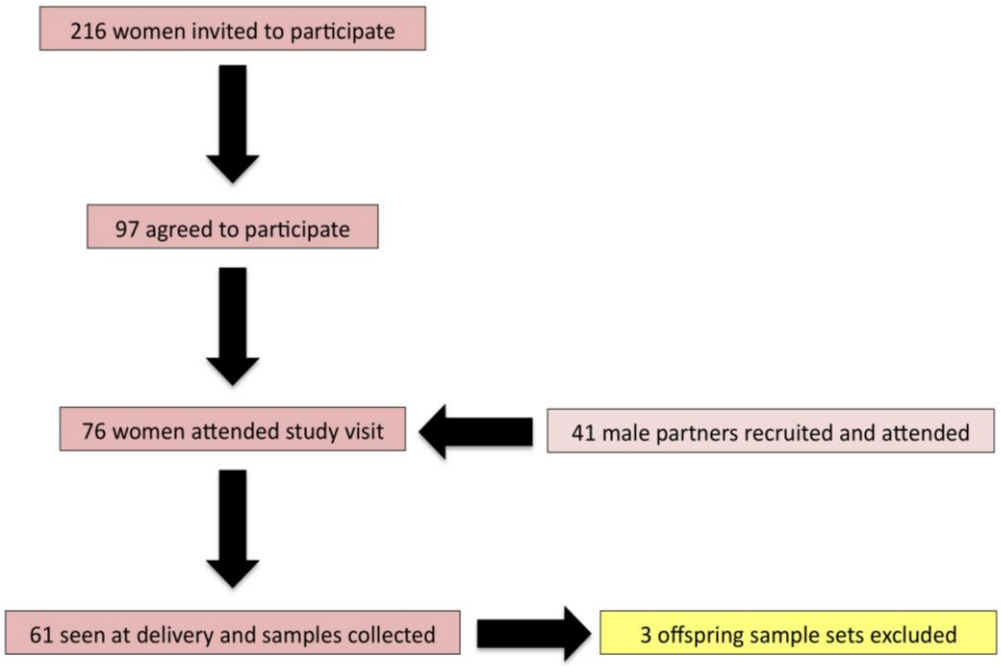


Figure 7b. Recruitment and sample numbers collected in the human GDM study

7.3.2 Maternal data

Maternal phenotype data and biochemistry on metabolic and micronutrient parameters are presented in table 7b. There was a statistically significant difference in the age of women with and without gestational diabetes, which is expected due age being a risk factor for the condition. Body mass index, parity and blood pressure parameters might have been expected to be different between the groups but were not, although the wide standard deviations

suggests that this is due to small sample size. The mean BMI across both groups was in the 'overweight' range, suggesting a population that is at risk of obesity and diabetes. As would be expected, women with gestational diabetes had, on average, higher fasting and 2 hour post-glucose challenge glucose measurements at diagnosis of gestational diabetes. In addition, serum fructosamine was significantly different, reflecting an increase in a glycosylated protein that is measurable in serum and reflects blood glucose concentrations over the previous 2-4 weeks. HOMA-IR scores, based on fasting glucose and insulin computed by a standard algorithm (www.dtu.ox.ac.uk/homacalculator/index.php) were not significantly different between experimental groups, but showed a wide standard deviation across groups and therefore a larger sample number would have been required to detect a significant difference. It is important to recognise glucose tolerance in the individuals studied as being a spectrum, and given that the diagnosis of GDM can be made on either a fasting or 2 hour OGTT glucose level, there is considerable overlap in glucose levels in the GDM and non-GDM groups (see figure 7c). Women with gestational diabetes were, on average, recruited one week later in gestation than control women. This difference therefore affects the timing of blood sampling, but did not have a significant effect on the biochemical or metabolic test results presented, as determined by regression analysis (data not shown).

There was a significant difference in serum holo-transcobalamin (or 'active B12') between women with and without GDM (median 50 pmol/L vs. 57.4 pmol/L, $p<0.05$), but in no other of the micronutrients measured. These effects were independent of metformin treatment, a drug used to treat type 2 and gestational diabetes known to induce B12 deficiency (via unclear mechanisms) (210). Serum B12 and methylmalonic acid showed differences that support the suspicion of B12 deficiency from the finding of lower holo-transcobalamin in GDM women, but these were not different at a statistically significant level themselves.

	Controls (n=36)		GDM (n=39)		t-test
	Mean (SD)		Mean (SD)		
Gestation at study visit (weeks)	32 (208)		33 (3.0)		*
Age (years)	28 (5.2)		31 (4.5)		**
Body mass index (m/kg^2)	27 (6.4)		27 (4.6)		ns
Parity (n)	1.5 (1.5)		1.8 (1.5)		ns
Systolic blood pressure at booking (mmHg)	111 (12.6)		110 (10.3)		ns
Diastolic blood pressure at booking (mmHg)	70 (11.3)		72 (9.9)		ns
28 week GTT 0 mins (mmol/L)	5.0 (0.5)		5.3 (1.1)		***
28 week GTT 120 mins (mmol/L)	5.8 (1.1)		9.6 (1.7)		***
Fasting insulin	16.8 (11.1)		16.6 (12.0)		ns
HOMA-IR	0.31 (0.21)		0.31 (0.23)		ns
Corrected fructosamine	282 (18.0)		294 (28.2)		*
	Mean (SD)	**Median (SEM)**	**Mean (SD)**	**Median (SEM)**	
Serum B12 (ng/L)	258 (84.4)	255 (14.3)	280 (151.6)	230 (24.0)	ns
Holotranscobalamin (pmol/L)	50 (15.6)	50 (2.6)	73.3 (52.7)	57.4 (8.5)	*
Methylmalonic acid (nmol/L)	260 (227.3)	168 (37.9)	248 (200.0)	222 (32)	ns
Serum folate (ug/L)	9.5 (5.1)	7.5 (0.9)	9.9 (5.4)	8.4 (0.9)	ns
Red cell folate (ug/L)	350 (209.6)	309 (36.5)	358 (137)	342 (22.0)	ns
5-methyltetrahydrofolate (nmol/L)	25.9 (26.3)	15.9 (4.4)	24.3 (19.8)	16.9 (3.2)	ns
Homocysteine (umol/L)	5.7 (2.2)	4.9 (0.4)	5.9 (2.0)	5.6 (0.3)	ns

25(OH)D (nmol/L	38.8 (25.4)	32 (4.2)	45.6 (31.8)	35.5 (5.0)	ns
	Mean (SD)		**Mean (SD)**		
Vitamin A (umol/L)	0.94 (0.26)		0.89 (0.3)		ns
Vitamin E (umol/L)	28.6 (9.3)		28.1 (7.7)		ns

Table 7b. Maternal metabolic and micronutrient parameters. Mean and standard deviation are presented and where means are significantly different *p<0.05, **p<0.005, ***p<0.0005. Where mean and median are presented, data is non-normally distributed and t-tests are performed on log-transformed data.

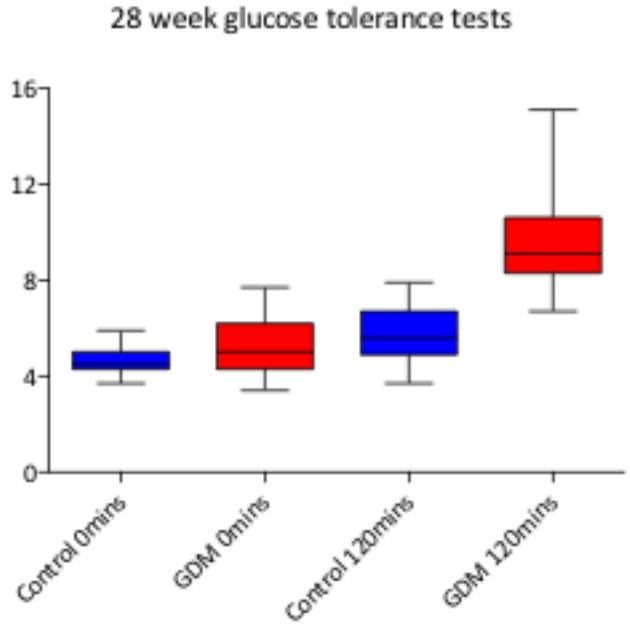


Figure 7c. Serum glucose measurements in control (non-GDM) and GDM women at their diagnostic glucose tolerance test.

Table 7c shows the prevalence of micronutrient deficiencies across both groups of women studied. Overall, there were high rates of vitamin D deficiency and insufficiency across both groups, with almost half of all women having a 25(OH)D level of less than 30nmol/L. The assays of vitamin B12 and its functional markers highlight the difficulty in diagnosing deficiency in pregnant women. Serum B12 (cobalamin) suggests that a quarter of women may be deficient; whereas holo-transcobalamin measures indicate that up to 41% of women could be deficient. Methylmalonic acid measurements suggest that this number is lower, but significant at 32%. Serum and red cell folate levels are consistent and suggest a low prevalence of deficiency, although this is a surprising proportion nevertheless, given that all women are supposed to take folic acid supplementation in early pregnancy. The measurement of 5-methyltetrahydrofolate reveals deficiency in 15% of women in this downstream metabolite of dietary tetrahydrofolate formed after catalysis using the methyltetrahydrofolate enzyme. These biochemical measures suggest that the excess deficiency of 5-methyltetrahydrofolate, compared to tetrahydrofolate, could be due to the presence of the MTHFR polymorphism in

those individuals. This is supported by the minor allele frequency of risk polymorphisms in the MTHFR gene (e.g. rs1801133) that are associated with hyperhomocysteinaemia and found in 33% of individuals of European origin. This high frequency of this allele has not been replicated in population-wide studies of South Asians, but is corroborated by some (211) but not all (212) smaller studies in this ethnic group.

Micronutrient	**Proportion of deficient mothers (n=75)**
Serum B12	24%
Holo-transcobalamin	True deficiency 4%
	'Deficiency not excluded' 41%
Methylmalonic acid (MMA)	32%
Serum folate	5%
Red cell folate	6%
5-methyltetrahydrofolate (5MTHF)	15%
Homocysteine	5%
25(OH)D	Deficient 49%
	Insufficient 40%

Table 7c. Prevalence of micronutrient deficiencies in all women

Homocysteine, can be thought of as a marker of the overall functioning of the one-carbon cycle; it is known to vary with B12 and folate status, as well as the presence or absence of MTHFR polymorphisms. Multiple linear regression modelling was performed to identify the predictors of homocysteine, and thus characterise what factors, in these pregnant women, were disruptors of the one-carbon cycle. Regression modelling, showed that only holo-transcobalamin concentrations were a significant (inverse) predictor of serum homocysteine (beta = -0.225, p = 0.012) (table 7d). Other markers of B12 status and/or folate status did not contribute to homocysteine levels and nor did maternal age, gestation, BMI or parity.

Independent variable: (log)homocysteine				
Covariates	**Unstandardised Coefficients**		**t**	**Sig.**
	B	**Std. Error**		
(Constant)	1.281	.434	2.953	.005
Red cell folate (log)	-.131	.145	-.904	.370
Holotranscobalamin (log)	**-.225**	**.087**	**-2.586**	**.012**
Methylmalonic acid (log)	.016	.052	.299	.766
5MTHF (log)	-.205	.228	-.898	.373
HOMA-IR	-.043	.084	-.506	.615
Serum B12 (log)	.055	.116	.474	.637
Serum folate (log)	.112	.317	.354	.725
Gestational age at delivery	.004	.006	.686	.495
Maternal age	.004	.003	1.043	.302
Maternal BMI	-.002	.003	-.605	.548
Parity	.002	.014	.154	.878
Model summary				
R	**R square**	**Adjusted R square**	**Std Error of the estimate**	**P value**
0.662	0.438	0.326	0.115	0.000

Table 7d. Multiple regression modelling to identify independent variables of (log)homocysteine. The 'log' prefix is used to denote data that is not normally distributed and has been log transformed prior to inclusion in the regression model.

Additional data was collected regarding treatment of gestational diabetes with metformin and insulin and is presented in table 7e and has been presented as cumulative dose exposures due to the variable timing of initiation and dosages of insulin and metformin treatment. No women in the control group were given metformin or insulin. The majority of participants (67%) with GDM were managed with diet-control alone.

	n	**%**
Diet control only	26	67
Mothers treated with metformin	10	26
Mothers treated with insulin	11	28
% of women receiving both metformin and insulin	8	21
Mean cumulative dose of metformin (g)	66 grams	
Mean cumulative dose of insulin (units)	592 units	

Table 7e. Summary of treatment methods and dosing GDM mothers (n=39).

7.3.3 Fetal data

Data collected at delivery is presented in table 7f. There were no significant differences in any of the phenotypic markers of birth weight/size between offspring born to GDM or control mothers and this was expected due to the small sample size. Although there is an apparent difference in customised birth weight centile between control and GDM offspring, this difference was not significant in parametric and non-parametric tests.

	Controls (n=34)		GDM (n=39)		Chi-squared test
	n (%)		n (%)		
Sex	21M, 13F		23M, 16F		ns
Macrosomia (>90th centile)	5		3		ns
Growth restriction (<10th centile)	9		7		ns
	Mean (SD)		**Mean (SD)**		**t-test**
Birth weight (grams)	3134 (473.9)		3156 (533.3)		ns
Gestational age at delivery (days)	274 (8.5)		271 (10.5)		ns
Ponderal index	28.9 (4.1)		29.7 (4.7)		ns
Cord blood haematocrit	0.45 (0.07)		0.46 (0.06)		ns
	Mean (SD)	**Median (SEM)**	**Mean (SD)**	**Median (SEM)**	
Customised birth weight centile	39.5 (32.1)	28.2 (5.5)	45.7 (30.3)	51.5 (4.9)	ns

Table 7f. Data collected at delivery of offspring

Regression modelling was performed to identify any predictors of birthweight, customised birth weight centile and ponderal index. There were no significant associations with customised birth weight centile (log transformed) after excluding the covariates already included in its own algorithm, nor were there any significant predictors of ponderal index. When birthweight was used as the independent variable, there was a significant association between the 0 minutes glucose level (log transformed) measured at the diagnostic glucose tolerance test and birthweight, as well as a positive association of parity and birth weight, both contributing to the overall model. Both of these factors are well known in multiple studies to have a positive association with birth weight. There was no association with any of the maternal nutritional parameters with birthweight and these have been excluded from the model in table 7g.

Independent variable: Birthweight				
Covariates	**Unstandardised Coefficients**		**t**	**Sig.**
	B	**Std. Error**		
(Constant)	417.744	1261.002	0.331	0.742
Corrected fructosamine	2.361	3.031	0.779	0.439
Gestational age at delivery	15.806	23.229	0.68	0.499
Log 0mins glucose (GTT)	**2480.563**	**1044.72**	**2.374**	**0.021**
Log 120mins glucose (GTT)	-86.313	565.987	-0.153	0.879
Maternal HOMA-R	-57.138	361.357	-0.158	0.875
Cumulative insulin dose	-0.076	0.22	-0.346	0.731
Cumulative metformin dose	-3.226	2.143	-1.506	0.138
Maternal age	-15.142	15.074	-1.005	0.32
Maternal BMI	11.857	14.845	0.799	0.428
Parity	**128.638**	**54.637**	**2.354**	**0.022**
Offspring sex	-98.414	123.676	-0.796	0.43
Model summary				
R	**R square**	**Adjusted R square**	**Std Error of the estimate**	**P value**
0.551	0.304	0.165	464	0.029

Table 7g. Multiple regression modelling to identify independent variables of birth weight. The 'log' prefix is used to denote data that is not normally distributed and has been log transformed prior to inclusion in the regression model.

7.3.4 Epigenomic data

After DNA extraction and bisulphite conversion, paired cord blood and placenta samples were hybridised to the Illumina 450k array in a single batch. Array data underwent standard QC, processing and normalisation. At the present time, this analysis is incomplete but preliminary data is presented. Matched cord blood and placenta samples show significant differences on a genome-wide scale, suggesting the presence of strong tissue-specific methylation patterns. Figure 7d is a multi-dimensional scaling plot shows genome-wide differences, at the 429,273 array probes that passed the QC, between cord blood and placenta pairs. The plot shows cord blood samples (in green) in a much tighter cluster than placenta samples (in orange). This is likely to reflect the greater heterogeneity of cell types in placenta samples than cord blood leading to varied methylation patterns

Figure 7d. Multi-dimensional scaling plot of normalised 450k data from cord blood (green) and placenta (orange) samples.

A subsequent MVP call has confirmed the presences of widespread methylation differences at genome-wide significance and withstanding FDR correction, at 338,062 probes ($p<0.05$) and 317,110 ($p<0.01$) and, at the present time, is being compared to a related study of cord blood and placenta sampling (described below).

7.4 Discussion

7.4.1 Nutritional data

This study has shown the wide variation in pregnancy phenotypes across a group of women of South Asian origin with and without gestational diabetes. The traditional risk factors for gestational diabetes, such as maternal BMI and parity were no more common in women with gestational diabetes recruited to this study, although on average they were 3 years older than controls. Micronutrient deficiencies were observed across control and GDM women, and of particular note, vitamin D deficiency was present in almost half of the women recruited. Other micronutrient deficiencies were identified and have highlighted a need for further study and characterisation. Folate deficiency in pregnancy is thought to be uncommon in the UK and peri-conceptual folate supplementation is advised to reduce the risk of neural tube defects across a population with mostly normal folate status. Data from East London has previously reported a 1% prevalence of folate deficiency of pregnant women (from European, African, Caribbean and Asian backgrounds), in contrast to the 6% identified in this study. This disparity may reflect a higher prevalence of folate deficiency in women of South Asian origin and requires further study with food frequency questionnaires to identify whether there is a clear dietary trigger to this deficiency, as well as understanding the adherence to folate supplements in early pregnancy. The observation that 15% women are deficient of the downstream folate metabolite, 5-methyltetrahydrofolate, suggests that the well-known genetic polymorphism of *MTHFR* may be common in the population studied and future studies should include genotyping of the known SNPs that are associated with folate deficiency and hyperhomocysteinaemia. The application of new assays of B12 metabolism, such as holo-transcobalamin and methylmalonic acid, has also highlighted possible deficiencies in as many as 41% of the population. The interpretation of these data in pregnancy is not well established and an important finding from this study is that holo-transcobalamin has an inverse relationship with homocysteine, suggesting the impact low B12 status on the overall functioning of the one-carbon cycle. Currently, holo-TC and MMA are not measured routinely in pregnancy and, other than an excess risk of neural tube defects, there is a lack of data suggesting what the direct risks of B12 deficiency to pregnancy outcomes might be. The identification of these one-carbon cycle defects in a 'healthy' UK population is a significant finding in relation to the findings of Yajnik et al showing their role in programming a childhood phenotype of insulin resistance and obesity. The prevalence of B12 deficiency in this London-based Bangladeshi population is somewhat surprising given that they do not follow lacto-vegetarian diets and are thought to have a regular consumption of fish. Analysis of food

frequency questionnaires will complement this data and provide an insight into whether B12 deficiency has a clear dietary cause or whether it may be due to reduced GI absorption or genetic factors affecting its metabolism. Future studies will also include characterisation of B12 status in women taking metformin in pregnancy to understand whether the excess risks of deficiency are increased with chronic metformin use, as has been shown in a non-pregnant population (213).

7.4.2 Metabolic data

Gestational diabetes represents the pathological glucose intolerance in pregnancy associated with excessive fetal growth and related maternal and fetal complications. A cross-section of pregnant women receiving antenatal care will exhibit a spectrum of glucose tolerance ranging from normal to that which is abnormal and diagnosed as gestational diabetes. The diagnostic criteria for gestational diabetes have raised debate due to uncertainty over the optimal glucose tolerance test cut-offs. Despite large population-based studies of the complications of gestational diabetes, there is still no consensus due to the linear association of maternal glucose status and adverse pregnancy outcome. This debate must be considered when designing studies of gestational diabetes, as the choice diagnostic criteria may affect the external validity of results. The linear association of glucose exposure and adverse fetal outcome (defined by cord blood C-peptide, macrosomia and neonatal hypoglycaemia) in the HAPO study was observed in a blinded group of women which provides strong proof of this association. However, using maternal glucose tolerance as a linear variable may not work in studies where women have received standard clinical care, including being managed as someone 'with' or 'without' gestational diabetes. Women who fall just within the diagnostic cut-offs for gestational diabetes and receive intensive clinic management based around reducing glycaemia may improve their glucose tolerance to a level that is better than a woman who was just outside the diagnostic criteria for GDM who did not receive this management. The higher corrected fructosamine levels in GDM women recruited to this study compared to controls (294 vs. 282 umol/L) does not support this theory, but this blood test will have only measured glycaemia around the first few weeks of gestational diabetes (as it was performed at recruitment to the study), and future studies should include a measure of maternal glycaemia, at the end of gestation or in cord blood to assess the cumulative glucose exposure throughout gestation to the fetus.

7.4.3 Potential interactions between gestational diabetes and one-carbon metabolism

As outlined in the introduction, it is important to understand the relationship between gestational diabetes and maternal nutrition in human fetal programming studies to elucidate their potential combined role. The data collected in this study suggests that women with gestational diabetes have lower levels of holo-transcobalamin, the 'active' vitamin B12 metabolite. Furthermore, the inverse relationship of holo-transcobalamin with homocysteine levels indicates that this may have a direct role on the overall functioning of the one-carbon cycle and, putatively, could disrupt the provision of methyl groups for DNA methylation.

There is little biochemical and molecular evidence to suggest a direct link between glucose (or diabetes) and the one-carbon cycle, apart from studies of the enzyme glycine N-methyltransferase (GNMT), reviewed by Luka et al. (214). GNMT is widely expressed in liver, pancreas and prostate and has a role in the functioning of the one-carbon cycle, in its transfer of a methyl group from S-adenosylmethionine to glycine to form sarcosine. In addition, folate is known to directly inhibit GNMT activity. The relevance of GNMT to diabetes is that it allows the utilisation of methionine for gluconeogenesis and it activated by glucagon. A study has shown that Zucker diabetic rats have increased GNMT activity, abundance and mRNA expression in liver and this is associated with differential methylation of certain genes with important epigenetic roles, such as *Dnmt1* (215). Whilst this evidence does not provide conclusive proof of a direct link between a diabetic state and DNA methylation via the one-carbon cycle, it does suggest a plausible mechanism where this interaction could lie. GNMT knockout mice have a phenotype of hepatic steatosis secondary to failure of gluconeogenesis, with hypoglycaemia and increased serum cholesterol (216). These observations suggest potential mechanisms whereby diabetic state, e.g. increased rate of gluconeogenesis, could affect epigenetic processes, or vice versa. It could be expected that increased gluconeogenesis could require an increased amount or activity of GNMT, and that this may increase the conversion of S-adenosylmethionine (SAM) to S-adenosylhomocysteine (SAH), increasing the SAM/SAH ratio and increasing amount of methyl groups for DNA methylation. However, it is important to note that these metabolic processes have only been studied in liver and could be tissue-specific. However, this hypothesis is supported by data published by Obeid et al. that assays multiple one-carbon cycle metabolites (including SAM and SAH) in individuals with and without type 2 diabetes (210). The researchers find a reduced serum SAM/SAH ratio in individuals with type 2 diabetes, compared to controls (8.2 vs. 9.1, $p = 0.006$, and $p = 0.023$ after adjustment for BMI) and described an association of this with a reduction in intracellular vitamin B12 concentration and increase in MMA. Future studies trying to elucidate the interaction between the one-carbon cycle and diabetes in pregnancy should

include further assessment of these metabolites. The advancement of metabolomic studies should allow assays of multiple metabolites to be performed with minimal sample quantity and high accuracy.

7.4.4 Epigenomic data

At the present time, the epigenomic data from this study is limited and has not yet focused on the identification of DNA methylation signatures associated with maternal GDM exposure. Sample collection and processing methods have been established as part of this study and the quality of samples in epigenomic studies has been high. To date, the analysis of DNA samples from cord blood and placenta have been bisulphite converted and hybridised to the Illumina 450k array with QC and normalisation of the data. Initial analysis has identified significant DNA methylation differences, on a genome-wide scale, between cord blood and placenta. These differences exist at the majority of probes on the 450k array and have strong statistical significance when examined using the analytical approach described in Chapter 2. It is hoped that understanding these tissue-specific differences will give a functional perspective on the epigenetic consequences of fetal exposure to maternal gestational diabetes. Regions of differential methylation between those exposed to GDM and those unexposed will be performed in cord blood primarily (chosen due to its greater homogeneity than placenta). It will then be possible to identify whether the same genomic positions that are differentially methylated in cord blood show variation in placenta, as well as performing independent analysis of placenta samples. Characterisation of any MVPs using pathway analysis will yield preliminary insights into the functional processes that may be affected by GDM exposure, and it may be possible to hypothesise how these relate to processes such as nutrient transfer in the placenta. Once MVPs have been identified in this human model, a comparison will be made with data from the mouse model presented in Chapter 6 to look for commonality that may represent stable epigenetic marks associated with this model of programming. This comparison will have to be between mouse livers and human cord blood/placenta, but additional replication can be performed in mouse placenta as this has been stored.

Analysis of methylation variation associated with GDM will also be performed using maternal glucose as a continuous variable. This may help identify whether there is a quantitative or threshold-based association of maternal glucose with the neonatal epigenome. As alluded to earlier, this approach will not be able to take into account the potential bias from GDM women being 'treated' and the possible improvement in their glucose parameters post-diagnosis through clinical management. Additional analysis will look for the effects of differences in one-

carbon status and DNA methylation. Whilst the numbers to do these additional analyses are small, it is hoped that it will yield sufficient insight to study further in new cohorts or existing cohorts in a more directed and hypothesis-driven manner.

Further analysis of the epigenomic data from this cohort will be performed in combination with a related study (PIs Sara Hillman, David Williams, UCL) that have collected clinical samples of women with normal and growth-restricted pregnancy. This study has been designed using the same sampling protocols as above so that direct comparisons may be made. Additionally, the cord blood and placenta samples that have been taken have undergone 450k array analysis in the same experimental batch. It is hoped that comparison of these datasets will be able to elucidate MVPs associated with ethnicity (the UCL samples come from a predominantly White European population), gestational age and growth restriction. Fetal programming studies of growth restriction and gestational diabetes exposure suggest a common end phenotype of cardiometabolic disease risk, therefore identification of common MVPs across both studies may elucidate epigenetic signatures that have stability and confer risk of that phenotype. It will not be possible to elucidate the association with phenotype in these studies, but this approach will guide hypothesis-driven studies of epigenotype-phenotype association using longitudinal sampling and detailed clinical phenotyping.

The study of epigenetic inheritance in model systems and humans has provided a tantalising possibility that epigenetic marks may be carried through successive generations (217). This theory suggests that a 'memory' of past environmental perturbations in pregnancy could be maintained, and may confer an adverse phenotypic outcome across these generations (218). This study has not been designed to study epigenetic inheritance, and indeed would require 3 successive generations to do so, but the availability of parental DNA from the recruits will allow follow-up of significant MVPs in parents. This may be of particular importance in SNP-associated methylation differences to understand whether these could be transmitted transgenerationally or whether they are a feature of the environmental exposure in pregnancy or stochastic processes.

7.4.5 Limitations of this study

There are several specific limitations of this study with regard to the proposed objectives set out in section 7.1.4. First, the identification of differential methylation in offspring born to mothers with and without gestational diabetes has not been completed and requires more analysis. The main limitations affecting all of the study objectives are (a) the small sample size, and (b) the recruitment of a group of women with gestational diabetes who had relatively

'mild' fasting hyperglycaemia. The sample size in this study was not designed to be sufficient to detect phenotypic outcomes associated with GDM exposure and therefore it will not be possible to identify epigenotype-phenotype associations. Furthermore, the phenotyping of offspring was limited and did not include anthropometric measures (e.g. skinfold thickness) that have been shown to be early markers of adiposity (40). Future studies should be designed with adequate sample sizes to detect phenotypic outcomes and longitudinal sampling from offspring would be a useful addition to uncover those epigenetic differences are primary effects of programming and which are secondary to the programmed phenotype. The sample size is also insufficient to examine genetic-epigenetic interactions in gestational diabetes, and it is conceivable that similar interactions occur as do in type 2 diabetes. The second main limitation of this study is that the women with GDM recruited to the study displayed 'mild' dysglycaemia, exemplified by the mean fasting glucose value at diagnosis of GDM (5.3 mmol/l) which is below the diagnostic thresholds for gestational diabetes. The majority of women met diagnostic criteria for GDM because of their 120 minutes (post-glucose challenge) glucose concentration, a finding which is not unexpected in women of Asian origin (24), but that is unusual compared to large studies such as HAPO. Further evidence for the 'mild' nature of the GDM in recruits comes from the small proportion of women who required metformin and/or insulin treatment. These findings could be due to selection bias, as individuals with severe dysglycaemia from gestational diabetes may require more medical attention and therefore be less likely to participate in a study, or because of the Hawthorne effect in which individuals participating in a study change their behaviour due to increased supervision or surveillance. The latter could have a significant effect on women with gestational diabetes with improvements in glucose tolerance via strict adherence to dietary control and physical activity recommendations. These biases could have significant effects on the expected outcomes, reducing the chance of finding significant differences between those with and without GDM with respect to micronutrient status, metabolic parameters and epigenetic outcomes. To best address this concern, a larger sample size and careful consideration of potential barriers to recruitment are important. In addition, the outcomes related to fetal programming may be better studied with additional recruitment of women with a pre-pregnancy diagnosis of type 2 diabetes or impaired glucose tolerance. Inclusion of these women would need careful consideration in terms of the expected outcomes, e.g. maternal hyperglycaemia would be a continuous fetal exposure rather than one limited to late gestation. However, this approach could maximise the potential to identify epigenetic differences associated with fetal programming as well as the possible collateral influences of micronutrient deficiency and gestational diabetes. Furthermore, recruitment of women with type 1 diabetes would expand

the possible outcomes related to maternal hyperglycaemia as a pure environmental influence, analogous to the approach of studies by Lindsay et al. (39).

Data collected on micronutrient status of the women in this study, particularly in relation to the one-carbon cycle, provides a new perspective on maternal B12 deficiency in an otherwise 'healthy' population of South Asian women in the UK. Whilst focus has been placed on folate and vitamin D supplementation in pregnancy for many years, there is little recognition of the possibility of B12 deficiency being prevalent. Further studies are required to evaluate this in larger numbers and to identify any association with adverse pregnancy outcomes or long-term programming risks. The data collected also suggests that current public health strategies around vitamin D supplementation in pregnancy are ineffective, given that the majority of women recruited to the study had either vitamin D deficiency or insufficiency. Other environmental factors that could play a role in programming or confound the maternal glucose-related effects include maternal lipids, obesity and macronutrient intake. These factors have not been fully evaluated in this study and should be included in future studies to elucidate their potential role. Future studies should also include people from different geographical areas and ethnicities to establish the role of other environmental and genetic factors that play a role in fetal programming.

Wider limitations that affect this study, such as sample size and power, will be discussed in Chapter 8 as it has relevance to the work presented in Chapters 3,4,5 and 6. Issues of tissue-specificity, timing of environmental exposures and the detection of genetic-epigenetic interactions will also be discussed in Chapter 8.

7.4.6 Next steps

The experimental plans set out in section 7.4.4 regarding the epigenomic data from this study will be used to guide further studies. It is hoped that this data, and the mouse model data in chapter 6, will yield sufficient insights to develop further hypothesis-driven research on the downstream functional consequences of such epigenetic variation and how this may predispose to disease onset in this programming model. This will guide the translational potential of these studies, and it is hoped that this will enable targeted intervention and therapeutic strategies to be developed. Future work will also focus on the possible 'double burden' of over and undernutrition (219), particularly in populations in transition where the prevalence of cardiometabolic disease is increasing. This approach could address important questions of whether gestational diabetes interacts with specific micronutrient deficiencies to confer risk, and if so, what mechanisms, e.g. placental nutrient transfer, are involved.

Chapter 8. Discussion

8.1 Main objectives

The main objectives of this thesis were as follows:

- To identify epigenetic factors involved in susceptibility to Type 2 diabetes as a route to understand disease aetiology and pathogenesis.
- To apply animal and human models to elucidate the role of epigenetic mechanisms in fetal programming of Type 2 diabetes and related cardiometabolic conditions.
- To use 'discovery-based' techniques to determine interactions between genetic, epigenetic and environmental influences, on an epigenomic scale.

8.2 Component studies

The main objectives have been investigated in a series of component studies, set out in chapters 3-7, and the main outcomes of these will now be discussed.

8.2.1 Type 2 diabetes study (chapter 3)

This study focused on the identification of epigenetic variation at specific regions of genetic susceptibility to Type 2 diabetes to characterise putative regions of genetic-epigenetic interactions. A Medip-chip approach was used to profile DNA methylation in nucleated blood cells on a genome-wide scale at regions determined by recent insights from type 2 diabetes and obesity genome-wide association studies. In this study, a genetic-epigenetic association was identified within the type 2 diabetes susceptibility haplotype of the *FTO* gene, characterised by the presence of a CpG-SNP at rs8050136 that created or abrogated a site for DNA methylation according to the phase of the SNP; this was conformed in DNA from adult hypothalamus and cerebral cortex. Our functional interpretation of this discovery suggests that this SNP-CpG site may have a dual role on *FTO* expression and in the functioning of an enhancer, *IRX3*, that has a role in development of the hypothalamus. The creation of a CpG site capable of methylation when the *FTO* risk allele is present may reduce the activity of this enhancer, suggesting a potential influence on brain systems involved in energy balance and

appetite regulation. The identification of this CpG-SNP has, in part, been replicated in a similar study performed by Toperoff (152), providing reassuring external validity of our findings.

In a wider context, the findings of this study have provided an important insight into how genetic and epigenetic variants may co-exist and provide a practical focus to determine which of the many genetic polymorphisms identified through GWAS have a functional role. This study also highlights the importance of integrating genetic and epigenetic techniques so as to avoid the assumption that an epigenetic variant is an independent, and perhaps stochastic event, when in fact it is driven purely by genetic variation. What this study lacks, most importantly, is an ability to identify allele-specific methylation and gene expression differences driven by this genetic-epigenetic variant. Determining allele-specific expression differences driven by genetic and methylation differences would require large and complex study, using epigenomic and transcriptomic techniques such as RNA-seq and BS-seq, and there is only a small body of published data taking this approach. One such study, performed by Chen et al. (220) has provided an important insight into the role of DNA methylation on transcriptional regulation in human embryonic stem cell development at *cis*-regulatory elements (e.g. transcription factor binding sites), and another study has identified the important influence genetic variation on DNA methylation patterns and subsequent gene expression (221). A different, and adequately powered, approach would be required to identify similar combined epigenomic and transcriptomic functions in studies of complex disease, such as type 2 diabetes.

8.2.2 Pune Maternal Nutrition Study (chapter 4)

In this study, epigenomic techniques were used to identify differentially methylated regions in the whole blood from offspring from this seminal fetal programming cohort in which maternal one-carbon cycle defects have been shown to induce a programmed phenotype of insulin resistance and obesity in offspring.

The aim of this study was to elucidate environmentally induced DNA methylation differences in the offspring epigenome with which to focus further studies of epiallelic variation and the relative influence on environmental and genetic factors. Medip-seq was used to identify DNA methylation differences in a super-selected group of offspring from the extremes of maternal exposure and associated programmed phenotype. Differentially methylated regions were identified between the control and case groups but this data has not yet been technically validated. The main limitations of this dataset are its small sample size and the difficulties of identifying the correct bioinformatic approach for to detect true methylation differences, itself

affected by sample size. More specifically, concerns over the bioinformatic analysis of Medip-seq datasets also include how to account for experimental biases (such as Medip enrichment) and false discovery. Since this project was started, the advent of newer whole genome and genome-wide technologies such as BS-seq and the 450k array have largely superseded Medip-seq experiments and offer a more robust platform with which to detect methylation without the experimental bias of an enrichment-based technique.

The Illumina 450k array was also used to identify differences between offspring groups in PMNS and has generated 2045 methylation variable positions (MVPs) between study groups, but none of these differences reach genome-wide significance. Unfortunately, the lack of overlaps between the 450k array and the Medip-seq DMRs meant that the former could not be used as technical validation of the latter. This dataset will require further work if conclusions are to be drawn from it about epigenomic differences in offspring exposed vs. unexposed to maternal vitamin B12 deficiency. Further experiments using the 450k array in additional offspring samples are planned, and a power calculation of the sample size required will be performed using data from the Matlab study.

Other possible limitations of this study exist and will be addressed in future studies. First, the selection of offspring enriched for the programmed phenotype as epigenetic differences secondary to the phenotype of insulin resistance and obesity may obscure primary epigenetic signatures of programming. This limitation is common to most cross-sectional studies designed to detect epigenetic variation with an aetiological role, and will be overcome in future studies by inclusion of offspring with and without the programmed phenotype as well as longitudinal sampling to study developmental and dynamic epigenetic differences before and after emerging phenotypic differences. The forthcoming intervention studies in PMNS, in which women will be randomly assigned to pre-pregnancy B12 supplementation or placebo, will also provide an invaluable insight into which epigenetic differences in offspring are related to the programming or other, perhaps stochastic, influences. Second, the use of whole blood DNA may be a poor surrogate for tissues with a more direct role in the pathogenesis of diabetes and obesity. Whole blood may a useful tissue for a preliminary discovery-based study such as this, especially as a periconceptual insult may mean that all 3 germ layers hold a molecular memory of it, it is not likely to inform downstream studies on gene function in relevant tissue. These kind of functional studies are important to develop a deeper understanding of the mechanisms by which the phenotype is induced and are likely to offer a greater potential for clinical translation and understanding the specific risks of adiposity and diabetes in this Indian population. To address these concerns, complementary studies are now taking place using an adipose tissue cell line (CHUB-S7) exposed to varying B12 and folate

environments (in collaboration with Gyan Tripathi and Sara Ponnusamy, University of Warwick). These studies have already shown that an environment deficient in vitamin B12 can promote adipogenesis via gene expression changes within the cholesterol biosynthesis pathway. Work is now underway to correlate these gene expression differences with genome-wide DNA methylation differences, using the 450k array. Preliminary data suggests that DNA methylation differences at 2 target genes (*SREBF1* and *LDLR*) exist and correlate to gene expression differences (Adaikalakoteswari A, Finer S et al., submitted to Circulation Research, 2012). Finally, another important consideration in all programming studies, such as PMNS, is to understand whether there are differences in the potential for programming according to the timing of the environmental insult in early development. The PMNS study is unique in its ability to define pre-pregnancy and early environmental factors, but is not designed to differentiate between these in the programmed outcome. Furthermore, the postnatal differences in childhood B12 and folate status were not studied until follow-up at 6 years, meaning that early life nutritional differences could have existed between the offspring study groups (and might be expected to be associated with the maternal nutritional differences) but are not accounted for.

8.2.3 Matlab Famine Study (chapter 5)

This study has generated data suggesting differences in DNA methylation in offspring directly exposed to famine in utero and early postnatal life, compared to those unexposed. DNA methylation differences, whilst not reaching genome-wide significance, were validated in a second sample set from the same Matlab study. Of particular interest is the finding that DNA methylation variants identified in a related model of fetal programming, in Gambia, seem to show commonality in this study. This finding supports the potential for there to be regions of the genome that are sensitive to environmental pertubation and that could play a central role in the development of a programmed cardiometabolic phenotype. The findings of this Matlab study will be used to generate sample size calculations in future studies as it provides an important insight into the numbers of samples required to detect statistically significant methylation differences using a genome-wide platform.

Gene ontology analysis of the MVPs identified in this study has produced findings that suggest that these epigenetic differences exert their effect via altered growth factor stimulation of cellular processes and signalling.

These findings can be considered to support an interaction between the early life environment and epigenetic profile of young adults. Interestingly, most programming models suggest that

the time of particular environmental susceptibility is in utero, whereas this supports the influence of postnatal life. The phenotypic outcome of the individuals in this study suggests that the strongest programmed influence is in those exposed in utero, rather than postnatally, but the small sample size and variation in the measurements taken may have obscured differences in this group. Alternatively, the programmed phenotype may not be fully evident in the postnatally exposed group, and if this is the case, it suggests that the epigenetic differences identified in these offspring may have a causal role, rather than be secondary to phenotypic outcome.

More detailed analysis of the phenotypic data from this study will be performed, and it is hoped that future studies will incorporate longitudinally collected samples and phenotypes from this cohort as well as studying the F2 offspring to look for possible transgenerational effects. It may also be possible to validate our data with that from other famine-associated programming cohorts, e.g. the Gambian, Dutch Winter Hunger or Chinese Great Leap Forward studies, and at the present time collaborations are being sought to do this.

8.2.4 Gestational diabetes studies (chapters 6 and 7)

The programming influence on the offspring epigenome from gestational diabetes exposure has been investigated using a mouse and human model. The mouse model, analogous to human gestational diabetes, was to generate tissue samples from offspring at the end of gestation and in adult life to study both tissue-specific and age-associated epigenomic differences. This model confirmed a programmed phenotype of glucose intolerance and adiposity in those mice exposed to GDM, and there were epigenomic differences in the fetal livers of those exposed vs. unexposed at the end of gestation. These findings have not yet been validated and are currently being followed up with gene expression studies in fetal liver and liver and adipose of aged offspring. A human study of gestational diabetes was set up and has generated samples for related epigenomic studies in cord blood and placenta from exposed. These samples have not yet been fully analysed, but the clinical data from this study has provided a useful insight into the complexity of studying human pregnancy due to the multiple different environmental 'exposures' that may occur in such a cohort. The women recruited to this study were predominantly vitamin D deficient and therefore it will be difficult to identify whether vitamin D status has an independent or contributory influence on any epigenomic differences in offspring. A focused study of one-carbon metabolism in the mothers recruited to this study has highlighted a prevalent deficiency of functional vitamin B12, using holo-transcobalamin and methylmalonic acid assays, with downstream effects on

homocysteine status and therefore possible influence on provision of methylation groups for epigenetic processes.

By combining an isogenic mouse model and human study, it is hoped that it will be possible to address important questions relating to tissue- sex- and age-specificity of epigenetic changes and their functional role, but also provide an important translational focus to guide future clinical studies and prevention strategies. Future experiments using this mouse model will enable a direct view of epigenetic-environment interactions, unaffected by genetic polymorphism, such a pure epialleles. In contrast, the human model, performed in women of Asian origin with a presumed high background genetic risk of diabetes, may yield insights into genetic-epigenetic-environmental interactions, but to assess these fully, much larger sample numbers will be required.

8.3 Discussion points relating to all studies

In addition to the summarised findings above, this thesis has generated points for discussion that relate to all of these studies and their ability to produce insights into type 2 diabetes aetiology, and these will now be discussed.

8.3.1 Methodology of epigenomic studies

The epigenomic studies performed first used Medip-seq as a means to assay and quantify DNA methylation across the genome (human and mouse). Two published studies have shown that Medip-seq, coupled with the Batman bioinformatic algorithm, is able to detect large methylation differences (>25%) between cancerous and non-cancerous human samples that technically validate using other experimental platforms including the Illumina 27k array and targeted BS-pyrosequencing (222). In this study, and others (223), the large methylation differences identified are thought to be driven by underlying genomic variants, e.g. cytogenetic differences, copy number variation and repeat sequences, associated with carcinogenesis itself. However, despite the large methylation differences detected in these studies, the DMRs identified did not withstand correction for multiple testing using the standard Benjamini and Hochberg FDR approach, although the statistical strength of methylation differences at individual DMRs identified were strong and analysis of the DMRs called showed individual

FDRs of <3.7%. The majority of DMRs identified in these studies were at repeat elements and areas of high CpG density, and this may be a true biological representation of where differential methylation occurs, but may also reflect the correction of the Batman algorithm for CpG density. The Medip-seq analyses presented in Chapters 3 and 6 used a different bioinformatic approach to call DMRs due to concerns that Batman was overcorrecting for CpG density and thereby biasing the DMR calling strategy. In these experiments, an approach developed by Thomas Down was used first to look for significant differences across the epigenome using simple t-test statistics in pair-wise and group-wise comparisons, and in the Pune experiment, pairs of samples were matched in such a way that possible enrichment biases would be reduced. This approach did not apply any correction for multiple testing and is prone to over-calling DMRs, and therefore an additional bioinformatic technique, USeq, was combined with this calling strategy. USeq incorporates the principles of a binomial distribution to the detection and comparison of enrichment peaks between groups of samples, and assumes a non-random distribution of sequencing reads across the genome. USeq may under call DMRs due to its stringent processing, and has not been validated in Medip-seq experiments, and therefore a combined approach of Thomas Down and USeq DMR calling strategies was used to generate a final 'top hit' list of DMRs in these experiments. To date, these top hits have not been individually validated and therefore it is not clear whether the Medip-seq data is technically sound and these are true DMRs or false discoveries. The FDR threshold in our USeq call was up to 20%, significantly higher than in the Saied and Feber studies, despite relatively similar sequencing depth in the former with our Pune dataset (the Saied study yielded 400 million reads for DMR calling from 12 cases vs. 4 controls; compared to 393 million reads in 8 cases vs. 8 controls in the Pune Study). The greater strength of the Saied and Feber DMRs that were called is likely to be due to the greater size of methylation difference at their DMRs due to their being driven by large genomic differences affecting methylation, in contrast to the smaller environmentally-determined differences expected in out programming studies. The importance of the sequencing depth and types of methylation difference expected for DMR detection is also highlighted between the mouse GDM study and Pune studies. The murine study, using inbred mice, identified a similar number of DMRs to the Pune study (outbred humans) with half the number of sequencing reads, suggesting that the CpG-SNPs identified in the Pune study may obscure the small, environmentally-induced methylation differences expected in these programming studies.

Technical validation of the Pune and Mouse GDM studies DMRs is underway, using different experimental platforms, and it is hoped that this may support the Medip-seq data, as is the

case in the Feber and Saied studies. Data using the Illumina 450k array suggests that where Medip-seq DMRs overlap 450k probes, there is a similar directionality of methylation difference. However, this analysis has not been performed across the Medip-seq dataset as a whole and the 450k array is not sufficiently epigenomic to validate the ability of Medip-seq to detect DMRs in the unbiased and discovery-based way that was originally intended in this study. For true validation of DMRs and the calling strategies that have been used, Medip-seq data should be compared to a dataset that is equivalently or more epigenomic, such as that derived from BS-seq. This type of validation analysis is also required to understand the sample size required in future studies to detect small methylation differences at genome-wide significance. It is likely that larger sample sizes and greater sequencing depth will be needed to overcome the issues of power, but they are unlikely to influence the possible bias from enrichment differences and non-random genomic coverage (i.e. bias towards highly methylated regions of the genome) in Medip-seq experiments. BS-seq is now a more accessible platform with which to study the epigenome, and it enables the direct quantification of DNA methylation at CpG sites across the epigenome, unaffected by antibody enrichment efficiency or CpG density. This now seems to offer a far better approach and will be used in future discovery-based epigenomic studies. Where a genome-wide approach is sufficient to test hypotheses, the Illumina 450k array offers a cheaper and reproducible means of generating DNA methylation data and can be readily integrated with genomic and expression profiling and used in so-called epigenome-wide association studies (EWAS) (224).

8.3.2 Other epigenetic signatures

In future studies, attention must also be paid to non-CpG associated DNA methylation, e.g. hydroxymethylation. This molecular event is thought to play an important role in developmental processes such as embryonic stem cell differentiation but cannot be detected in all epigenomic approaches (225). Techniques that use bisulphite conversion, e.g. 450k array and BS-seq, are unable to detect hydroxymethylation due to their reliance on genetic basis of CpG sites to assay methylation. The role of hydroxymethylation in fetal programming has not been determined, but would be an interesting molecular mechanism to study in this context due to its important developmental role.

Higher order epigenetic modifications, such as histone modifications, have not been studied in the experiments performed and would be a useful component of any follow-up studies and could provide an important means to integrate epigenomic data with studies of gene expression. Specific histone marks, and other post-translational modifications such as

transcription factor binding sites can be assayed on an epigenomic scale using chromatin immunoprecipitation with Next Generation Sequencing in a technique called ChIP-seq. Such experiments have provided important insights into the understanding of complex disease, for example in understanding the widespread genomic functions of vitamin D through its receptor binding (226), and its interaction with genomic variation and gene expression. Other regulatory features, e.g. Dnase hypersensitivity sites, can be assayed on a genomic scale using an enzyme-based approach coupled with sequencing, to identify their role in chromatin structure and its role in epigenomic regulation of transcription alongside specific histone marks (227). Recently, the ENCODE project has highlighted the importance of integrating studies of the chromatin landscape with DNA methylation maps to understand *cis*- and *trans*-regulatory elements if the functions of the genome, and especially its non-coding parts, are to be fully understood (228). One recent study by Sandovici et al. highlights the importance of studying higher order epigenetic marks to understand how fetal programming may disrupt transcriptional regulation of genes involved in diabetes pathogenesis (229). The authors identified that the promoter-enhancer interaction in Hnf4a was altered through diet-induced fetal programming and ageing, and that gene expression was associated with differences in active and repressive histone marks at the enhancer region, and DNA methylation differences were not identified at the same region. Beyond this study, the role of higher order epigenomic factors has not been well-studied in the context of fetal programming, but is likely to yield deeper insights into how DNA methylation changes exert their function and whether the dynamic chromatin landscape is susceptible to environmental influences and if so how by what processes. DMRs identified in the Pune study are seen to over-represent regions transcription factor binding sites and DNase hypersensitivity sites, suggesting an important and integrated role for these epigenomic processes in programming.

Studies of higher order marks are complex to perform as each histone mark requires a separate ChIP-seq protocol and it is therefore costly and bioinformatically complex to provide epigenomic data sets for integration. Furthermore, sample collection methods for chromatin immunoprecipitation studies are not easily applied to clinical cohorts, requiring greater amounts of DNA and immediate laboratory processing to cross-link chromatin and yield it stable and suitable for downstream experiments.

Another important factor in post-translational regulation of gene expression is the role of microRNAs in gene silencing. These nuclear DNA-encoded small RNA molecules have been shown to have a functional role in disease pathogenesis and their dysregulation may be induced through the maternal environment with functional consequences on adipose tissue and possible cardiometabolic disease pathogenesis (230).

8.3.3 Detection of environment-epigenetic interactions

The notion of pure epialleles, or epigenetic variants induced through environmental and/or stochastic factors, and their putative role in fetal programming, has been described in section 1.5.2. The human studies of fetal programming performed in this thesis have highlighted the difficulties of detecting epigenetic variants associated purely with the environment and uninfluenced by genetic variation. The Pune, Matlab and human GDM studies described in chapters 4, 5 and 7 incorporate studies of outbred individuals who can be assumed to vary genetically each other, especially with regard to SNPs. The DNA methylation studies performed in these experiments identify multiple regions of differential methylation located at SNPs where the creation or abrogation of a CpG site changes the ability to methylate, like the *FTO* SNP-CpG site described in Chapter 3. The variation in methylation at these regions, either 0, 50 or 100%, is large and therefore easily detected and may obscure the smaller differences in methylation induced through the environment. Future studies require greater power to detect smaller differences and this is best achieved through increasing sample size. Other options to identify pure environment-epigenome interactions would be to include assays of genetic variation, e.g. SNP genotyping and CNV arrays, so that these regions could be defined and analysed separately. Alternatively, this problem could be improved by reducing the amount of inter-individual genetic variation, e.g. by using sibling controls or twin studies, or before and after intervention studies of the same individual.

8.3.4 Detection of genetic-epigenetic interactions

The study of genetic-epigenetic interactions in type 2 diabetes described in Chapter 3 identified a CpG-SNP associated with the FTO risk allele for obesity with a putative functional role via enhancer activity. This study has shown that integrated detection of molecular variants may help direct outputs of association-based studies towards functional understanding and a similar approach is now used in many genomic studies of complex diseases. What future studies need to incorporate is the ability to detect allele-specific differences in DNA methylation and gene expression to understand the role of these genetic-epigenetic variants fully. Allele-specific methylation was studied by Li et al. (181) using high-density BS-sequencing across the genome of one Asian individual using DNA from peripheral blood mononuclear cells. This study identified multiple regions of allele-specific methylation associated with regions of imprinting and sequence variation, and that these were associated with allele-specific expression. Our Type 2 diabetes did not detect allele-specific methylation associated with the *FTO* CpG-SNP, i.e. it was a haplotype-driven effect rather than one with

wider allelic differences in methylation beyond the CpG-SNP. However, Shoemaker et al. have suggested that there is a concentration of highly correlated methylation patterns around some CpG-SNPs (and their LD blocks) in the genome, and that these tend to be represented by intermediate methylation (25-75%) (231). The authors of this paper suggest that CpG-SNPs are an important component of cis-regulation of the genome and epigenome, and beyond this study these findings highlight the importance of genetic control over the epigenome and the need for specific experimental approaches to define this fully.

Other genetic variants, such as repeat elements and transposons, have been shown to be associated with major sites of methylation variation (179). The identification of differentially methylation regions at repeat-rich regions has been confirmed in the Pune Study, in which 88% of DMRs were located at repeat-rich regions. Furthermore, the study of methylation differences in Gambian offspring exposed to seasonal nutritional differences in the Gambia showed preponderance at hypervariable regions, suggesting that these genetic-epigenetic interactions may have a role in fetal programming, although the mechanisms are not clear. These associations may have evolutionarily significant functions, such as the development of genomic imprinting or genome defence via transposable elements. It is therefore conceivable that the effects of methylation variation at these positions could be protective to subsequent generations by inducing parent-of-origin effects and/or minimising their effects via the accumulation of transposons, or could have the opposite effect by maladaptive influences on these evolutionary mechanisms. However, other data suggests that fetal programming induced through maternal nutrition may not exert influence via imprinting (99), and the bias of enrichment-based methylation assays towards regions of high CpG density may over-represent these genomic features in their outputs.

8.3.5 Biological relevance of epigenetic variants

The data presented in this thesis is significantly limited by the absence of gene expression studies to investigate the functional role of the differential methylation identified. Future studies in the mouse GDM study will be performed to look at genome-wide expression and integrate this with epigenomic data. This study lends itself well to this kind of investigation, as sample collection was tissue-specific and performed with methods suitable for high quality RNA extraction. Expression studies may also be performed in cord blood and placenta samples collected from the human GDM study but will be limited by the mixed cell populations in the samples collected. Future approaches will need to incorporate micro-dissection of tissues

and/or cell sorting of blood samples if detailed biological studies of these molecular differences are to be performed.

Other means by which these molecular studies may expand their biological relevance is to incorporate different study designs. Longitudinal sampling from cohorts is going to be an important means by which cause versus consequence are disentangled in studies, using the temporal stability of epigenetic variants and temporal association with phenotype to do so. Identification of quantitative traits and exposures and their effects on epigenetic variation may also yield useful functional insights and this approach will be taken with the human GDM study which will analyses methylation differences according to maternal glycaemia as a linear variable.

8.4 Wider discussion points

8.4.1 Does fetal programming underlie the missing heritability of type 2 diabetes?

Chapters 4-7 concentrate on the concept of fetal programming, a theory that is thought by many to provide a partial explanation of the missing heritability of complex diseases such as type 2 diabetes and obesity. There are many criticisms of this theory that need to be taken into account when considering the wider consequences of these studies. First and foremost that despite the wealth of cohort studies, there are no studies assessing the population attributable risk of fetal programming on complex disease, and therefore its global impact cannot be quantified. This may be due to the second, and more commonly voiced criticism of fetal programming studies, which is that despite high quality epidemiological, clinical and now molecular evidence of fetal programming, fetal programming is only supported by evidence from specific and extreme environmental contexts and have limited external validity. This second criticism, and related points, will be discussed in more detail as follows:

8.4.1.1 Environmental exposures

The environmental exposures studied in fetal programming studies are wide-ranging, and include nutritional, metabolic and toxic exposures before, during and after pregnancy. The nutritional exposures that have been studied most commonly in human cohorts are

generalised nutritional deficiencies (e.g. famine exposure), single micronutrient (e.g. vitamin B12) deficiencies. The major limitation of these studies is that studies of famine exposure have not quantified the specific nutritional insult that pregnant women are exposed to, and therefore it is difficult to interpret the outcome of these studies in a more general light. In contrast, the studies of vitamin B12 deficiency, e.g. PMNS, are well quantified using biomarkers of nutritional status from pre-conception to the post-natal period, but these studies may have limited applicability outside of India where a lacto-vegetarian diet is rare. Since the PMNS studies were performed, new data on the effects of other one-carbon metabolites, e.g. choline and betaine, has been generated and shows a potentially important role in developmental processes and offspring phenotype (e.g. (232)). The limitations of these human studies affect the potential to combine understanding of the pathogenic processes involved in these models, but are partly overcome by animal models where nutritional insults can be quantified and regulated. In the future, the application of metabolomic approaches may improve the ability to quantify multiple nutritional makers in these cohort studies and their applicability to small sample volumes may enable testing at multiple time points in order to understand specific periods of susceptibility. Additionally, data collected on maternal diet, even using the most sensitive of assays, may not fully reflect nutrient delivery to the fetus that may be affected by other factors such as placental function. Studies of nutritional programming should incorporate cord blood assays as an additional window to characterise the in utero environment.

Programming via metabolic dysregulation, e.g. diabetes in pregnancy, has been studied in many human contexts and the mouse model presented in Chapter 6 adds to a previously minimal body of literature using a murine model. Again, these models are limited by the detail in which the maternal phenotype is studied, and many of the studies of diabetes in pregnancy (including those in Chapter 6 and 7) have a glucocentric approach that does not take into account the potential programming of fat mass via leptin and other processes. Studies have also shown that 'stress' in utero may programme cardiometabolic phenotypes, with a putative role via cortisol, (e.g. (198)) but do not sufficiently investigate the possible confounding influences on glucose and insulin metabolism, altered cortisol binding globulin dynamics, nor the possible confounding of maternal stress with poor nutritional status, disruptions in circadian rhythmicity and adverse health behaviours. Finally, the increasing prevalence of gestational diabetes in transitional populations may result in a dichotomy of an excessive glucose environment in combination with specific micronutrient deficiencies.

Toxin exposure in pregnancy is less well studied but may have a role in programming of adult disease. Studies of arsenic exposure in Bangladesh have produced variable results, but suggest

that this may have an impact on thymic development (189) and cognitive development (233). These studies have been performed in the same region as the Matlab famine study, thus highlighting the possible combined or confounding influences of arsenic and nutrition in the offspring studied, not least because arsenic has multiple influences on the one-carbon cycle.

The overlap found between DNA methylation variants in the Matlab study and published Gambian study suggests possible common pathways that could underlie fetal programming via different environmental exposures (80). These pathways should be studied further in future investigations, and will be optimised by using similar experimental platforms to allow direct comparison.

8.4.1.2 Timing of exposure

The Dutch Winter Hunger studies suggest that the epigenetic changes associated with their fetal programming model are specific to the peri-conceptual period (109). In contrast, the original study that identified the diabetic phenotype in offspring exposed to famine showed that late gestational exposure conferred the highest risk (47). This study also identified an excess risk of major affective disorder in offspring exposed to famine in the second and third trimester, compared to those unexposed. The discordance of these findings suggest that the epigenetic memory of this late gestational exposure may be specific to tissues directly involved in pathogenesis, rather than the whole blood used in their epigenetic studies. Periconceptual exposure may suggest that environmental factors have influenced early epigenetic reprogramming and/or set down epigenetic variants that are perpetuated through all 3 developing germ layers.

In contrast, to the Dutch Winter Hunger findings, the Matlab study in Chapter 4 identified whole blood DNA methylation differences in those exposed to famine in early postnatal life as well as in utero (the specific timing in utero was not determined) suggesting that whole blood may still be a useful marker of epigenetic signatures involved in fetal programming.

8.4.1.3 Phenotypic outcomes

The phenotypic outcomes measured in fetal programming studies vary in their detail and scope. Many researchers have used birth weight as a window to understand the in utero environment, fetal growth and risk of disease in later life. Assumptions are frequently made that low birth weight reflects fetal growth restriction and future cardiometabolic disease risk.

However, birth weight is a variable that is affected by non-pathological factors including parental height, weight and ethnicity, maternal parity and smoking, and gestational age at delivery, as well as pathological factors such as prematurity, hypoxia, nutritional deficit and genetic aberrations. The inconsistent and non-linear relationship between perinatal mortality and birth weight has been well reported and is thought to be due to non-causal pathways such as maternal smoking (234). Similar non-causal relationships may underlie the association between birth weight and adult disease in models of fetal programming and low birth weight should not be assumed to reflect and need further investigation. The wide range of programming studies performed by Fall and Yajnik in India (e.g. PMNS) show the importance of detailed neonatal anthropometry to measure adiposity, as well as longitudinal phenotyping with more detailed biochemical and metabolic studies of older children. Newer techniques to measure fat mass and body composition, e.g. the Peapod, could provide important insights for future studies. In addition, longer-term phenotypic outcome measures relating to cardiovascular health, bone strength, cognitive function and physical capability are included in some cohort studies and will provide an invaluable insight into the outcomes of fetal programming and confounding influences.

8.4.1.4 Postnatal influences

Postnatal influences on fetal programming, and its molecular basis, are well studied in animal models but not in human studies. The Matlab famine study presented in Chapter 4 suggests that postnatal life is itself a susceptible developmental stage to epigenomic change induced through famine exposure. Future studies must include a postnatal phase to identify whether exposure to nutritional deficits are solely maternal influences in utero, or whether they continue during breastfeeding and post-weaning. These factors are important given also the role of 'catch-up growth' in mediating further excess risk towards cardiometabolic risk (235) and are rarely studied in combination with in utero programming studies in humans.

8.4.1.5 Stochastic environmental influences in later life

Beyond developmental life, a range of stochastic influences may give rise to epigenomic variation in different tissues, suggested by the diverging epigenome of ageing twins throughout the lifecourse. Stochastic factors are, by definition, ill defined and difficult to characterise. One such stochastic environmental factor that has gained recent notoriety as a possible influence on diabetes risk is that of circadian rhythm disruption. Epidemiological

studies have identified an increased risk of metabolic syndrome in individuals who work shifts (236). Recent mechanistic studies suggest that this may be due to disruption of circadian rhythms and a consequent decrease in metabolic rate and pancreatic insulin secretion (237). Circadian control may be under epigenetic control by sirtuins, a family of proteins with histone-modifying potential via NAD^+-dependent deacetylation. One such sirtuin, SIRT1, has also been implicated in both ageing and is down regulated in insulin resistant individuals (238). Further insights into this area are likely to elucidate the common mechanisms behind ageing and diseases that become increasingly common in old age, such as type 2 diabetes, and may yield translational insights via agents that enhance *SIRT1* expression such as resveratrol (239) and metformin (240) (241). However, these studies highlight the need to characterise stochastic environmental exposures in clinical cohorts as well as the importance of studying different epigenetic modifications, including those carried by histones. An interesting observation by Sandovici et al. shows that the combined effects of maternal diet-induced fetal programming and ageing of exposed offspring has a combined, and detrimental, effect on Hnf4a epigenetic regulation expression in rat islets, and hypothesise that this may have a role in the development of type 2 diabetes (229). This study highlights the importance of examining a range of different stochastic influences in a longitudinal manner to maximise the understanding of how genetic, epigenetic and environmental influences interact through the lifecourse.

8.4.1.6 Genetic influences

As discussed earlier, the molecular basis of fetal programming is best studied with an integrated approach using epigenomic and genomic studies and functional analysis. The role of the fetal genotype in modulating fetal growth has been well characterised by GWAS studies (e.g. (21)) but the role of these genetic variants, e.g. glucokinase polymorphisms, in programming long-term cardiometabolic disease risk are not known. This angle is an important one to pursue in future studies and the collection of parental DNA samples will help understand the relative contribution of genotypic factors, and the potential conflicting influences of certain mother-child genetic polymorphisms.

Comparison of epigenomic datasets to existing genomic insights is difficult in studies of people of Asian origin, such as those in this thesis, due to the paucity of GWAS studies in these populations. At the present time, the genetic risk of type 2 diabetes in Asians is assumed but not characterised. Whilst many assume that Asians have a higher genetic risk towards cardiometabolic disease compared to White Europeans, this is far from confirmed. Chauhan et

al. have suggested that the GWAS hits derived in European studies have stronger effect sizes in Asians (242). GWAS hits may also perform different functions in Asian populations, and studies of *FTO* in South Asians suggest that FTO risk alleles increase the risk of type 2 diabetes independent of obesity, unlike in Europeans (243). A study of genetic variants regulating the one-carbon cycle was performed in Indians to understand whether the programming influences of B12 deficiency could have a genetic basis, but no association was found between these variants and risk of type 2 diabetes in the Indians studied (244). Genetic variation between South Asian groups is another important consideration in these studies due to the differences in populations with Dravidian (indigenous) and European ancestries, and these differences are thought to affect SNP variation, CNVs and mitochondrial inheritance (245). The Indian Genome Variation Consortium should soon characterise these differences, and the results of the latter will be released through HapMap.

8.4.1.7 Paternal effects

Recent evidence suggests that paternal insulin resistance and diabetes has an inverse relationship with newborn size and adiposity (246), but does not elucidate the molecular mechanism whereby this occurs. Paternal genetic influences for this association are plausible, but a study by Carone et al. suggests that the paternal diet may influence offspring hepatic gene expression towards a cardiometabolic disease phenotype (247), and is an important factor to consider in future studies.

8.4.2 Epigenetic reprogramming and inheritance

The experiments designed to elucidate the epigenomic basis of fetal programming set out in this thesis concentrate on identifying DNA methylation differences in F_1 offspring exposed and unexposed to an environmental insult. This approach yields insights into the possible role of epigenetic variation in the programmed phenotype and suggests a direct role on the developing epigenome, but does provide insights into the processes of epigenetic reprogramming. Understanding the mechanisms of epigenetic reprogramming is a complex field and was beyond the scope of this thesis, but is an important consideration in understanding the putative mechanisms by which disease risk may be perpetuated across generations. Epigenetic reprogramming is understood as a process by which primordial germ cells erase their DNA methylation and histone modifications on a genomic scale and in complex stages that are poorly understood and is summarised in a recent paper studying its

relationship with the onset of pluripotency (248). The stages in which demethylation occurs varies according to the epigenomic landscape, with specific certain genomic features (e.g. imprinted regions) that carry long-term epigenetic memory undergoing demethylation later than other regions. Demethylation is thought to include both passive and active processes and the mediators of the latter are elusive, and poorly characterised, although recent studies have suggested a role for DNA deaminases (249). Recent studies also suggest that some regions of the genome, e.g. retro-elements, may resist demethylation leading to only partial demethylation during primordial germ cell development (250). Remethylation of the genome occurs at around the time of embryonic implantation and occurs with the assistance of methyltransferases, e.g. Dnmt3 and Dnmt3a.

These processes of epigenetic reprogramming may have the potential to induce epigenetic inheritance, via the stable transmission of epigenetic marks through successive generations. Animal models using the A^{vy} and $Axin^{Fu}$ mice have identified that retro-transposon regions that are resistant to full demethylation during reprogramming are locations that seem to permit epigenetic inheritance to occur (74). In addition to these retrotransposons, Seisenberger et al. have recently identified specific CpG islands that are variably erased during epigenetic reprogramming, and note that one of these was within a gene, *Exoc4*, that is involved in insulin-stimulated glucose transport and associated with type 2 diabetes (248).

In contrast to these animal models, there is, there is scant evidence that true epigenetic inheritance occurs in humans, despite a great deal of suggestion that it may occur and underlie the transgenerational patterns of complex disease risk. The paucity of evidence of the molecular mechanisms of epigenetic reprogramming in humans reflects the technical difficulties of studying these processes in primordial germ cells, gametes and developing embryos. Without the opportunity to study these processes in human development, the question of whether environmental influences on fetal programming are mediated through epigenetic reprogramming will remain an unknown until molecular studies can be performed in multi-generational cohorts.

8.4.3 Evolutionary perspectives

The potential evolutionary roles of the epigenetic modifications described in these studies are important to consider in the wider context of changing disease susceptibility. The identification of a CpG-SNP in the Type 2 diabetes study is interesting to consider in the light of

a recent suggestion that genetic-epigenetic variants have a role in evolutionary adaptation. Authors have suggested that the loss or gain of CpG dinucleotides over time can add to the variance at genetic features, e.g. SNPs, and allow evolutionary adaptation to the environment or changing disease susceptibility over time (173). Furthermore, the deamination of methylated CpGs (as was seen in this study) is thought to have a potential role in the formation of transcription factor binding sites, itself an important evolutionary process, and has previously been noted in the p53 binding sites which themselves have a critical role in regulation of insulin resistance (175).

The wider evolutionary context of epigenomic variation in fetal programming has already been discussed (section 8.3.4), and relates to the possible role of parent-of-origin effects in mediating adaptation to the environment.

An interesting observation across many fetal programming studies is that famine exposure in utero may influence sex ratio at birth (251). In over 300,000 people studied during and after the Great Leap Forward famine, there was an abrupt decline in the numbers of female births in the famine-exposed mothers. This observation fits with the hypothesis that mothers in poor condition are likely to give birth to daughters and those in healthy condition are more likely to give birth to sons; this would serve an evolutionary purpose to preserve a population by promoting reproduction of the famine-survivors. These evolutionarily adaptive mechanisms have not been studied on a molecular level but could provide a fascinating angle to future epigenomic studies in famine-based cohorts.

8.5 Final conclusions and future direction

The studies performed in this thesis have provided novel insights into epigenetic variants associated with genetic variation in type 2 diabetes and in studies of fetal programming and have achieved most of the original objectives set out in the introduction. First, the identification of an *FTO*-associated CpG-SNP has elucidated a possible mechanism by which genetic and epigenetic variation can have combined functional consequence. The Matlab study shows reproducible variation in DNA methylation in offspring exposed to famine in utero and in postnatal life, compared to those unexposed and sheds light on the potential molecular mechanisms of fetal programming of adult cardiometabolic disease. The related study of offspring in the Pune Maternal Nutrition study has been hampered by difficulties in analysing

and reproducing a Medip-seq dataset, but it is hoped that additional studies using the Illumina 450k array and combined cell line work will elucidate the mechanisms by which insulin resistance and adiposity are programmed. A studies of gestational diabetes in mice has confirmed the long-term programmed phenotype in offspring exposed to maternal hyperglycaemia and epigenomic studies in embryonic offspring have identified DNA methylation differences in exposed vs. unexposed. This study has provided a platform for investigations of functional pathways and the tissue- and age-specific effects of this programming, all of which are currently underway. A small human cohort of women of Asian origin with and without gestational diabetes has provided an invaluable opportunity to identify whether nutritional and metabolic disturbances coexist in pregnant women and has suggested that an apparently 'healthy' UK population has a high prevalence of micronutrient deficiencies. This study of gestational diabetes is now being used to generate epigenomic insights into the effects on offspring of adverse maternal pregnancy environments and will be used to complement the murine model.

The difficulties of taking an unbiased 'discovery-based' approach to studying the epigenome has been highlighted several times in this thesis and has been hindered by small sample sizes and technical difficulties. Methylation array based studies have offered a robust alternative to Medip-seq, but do not offer the same discovery-based approach to investigate epigenomic variation associated with programming. The future of these kinds of studies could take two directions, first to expand on the association-based approach using larger sample numbers to identify epigenetic variants associated with a programmed environment or outcome. This approach, best served by a genome-wide or epigenomic experimental platform, would need careful study design and estimation of sample size to overcome the difficulties encountered in these studies. Such epigenome-wide association studies could be integrated with genomic and gene expression studies to expand their functional relevance. Second, this area of research could take a more focused and hypothesis-driven approach, perhaps following up candidate regions of epigenomic variation and studying these in different environments and sample types. A combination of both approaches, and their application to model systems, such as cell lines, looks set to expand the understanding of gene-environment interactions in developmental life and their impact on future cardiometabolic disease risk. This understanding is crucial to elucidate the mechanisms by which diabetes prevalence is increasing and it is hoped that these molecular insights, coupled with intervention studies, and advancements in –omic studies, will enable their direct translation to targeted prevention studies and the design of specific therapeutic interventions.

References

1. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants. Lancet. 2011 Jul 2;378(9785):31–40.
2. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2013 Dec 15;380(9859):2095–128.
3. Wellcome T, Case T, Consortium C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 Jun 7;447(7145):661–78.
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747–53.
5. Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, et al. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. The New England journal of medicine. 2001 Sep 13;345(11):790–7.
6. Klein S, Sheard NF, Pi-Sunyer X, Daly A, Wylie-Rosett J, Kulkarni K, et al. Weight management through lifestyle modification for the prevention and management of type 2 diabetes: rationale and strategies. A statement of the American Diabetes Association, the North American Association for the Study of Obesity, and the American So. The American journal of clinical nutrition. 2004 Aug;80(2):257–63.
7. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. Nature reviews. Genetics. 2007 Apr;8(4):253–62.
8. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nature genetics. 2007 Jan;39(1):61–9.
9. DeFronzo RA. Pathogenesis of type 2 diabetes mellitus. The Medical clinics of North America. 2004 Jul;88(4):787–835, ix.
10. Muoio DM, Newgard CB. Mechanisms of disease: molecular and metabolic mechanisms of insulin resistance and beta-cell failure in type 2 diabetes. Nature reviews. Molecular cell biology. 2008 Mar;9(3):193–205.
11. Florez JC. Newly identified loci highlight beta cell dysfunction as a key cause of type 2 diabetes: where are the insulin resistance genes? Diabetologia. 2008 Jul;51(7):1100–10.
12. Gloyn AL, Braun M, Rorsman P. Type 2 diabetes susceptibility gene TCF7L2 and its role in beta-cell function. Diabetes. 2009 Apr;58(4):800–2.
13. Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ (Clinical research ed.). 2000 Aug 12;321(7258):405–12.
14. Skyler JS, Bergenstal R, Bonow RO, Buse J, Deedwania P, Gale EAM, et al. Intensive glycemic control and the prevention of cardiovascular events: implications of the ACCORD, ADVANCE, and VA diabetes trials: a position statement of the American Diabetes Association and a scientific statement of the American College of Cardiology. Circulation. 2009 Jan 20;119(2):351–7.
15. Holman RR, Paul SK, Bethel MA, Matthews DR, Neil HAW. 10-year follow-up of intensive glucose control in type 2 diabetes. The New England journal of medicine. 2008 Oct 9;359(15):1577–89.
16. Brownlee M. The pathobiology of diabetic complications: a unifying mechanism. Diabetes. 2005 Jun;54(6):1615–25.
17. Lindström J, Ilanne-Parikka P, Peltonen M, Aunola S, Eriksson JG, Hemiö K, et al. Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study. Lancet. 2006 Nov 11;368(9548):1673–9.
18. Hales CN, Barker DJ, Clark PM, Cox LJ, Fall C, Osmond C, et al. Fetal and infant growth and impaired glucose tolerance at age 64. BMJ (Clinical research ed.). 1991 Oct 26;303(6809):1019–22.
19. Osmond C, Barker DJ, Winter PD, Fall CH, Simmonds SJ. Early growth and death from cardiovascular disease in women. BMJ (Clinical research ed.). 1993 Dec 11;307(6918):1519–24.
20. Hattersley AT, Tooke JE. The fetal insulin hypothesis: an alternative explanation of the association of low birthweight with diabetes and vascular disease. Lancet. 1999 May 22;353(9166):1789–92.

21. Freathy RM, Mook-Kanamori DO, Sovio U, Prokopenko I, Timpson NJ, Berry DJ, et al. Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. Nature genetics. 2010 May;42(5):430–5.
22. Diagnosis and classification of diabetes mellitus. Diabetes care. 2011 Jan;34 Suppl 1:S62–9.
23. Agarwal MM, Dhatt GS, Shah SM. Gestational diabetes mellitus: simplifying the international association of diabetes and pregnancy diagnostic algorithm using fasting plasma glucose. Diabetes care. 2010 Sep;33(9):2018–20.
24. Wong VW. Gestational diabetes mellitus in five ethnic groups: a comparison of their clinical characteristics. Diabetic medicine : a journal of the British Diabetic Association. 2012 Mar;29(3):366–71.
25. Carpenter MW, Coustan DR. Criteria for screening tests for gestational diabetes. American journal of obstetrics and gynecology. 1982 Dec 1;144(7):768–73.
26. Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabetic medicine : a journal of the British Diabetic Association. 1998 Jul;15(7):539–53.
27. Kwak SH, Kim S-H, Cho YM, Go MJ, Cho YS, Choi SH, et al. A Genome-Wide Association Study of Gestational Diabetes Mellitus in Korean Women. Diabetes. 2012 Jan 10;61(2):531–41.
28. Novakovic B, Gordon L, Robinson WP, Desoye G, Saffery R. Glucose as a fetal nutrient: dynamic regulation of several glucose transporter genes by DNA methylation in the human placenta across gestation. The Journal of nutritional biochemistry. 2013 Jan;24(1):282–8.
29. Kim H, Toyofuku Y, Lynn FC, Chak E, Uchida T, Mizukami H, et al. Serotonin regulates pancreatic beta cell mass during pregnancy. Nature medicine. 2010 Jul;16(7):804–8.
30. Miehle K, Stepan H, Fasshauer M. Leptin, adiponectin and other adipokines in gestational diabetes mellitus and pre-eclampsia. Clinical endocrinology. 2012 Jan;76(1):2–11.
31. Yamashita H, Shao J, Ishizuka T, Klepcyk PJ, Muhlenkamp P, Qiao L, et al. Leptin administration prevents spontaneous gestational diabetes in heterozygous Lepr(db/+) mice: effects on placental leptin and fetal growth. Endocrinology. 2001 Jul;142(7):2888–97.
32. Urbanek M, Hayes MG, Lee H, Freathy RM, Lowe LP, Ackerman C, et al. The role of inflammatory pathway genetic variation on maternal metabolic phenotypes during pregnancy. PloS one. 2012 Jan;7(3):e32958.
33. Metzger BE, Lowe LP, Dyer AR, Trimble ER, Chaovarindr U, Coustan DR, et al. Hyperglycemia and adverse pregnancy outcomes. The New England journal of medicine. 2008 May 8;358(19):1991–2002.
34. Fraser A, Tilling K, Macdonald-Wallis C, Sattar N, Brion M-J, Benfield L, et al. Association of maternal weight gain in pregnancy with offspring obesity and metabolic and vascular traits in childhood. Circulation. 2010 Jun 15;121(23):2557–64.
35. Pettitt DJ, Bennett PH, Saad MF, Charles MA, Nelson RG, Knowler WC. Abnormal glucose tolerance during pregnancy in Pima Indian women. Long-term effects on offspring. Diabetes. 1991 Dec;40 Suppl 2:126–30.
36. Gauguier D, Bihoreau MT, Ktorza A, Berthault MF, Picon L. Inheritance of diabetes mellitus as consequence of gestational hyperglycemia in rats. Diabetes. 1990 Jun;39(6):734–9.
37. Yamashita H, Shao J, Qiao L, Pagliassotti M, Friedman JE. Effect of spontaneous gestational diabetes on fetal and postnatal hepatic insulin resistance in Lepr(db/+) mice. Pediatric research. 2003 Mar;53(3):411–8.
38. Dabelea D, Hanson RL, Lindsay RS, Pettitt DJ, Imperatore G, Gabir MM, et al. Intrauterine exposure to diabetes conveys risks for type 2 diabetes and obesity: a study of discordant sibships. Diabetes. 2000 Dec;49(12):2208–11.
39. Lindsay RS, Nelson SM, Walker JD, Greene SA, Milne G, Sattar N, et al. Programming of adiposity in offspring of mothers with type 1 diabetes at age 7 years. Diabetes care. 2010 May;33(5):1080–5.
40. Krishnaveni G V, Veena SR, Hill JC, Kehoe S, Karat SC, Fall CHD. Intrauterine exposure to maternal diabetes is associated with higher adiposity and insulin resistance and clustering of cardiovascular risk markers in Indian children. Diabetes care. 2010 Feb;33(2):402–4.
41. James SJ, Melnyk S, Pogribna M, Pogribny IP, Caudill MA. Elevation in S-adenosylhomocysteine and DNA hypomethylation: potential epigenetic mechanism for homocysteine-related pathology. The Journal of nutrition. 2002 Aug;132(8 Suppl):2361S–2366S.

42. Yi P, Melnyk S, Pogribna M, Pogribny IP, Hine RJ, James SJ. Increase in plasma homocysteine associated with parallel increases in plasma S-adenosylhomocysteine and lymphocyte DNA hypomethylation. The Journal of biological chemistry. 2000 Sep 22;275(38):29318–23.
43. Yajnik CS, Deshpande SS, Jackson a a, Refsum H, Rao S, Fisher DJ, et al. Vitamin B12 and folate concentrations during pregnancy and insulin resistance in the offspring: the Pune Maternal Nutrition Study. Diabetologia. 2008 Jan;51(1):29–38.
44. Sinclair KD, Allegrucci C, Singh R, Gardner DS, Sebastian S, Bispham J, et al. DNA methylation, insulin resistance, and blood pressure in offspring determined by maternal periconceptional B vitamin and methionine status. Proceedings of the National Academy of Sciences of the United States of America. 2007 Dec 4;104(49):19351–6.
45. McKay JA, Groom A, Potter C, Coneyworth LJ, Ford D, Mathers JC, et al. Genetic and non-genetic influences during pregnancy on infant global and site specific DNA methylation: role for folate gene variants and vitamin B12. PloS one. 2012 Jan 30;7(3):e33290.
46. Hoyo C, Murtha AP, Schildkraut JM, Jirtle RL, Demark-Wahnefried W, Forman MR, et al. Methylation variation at IGF2 differentially methylated regions and maternal folic acid use before and during pregnancy. Epigenetics. 2011 Jul 1;6(7):928–36.
47. Ravelli a C, Van der Meulen JH, Michels RP, Osmond C, Barker DJ, Hales CN, et al. Glucose tolerance in adults after prenatal exposure to famine. Lancet. 1998 Jan 17;351(9097):173–7.
48. De Rooij SR, Painter RC, Phillips DIW, Osmond C, Michels RPJ, Godsland IF, et al. Impaired insulin secretion after prenatal exposure to the Dutch famine. Diabetes care. 2006 Aug;29(8):1897–901.
49. Yajnik CS. Nutrient-mediated teratogenesis and fuel-mediated teratogenesis: two pathways of intrauterine programming of diabetes. International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics. 2009 Mar;104 Suppl S27–31.
50. Ozanne SE, Wang CL, Coleman N, Smith GD. Altered muscle insulin sensitivity in the male offspring of protein-malnourished rats. The American journal of physiology. 1996 Dec;271(6 Pt 1):E1128–34.
51. Li Y, He Y, Qi L, Jaddoe VW, Feskens EJM, Yang X, et al. Exposure to the Chinese famine in early life and the risk of hyperglycemia and type 2 diabetes in adulthood. Diabetes. 2010 Oct;59(10):2400–6.
52. Stanner SA, Bulmer K, Andrès C, Lantseva OE, Borodina V, Poteen V V, et al. Does malnutrition in utero determine diabetes and coronary heart disease in adulthood? Results from the Leningrad siege study, a cross sectional study. BMJ (Clinical research ed.). 1997 Nov 22;315(7119):1342–8.
53. Martin-Gronert MS, Ozanne SE. Mechanisms linking suboptimal early nutrition and increased risk of type 2 diabetes and obesity. The Journal of nutrition. 2010 Mar;140(3):662–6.
54. A haplotype map of the human genome. Nature. 2005 Oct 27;437(7063):1299–320.
55. Kaminsky Z a, Tang T, Wang S-C, Ptak C, Oh GHT, Wong AHC, et al. DNA methylation profiles in monozygotic and dizygotic twins. Nature genetics. 2009 Feb;41(2):240–5.
56. Cheverud JM, Hager R, Roseman C, Fawcett G, Wang B, Wolf JB. Genomic imprinting effects on adult body composition in mice. Proceedings of the National Academy of Sciences of the United States of America. 2008 Mar 18;105(11):4253–8.
57. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar a H, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008 May 2;133(3):523–36.
58. Rakyan VK, Down T a, Thorne NP, Flicek P, Kulesha E, Gräf S, et al. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome research. 2008 Sep;18(9):1518–29.
59. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007 May 18;129(4):823–37.
60. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature methods. 2007 Aug;4(8):651–7.
61. Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature. 2010 Apr 15;464(7291):1082–6.
62. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nature reviews. Genetics. 2009 Mar;10(3):161–72.
63. Richards EJ. Inherited epigenetic variation--revisiting soft inheritance. Nature reviews. Genetics. 2006 May;7(5):395–401.

64. Heijmans BT, Kremer D, Tobi EW, Boomsma DI, Slagboom PE. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. Human molecular genetics. 2007 Mar 1;16(5):547–54.
65. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, et al. Parental origin of sequence variants associated with complex diseases. Nature. Nature Publishing Group; 2009 Dec 17;462(7275):868–74.
66. Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nature genetics. 2008 Jul;40(7):904–8.
67. Schalkwyk LC, Meaburn EL, Smith R, Dempster EL, Jeffries AR, Davies MN, et al. Allelic skewing of DNA methylation is widespread across the genome. American journal of human genetics. The American Society of Human Genetics; 2010 Feb 12;86(2):196–212.
68. Schilling E, El Chartouni C, Rehli M. Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. Genome research. 2009 Nov;19(11):2028–35.
69. Kadota M, Yang HH, Hu N, Wang C, Hu Y, Taylor PR, et al. Allele-specific chromatin immunoprecipitation studies show genetic influence on chromatin state in human genome. PLoS genetics. 2007 May 18;3(5):e81.
70. McDaniell R, Lee B-K, Song L, Liu Z, Boyle AP, Erdos MR, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science (New York, N.Y.). 2010 Apr 9;328(5975):235–9.
71. Ashe A, Morgan DK, Whitelaw NC, Bruxner TJ, Vickaryous NK, Cox LL, et al. A genome-wide screen for modifiers of transgene variegation identifies genes with critical roles in development. Genome biology. 2008 Jan;9(12):R182.
72. Blewitt ME, Gendrel A-V, Pang Z, Sparrow DB, Whitelaw N, Craig JM, et al. SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. Nature genetics. 2008 May;40(5):663–9.
73. Morgan HD, Sutherland HG, Martin DI, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. Nature genetics. 1999 Nov;23(3):314–8.
74. Rakyan VK, Chong S, Champ ME, Cuthbert PC, Morgan HD, Luu KVK, et al. Transgenerational inheritance of epigenetic states at the murine Axin(Fu) allele occurs after maternal and paternal transmission. Proceedings of the National Academy of Sciences of the United States of America. 2003 Mar 4;100(5):2538–43.
75. Stöger R, Kajimura TM, Brown WT, Laird CD. Epigenetic variation illustrated by DNA methylation patterns of the fragile-X gene FMR1. Human molecular genetics. 1997 Oct;6(11):1791–801.
76. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proceedings of the National Academy of Sciences of the United States of America. 2005 Jul 26;102(30):10604–9.
77. Lillycrop K a, Phillips ES, Torrens C, Hanson M a, Jackson A a, Burdge GC. Feeding pregnant rats a protein-restricted diet persistently alters the methylation of specific cytosines in the hepatic PPAR alpha promoter of the offspring. The British journal of nutrition. 2008 Aug;100(2):278–82.
78. Lillycrop KA, Phillips ES, Jackson AA, Hanson MA, Burdge GC. Dietary protein restriction of pregnant rats induces and folic acid supplementation prevents epigenetic modification of hepatic gene expression in the offspring. The Journal of nutrition. 2005 Jun;135(6):1382–6.
79. Burns SP, Desai M, Cohen RD, Hales CN, Iles RA, Germain JP, et al. Gluconeogenesis, glucose handling, and structural changes in livers of the adult offspring of rats partially deprived of protein during pregnancy and lactation. The Journal of clinical investigation. 1997 Oct 1;100(7):1768–74.
80. Waterland RA, Kellermayer R, Laritsky E, Rayco-Solon P, Harris RA, Travisano M, et al. Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. Whitelaw E, editor. PLoS genetics. Public Library of Science; 2010 Jan;6(12):e1001252.
81. Finer S, Holland ML, Nanty L, Rakyan VK. The hunt for the epiallele. Environmental and molecular mutagenesis. 2011 Jan;52(1):1–11.
82. Mitchell KJ. The genetics of brain wiring: from molecule to mind. PLoS biology. 2007 Apr;5(4):e113.
83. Sasaki H, Matsui Y. Epigenetic events in mammalian germ-cell development: reprogramming and beyond. Nature reviews. Genetics. 2008 Feb;9(2):129–40.
84. Oswald J, Engemann S, Lane N, Mayer W, Olek A, Fundele R, et al. Active demethylation of the paternal genome in the mouse zygote. Current biology : CB. 2000 Apr 20;10(8):475–8.

85. Suter CM, Martin DIK, Ward RL. Germline epimutation of MLH1 in individuals with multiple cancers. Nature genetics. 2004 May;36(5):497–501.
86. Cropley JE, Suter CM, Beckman KB, Martin DIK. Germ-line epigenetic modification of the murine A vy allele by nutritional supplementation. Proceedings of the National Academy of Sciences of the United States of America. 2006 Nov 14;103(46):17308–12.
87. Novakovic B, Sibson M, Ng HK, Manuelpillai U, Rakyan V, Down T, et al. Placenta-specific methylation of the vitamin D 24-hydroxylase gene: implications for feedback autoregulation of active vitamin D levels at the fetomaternal interface. The Journal of biological chemistry. 2009 May 29;284(22):14838–48.
88. Bourque DK, Avila L, Peñaherrera M, Von Dadelszen P, Robinson WP. Decreased placental methylation at the H19/IGF2 imprinting control region is associated with normotensive intrauterine growth restriction but not preeclampsia. Placenta. 2010 Mar;31(3):197–202.
89. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proceedings of the National Academy of Sciences of the United States of America. 2008 Nov 4;105(44):17046–9.
90. Waterland R a, Lin J-R, Smith C a, Jirtle RL. Post-weaning diet affects genomic imprinting at the insulin-like growth factor 2 (Igf2) locus. Human molecular genetics. 2006 Mar 1;15(5):705–16.
91. Weaver ICG, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, Seckl JR, et al. Epigenetic programming by maternal behavior. Nature neuroscience. 2004 Aug;7(8):847–54.
92. Rakyan VK, Down TA, Maslau S, Andrew T, Yang T-P, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome research. 2010 Apr;20(4):434–9.
93. Gicquel C, Rossignol S, Cabrol S, Houang M, Steunou V, Barbu V, et al. Epimutation of the telomeric imprinting center region on chromosome 11p15 in Silver-Russell syndrome. Nature genetics. 2005 Sep;37(9):1003–7.
94. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. Nature reviews. Cancer. 2011 Oct;11(10):726–34.
95. Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, et al. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. PLoS genetics. 2011 Sep;7(9):e1002300.
96. Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, Kong A, et al. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nature genetics. 2011 Jun;43(6):561–4.
97. Demars J, Shmela ME, Rossignol S, Okabe J, Netchine I, Azzi S, et al. Analysis of the IGF2/H19 imprinting control region uncovers new genetic defects, including mutations of OCT-binding sequences, in patients with 11p15 fetal growth disorders. Human molecular genetics. 2010 Mar 1;19(5):803–14.
98. Mackay DJG, Callaway JLA, Marks SM, White HE, Acerini CL, Boonen SE, et al. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. Nature genetics. 2008 Aug;40(8):949–51.
99. Ivanova E, Chen J-H, Segonds-Pichon A, Ozanne SE, Kelsey G. DNA methylation at differentially methylated regions of imprinted genes is resistant to developmental programming by maternal nutrition. Epigenetics : official journal of the DNA Methylation Society. 2012 Oct;7(10):1200–10.
100. Yang BT, Dayeh TA, Volkov PA, Kirkpatrick CL, Malmgren S, Jing X, et al. Increased DNA methylation and decreased expression of PDX-1 in pancreatic islets from patients with type 2 diabetes. Molecular endocrinology (Baltimore, Md.). 2012 Jul;26(7):1203–12.
101. Beyan H, Down TA, Ramagopalan S V, Uvebrant K, Nilsson A, Holland ML, et al. Guthrie card methylomics identifies temporally stable epialleles that are present at birth in humans. Genome research. 2012 Nov;22(11):2138–45.
102. Ling C, Poulsen P, Simonsson S, Rönn T, Holmkvist J, Almgren P, et al. Genetic and epigenetic factors are associated with expression of respiratory chain component NDUFB6 in human skeletal muscle. The Journal of clinical investigation. 2007 Nov;117(11):3427–35.
103. Ling C, Del Guerra S, Lupi R, Rönn T, Granhall C, Luthman H, et al. Epigenetic regulation of PPARGC1A in human type 2 diabetic islets and effect on insulin secretion. Diabetologia. 2008 Apr;51(4):615–22.
104. El-Osta A, Brasacchio D, Yao D, Pocai A, Jones PL, Roeder RG, et al. Transient high glucose causes persistent epigenetic changes and altered gene expression during subsequent normoglycemia. The Journal of experimental medicine. 2008 Sep 29;205(10):2409–17.

105. Brasacchio D, Okabe J, Tikellis C, Balcerczyk A, George P, Baker EK, et al. Hyperglycemia induces a dynamic cooperativity of histone methylase and demethylase enzymes associated with gene-activating epigenetic marks that coexist on the lysine tail. Diabetes. 2009 May;58(5):1229–36.
106. Pirola L, Balcerczyk A, W. Tothill R, Haviv I, Kaspi A, Lunke S, et al. Genome-wide analysis distinguishes hyperglycemia regulated epigenetic signatures of primary vascular cells. Genome Research. 2011 Sep 2;21(10):1601–15.
107. Bouchard L, Thibault S, Guay SP, Santure M, Monpetit A, St-Pierre J, et al. Leptin gene epigenetic adaptation to impaired glucose metabolism during pregnancy. Diabetes care. Am Diabetes Assoc; 2010;33(11):2436.
108. Bell CG, Teschendorff AE, Rakyan VK, Maxwell AP, Beck S, Savage D a. Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. BMC medical genomics. 2010 Jan;3:33.
109. Tobi EW, Lumey LH, Talens RP, Kremer D, Putter H, Stein AD, et al. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. Human molecular genetics. 2009 Nov 1;18(21):4046–53.
110. Brøns C, Jacobsen S, Nilsson E, Rönn T, Jensen CB, Storgaard H, et al. Deoxyribonucleic acid methylation and gene expression of PPARGC1A in human muscle is influenced by high-fat overfeeding in a birth-weight-dependent manner. The Journal of clinical endocrinology and metabolism. 2010 Jun;95(6):3048–56.
111. Park JH, Stoffers DA, Nicholls RD, Simmons RA. Development of type 2 diabetes following intrauterine growth retardation in rats is associated with progressive epigenetic silencing of Pdx1. 2008;118(6).
112. Bhandare R, Schug J, Le Lay J, Fox A, Smirnova O, Liu C, et al. Genome-wide analysis of histone modifications in human pancreatic islets. Genome research. 2010 Apr;20(4):428–33.
113. Zhang C, Qiu C, Hu FB, David RM, Van Dam RM, Bralley A, et al. Maternal plasma 25-hydroxyvitamin D concentrations and the risk for gestational diabetes mellitus. PloS one. 2008 Jan;3(11):e3753.
114. Down T a, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nature biotechnology. 2008 Jul;26(7):779–85.
115. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008 Mar 13;452(7184):215–9.
116. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009 Dec 19;462(7271):315–22.
117. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008 Aug 7;454(7205):766–70.
118. Bjornsson HT, Albert TJ, Ladd-Acosta CM, Green RD, Rongione MA, Middle CM, et al. SNP-specific array-based allele-specific expression analysis. Genome research. 2008 May;18(5):771–9.
119. Pogribny IP, Tryndyak VP, Bagnyukova T V, Melnyk S, Montgomery B, Ross S a, et al. Hepatic epigenetic phenotype predetermines individual susceptibility to hepatic steatosis in mice fed a lipogenic methyl-deficient diet. Journal of hepatology. European Association for the Study of the Liver; 2009 Jul;51(1):176–86.
120. Wells JCK. Historical cohort studies and the early origins of disease hypothesis: making sense of the evidence. The Proceedings of the Nutrition Society. 2009 May;68(2):179–88.
121. Shames DS, Girard L, Gao B, Sato M, Lewis CM, Shivapurkar N, et al. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. PLoS medicine. 2006 Dec;3(12):e486.
122. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010 Apr 1;464(7289):768–72.
123. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010 Apr 1;464(7289):773–7.
124. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011 Oct;98(4):288–95.

125. Ordway JM, Curran T. Methylation matters: modeling a manageable genome. Cell growth & differentiation : the molecular biology journal of the American Association for Cancer Research. 2002 Apr;13(4):149–62.
126. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behavioural brain research. 2001 Nov 1;125(1-2):279–84.
127. Qian H-R, Huang S. Comparison of false discovery rate methods in identifying genes with differential expression. Genomics. 2005 Oct;86(4):495–503.
128. De S, Michor F. DNA secondary structures and epigenetic determinants of cancer genome evolution. Nature structural & molecular biology. 2011 Aug;18(8):950–5.
129. Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, et al. MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. Genome research. 2008 Oct;18(10):1652–9.
130. Nix D a, Courdy SJ, Boucher KM. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC bioinformatics. 2008 Jan;9:523.
131. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America. 2003 Aug 5;100(16):9440–5.
132. Vining KJ, Pomraning KR, Wilhelm LJ, Priest HD, Pellegrini M, Mockler TC, et al. Dynamic DNA cytosine methylation in the Populus trichocarpa genome: tissue-level variation and relationship to gene expression. BMC genomics. 2012 Jan;13:27.
133. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000 May;25(1):25–9.
134. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC bioinformatics. 2010 Jan;11:587.
135. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome biology. 2003 Jan;4(4):210.
136. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Human molecular genetics. 2008 Oct 15;17(R2):R156–65.
137. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nature genetics. 2010 Feb;42(2):105–16.
138. Prokopenko I, McCarthy MI, Lindgren CM. Type 2 diabetes: new genes, new understanding. Trends in genetics : TIG. 2008 Dec;24(12):613–21.
139. Glaser RL, Ramsay JP, Morison IM. The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. Nucleic acids research. 2006 Jan 1;34(Database issue):D29–31.
140. Goeman JJ, Van de Geer SA, De Kort F, Van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics (Oxford, England). 2004 Jan 1;20(1):93–9.
141. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science (New York, N.Y.). 2002 Jul 21;296(5576):2225–9.
142. Forsberg EC, Downs KM, Christensen HM, Im H, Nuzzi PA, Bresnick EH. Developmentally dynamic histone acetylation pattern of a tissue-specific chromatin domain. Proceedings of the National Academy of Sciences of the United States of America. 2000 Dec 19;97(26):14494–9.
143. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov 27;456(7221):470–6.
144. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science (New York, N.Y.). 2007 May 11;316(5826):889–94.
145. Arndt PF, Petrov DA, Hwa T. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. Molecular biology and evolution. 2003 Nov;20(11):1887–96.
146. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome research. 2005 Jul;15(7):901–13.
147. Helgason A, Pálsson S, Thorleifsson G, Grant SFA, Emilsson V, Gunnarsdottir S, et al. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nature genetics. 2007 Feb;39(2):218–25.

148. Ragvin A, Moro E, Fredman D, Navratilova P, Drivenes Ø, Engström PG, et al. Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. Proceedings of the National Academy of Sciences of the United States of America. 2010 Jan 12;107(2):775–80.
149. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science (New York, N.Y.). 2007 Feb 9;315(5813):848–53.
150. Heap GA, Yang JHM, Downes K, Healy BC, Hunt KA, Bockett N, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Human molecular genetics. 2010 Jan 1;19(1):122–34.
151. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell. 2006 Sep 22;126(6):1189–201.
152. Toperoff G, Aran D, Kark JD, Rosenberg M, Dubnikov T, Nissan B, et al. Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. Human molecular genetics. 2012 Jan 15;21(2):371–83.
153. Almén MS, Jacobsson JA, Moschonis G, Benedict C, Chrousos GP, Fredriksson R, et al. Genome wide analysis reveals association of a FTO gene variant with epigenetic changes. Genomics. 2012 Mar;99(3):132–7.
154. Dina C, Meyre D, Gallina S, Durand E, Körner A, Jacobson P, et al. Variation in FTO contributes to childhood obesity and severe adult obesity. Nature genetics. 2007 Jun;39(6):724–6.
155. López-Bermejo A, Petry CJ, Díaz M, Sebastiani G, De Zegher F, Dunger DB, et al. The association between the FTO gene and fat mass in humans develops by the postnatal age of two weeks. The Journal of clinical endocrinology and metabolism. 2008 Apr;93(4):1501–5.
156. Boissel S, Reish O, Proulx K, Kawagoe-Takaki H, Sedgwick B, Yeo GSH, et al. Loss-of-function mutation in the dioxygenase-encoding FTO gene causes severe growth retardation and multiple malformations. American journal of human genetics. 2009 Jul;85(1):106–11.
157. Grunnet LG, Nilsson E, Ling C, Hansen T, Pedersen O, Groop L, et al. Regulation and function of FTO mRNA expression in human skeletal muscle and subcutaneous adipose tissue. Diabetes. 2009 Oct;58(10):2402–8.
158. Verlaan DJ, Ge B, Grundberg E, Hoberman R, Lam KCL, Koka V, et al. Targeted screening of cis-regulatory variation in human haplotypes. Genome research. 2009 Jan;19(1):118–27.
159. Meyre D, Proulx K, Kawagoe-takaki H, Vatin V, Gutie R, Lyon D, et al. Prevalence of Loss-of-Function FTO Mutations in Lean and Obese Individuals. Obesity. 2010;59(January):311–8.
160. Fischer J, Koch L, Emmerling C, Vierkotten J, Peters T, Brüning JC, et al. Inactivation of the Fto gene protects from obesity. Nature. 2009 Apr 16;458(7240):894–8.
161. Church C, Lee S, Bagg E a L, McTaggart JS, Deacon R, Gerken T, et al. A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. PLoS genetics. 2009 Aug;5(8):e1000599.
162. Ma M, Harding HP, O'Rahilly S, Ron D, Yeo GSH. Kinetic analysis of FTO (fat mass and obesity-associated) reveals that it is unlikely to function as a sensor for 2-oxoglutarate. The Biochemical journal. 2012 Jun 1;444(2):183–7.
163. Fawcett K a, Barroso I. The genetics of obesity: FTO leads the way. Trends in genetics : TIG. 2010 Jun;26(6):266–74.
164. Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, et al. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nature chemical biology. 2011 Dec;7(12):885–7.
165. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009 Sep 10;461(7261):199–205.
166. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009 May 7;459(7243):108–12.
167. Schmidl C, Klug M, Boeld TJ, Andreesen R, Hoffmann P, Edinger M, et al. Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. Genome research. 2009 Jul;19(7):1165–74.
168. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature genetics. 2009 Mar;41(2):178–86.
169. Lingohr MK, Buettner R, Rhodes CJ. Pancreatic beta-cell growth and survival--a role in obesity-linked type 2 diabetes? Trends in molecular medicine. 2002 Aug;8(8):375–84.

170. O'Rahilly S. Human genetics illuminates the paths to metabolic disease. Nature. 2009 Nov 19;462(7271):307–14.
171. Caqueret A, Boucher F, Michaud JL. Laminar organization of the early developing anterior hypothalamus. Developmental biology. 2006 Oct 1;298(1):95–106.
172. Braun MM, Etheridge A, Bernard A, Robertson CP, Roelink H. Wnt signaling is required at distinct stages of development for the induction of the posterior forebrain. Development (Cambridge, England). 2003 Dec;130(23):5579–87.
173. Feinberg AP, Irizarry RA. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proceedings of the National Academy of Sciences of the United States of America. 2010 Jan 26;107 Suppl 1757–64.
174. Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome research. 2009 May;19(5):711–22.
175. Zemojtel T, Kielbasa SM, Arndt PF, Chung H-R, Vingron M. Methylation and deamination of CpGs generate p53-binding sites on a genomic scale. Trends in genetics : TIG. 2009 Feb;25(2):63–6.
176. Scott JM, Weir DG. The methyl folate trap. A physiological response in man to prevent methyl group deficiency in kwashiorkor (methionine deficiency) and an explanation for folic-acid induced exacerbation of subacute combined degeneration in pernicious anaemia. Lancet. 1981 Aug 15;2(8242):337–40.
177. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nature reviews. Genetics. 2008 Jun;9(6):465–76.
178. Hellman A, Chess A. Gene body-specific methylation on the active X chromosome. Science (New York, N.Y.). 2007 Feb 23;315(5815):1141–3.
179. Ekram MB, Kang K, Kim H, Kim J. Retrotransposons as a major source of epigenetic variations in the mammalian genome. Epigenetics : official journal of the DNA Methylation Society. 2012 Apr;7(4):370–82.
180. Molloy AM, Kirke PN, Troendle JF, Burke H, Sutton M, Brody LC, et al. Maternal vitamin B12 status and risk of neural tube defects in a population with high neural tube defect prevalence and no folic Acid fortification. Pediatrics. 2009 Mar;123(3):917–23.
181. Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, et al. The DNA methylome of human peripheral blood mononuclear cells. PLoS biology. 2010 Jan;8(11):e1000533.
182. Rao S, Yajnik CS, Kanade A, Fall CH, Margetts BM, Jackson AA, et al. Intake of micronutrient-rich foods in rural Indian mothers is associated with the size of their babies at birth: Pune Maternal Nutrition Study. The Journal of nutrition. 2001 Apr;131(4):1217–24.
183. Tanaka T, Scheet P, Giusti B, Bandinelli S, Piras MG, Usala G, et al. Genome-wide association study of vitamin B6, vitamin B12, folate, and homocysteine blood concentrations. American journal of human genetics. The American Society of Human Genetics; 2009 Apr;84(4):477–82.
184. Kumar KA, Lalitha A, Pavithra D, Padmavathi IJN, Ganeshan M, Rao KR, et al. Maternal dietary folate and/or vitamin B(12) restrictions alter body composition (adiposity) and lipid metabolism in Wistar rat offspring. The Journal of nutritional biochemistry. 2012 Jun 13;null(null).
185. Smith RM, Osborne-White WS. Folic acid metabolism in vitamin B12-deficient sheep. Depletion of liver folates. The Biochemical journal. 1973 Oct;136(2):279–93.
186. Hussain A, Rahim MA, Azad Khan AK, Ali SMK, Vaaler S. Type 2 diabetes in rural and urban population: diverse prevalence and associated risk factors in Bangladesh. Diabetic medicine : a journal of the British Diabetic Association. 2005 Jul;22(7):931–6.
187. Persson LÅ, Arifeen S, Ekström E-C, Rasmussen KM, Frongillo EA, Yunus M. Effects of prenatal micronutrient and early food supplementation on maternal hemoglobin, birth weight, and infant mortality among children in Bangladesh: the MINIMat randomized trial. JAMA : the journal of the American Medical Association. 2012 May 16;307(19):2050–9.
188. Tobi EW, Slagboom PE, Van Dongen J, Kremer D, Stein AD, Putter H, et al. Prenatal famine and genetic variation are independently and additively associated with DNA methylation at regulatory loci within IGF2/H19. PloS one. 2012 Jan;7(5):e37933.
189. Moore SE, Prentice a M, Wagatsuma Y, Fulford a JC, Collinson a C, Raqib R, et al. Early-life nutritional and environmental determinants of thymic size in infants born in rural Bangladesh. Acta paediatrica (Oslo, Norway : 1992). 2009 Jul;98(7):1168–75.
190. Schläwicke Engström K, Nermell B, Concha G, Strömberg U, Vahter M, Broberg K. Arsenic metabolism is influenced by polymorphisms in genes involved in one-carbon metabolism and reduction reactions. Mutation research. 2009 Jul 10;667(1-2):4–14.

191. Mukherjee S, Das D, Mukherjee M, Das AS, Mitra C. Synergistic effect of folic acid and vitamin B12 in ameliorating arsenic-induced oxidative damage in pancreatic tissue of rat. The Journal of nutritional biochemistry. 2006 May;17(5):319–27.
192. Tsang V, Fry RC, Niculescu MD, Rager JE, Saunders J, Paul DS, et al. The epigenetic effects of a high prenatal folate intake in male mouse fetuses exposed in utero to arsenic. Toxicology and applied pharmacology. 2012 Nov 1;264(3):439–50.
193. Grissa O, Yessoufou A, Mrisak I, Hichami A, Amoussou-Guenou D, Grissa A, et al. Growth factor concentrations and their placental mRNA expression are modulated in gestational diabetes mellitus: possible interactions with macrosomia. BMC pregnancy and childbirth. 2010 Jan;10:7.
194. Ye S, Dhillon S, Ke X, Collins AR, Day IN. An efficient procedure for genotyping single nucleotide polymorphisms. Nucleic acids research. 2001 Sep 1;29(17):E88–8.
195. Franko KL, Forhead AJ, Fowden AL. Differential effects of prenatal stress and glucocorticoid administration on postnatal growth and glucose metabolism in rats. The Journal of endocrinology. 2010 Mar;204(3):319–29.
196. Tang W, Jia L, Ma Y, Xie P, Haywood J, Dawson PA, et al. Ezetimibe restores biliary cholesterol excretion in mice expressing Niemann-Pick C1-Like 1 only in liver. Biochimica et biophysica acta. 2011 Sep;1811(9):549–55.
197. Yang H, Wang JR, Didion JP, Buus RJ, Bell T a, Welsh CE, et al. Subspecific origin and haplotype diversity in the laboratory mouse. Nature genetics. Nature Publishing Group; 2011 May 29;43(7):648–55.
198. Kapoor A, Dunn E, Kostaki A, Andrews MH, Matthews SG. Fetal programming of hypothalamo-pituitary-adrenal function: prenatal stress and glucocorticoids. The Journal of physiology. 2006 Apr 1;572(Pt 1):31–44.
199. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome research. 2010 Jun;20(6):861–73.
200. PEDERSEN J. Weight and length at birth of infants of diabetic mothers. Acta endocrinologica. 1954 Aug;16(4):330–42.
201. Landon MB, Spong CY, Thom E, Carpenter MW, Ramin SM, Casey B, et al. A multicenter, randomized trial of treatment for mild gestational diabetes. The New England journal of medicine. 2009 Oct 1;361(14):1339–48.
202. Krishnaveni G V, Hill JC, Veena SR, Bhat DS, Wills AK, Karat CLS, et al. Low plasma vitamin B12 in pregnancy is associated with gestational “diabesity” and later diabetes. Diabetologia. 2009 Nov;52(11):2350–8.
203. Malhotra A, Kobes S, Knowler WC, Baier LJ, Bogardus C, Hanson RL. A genome-wide association study of BMI in American Indians. Obesity (Silver Spring, Md.). 2011 Oct;19(10):2102–6.
204. Poel YHM, Hummel P, Lips P, Stam F, Van der Ploeg T, Simsek S. Vitamin D and gestational diabetes: a systematic review and meta-analysis. European journal of internal medicine. 2012 Jul;23(5):465–9.
205. Seghieri G, Breschi MC, Anichini R, De Bellis A, Alviggi L, Maida I, et al. Serum homocysteine levels are increased in women with gestational diabetes mellitus. Metabolism: clinical and experimental. 2003 Jun;52(6):720–3.
206. Wheeler S. Assessment and interpretation of micronutrient status during pregnancy. The Proceedings of the Nutrition Society. 2008 Nov;67(4):437–50.
207. Murphy MM, Molloy AM, Ueland PM, Fernandez-Ballart JD, Schneede J, Arija V, et al. Longitudinal Study of the Effect of Pregnancy on Maternal and Fetal Cobalamin Status in Healthy Women and Their Offspring. J. Nutr. 2007 Aug 1;137(8):1863–7.
208. Sayeed MA, Mahtab H, Khanam PA, Begum R, Banu A, Azad Khan AK. Diabetes and hypertension in pregnancy in a rural community of Bangladesh: a population-based study. Diabetic medicine : a journal of the British Diabetic Association. 2005 Sep;22(9):1267–71.
209. Gardosi J, Chang A, Kalyan B, Sahota D, Symonds EM. Customised antenatal growth charts. Lancet. 1992 Feb 1;339(8788):283–7.
210. Obeid R, Jung J, Falk J, Herrmann W, Geisel J, Friesenhahn-Ochs B, et al. Serum vitamin B12 not reflecting vitamin B12 status in patients with type 2 diabetes. Biochimie. 2012 Nov 17;null(null).
211. Yakub M, Moti N, Parveen S, Chaudhry B, Azam I, Iqbal MP. Polymorphisms in MTHFR, MS and CBS genes and homocysteine levels in a Pakistani population. PloS one. 2012 Jan;7(3):e33222.
212. Refsum H, Yajnik CS, Gadkari M, Schneede J, Vollset SE, Orning L, et al. Hyperhomocysteinemia and elevated methylmalonic acid indicate a high prevalence of cobalamin deficiency in Asian Indians. The American journal of clinical nutrition. 2001 Aug;74(2):233–41.

213. Reinstatler L, Qi YP, Williamson RS, Garn J V, Oakley GP. Association of biochemical B_{12} deficiency with metformin therapy and vitamin B_{12} supplements: the National Health and Nutrition Examination Survey, 1999-2006. Diabetes care. 2012 Feb;35(2):327–33.
214. Luka Z, Mudd SH, Wagner C. Glycine N-methyltransferase and regulation of S-adenosylmethionine levels. The Journal of biological chemistry. 2009 Aug 21;284(34):22507–11.
215. Williams KT, Schalinske KL. Tissue-specific alterations of methyl group metabolism with DNA hypermethylation in the Zucker (type 2) diabetic fatty rat. Diabetes/metabolism research and reviews. 2012 Feb;28(2):123–31.
216. Liu S-P, Li Y-S, Chen Y-J, Chiang E-P, Li AF-Y, Lee Y-H, et al. Glycine N-methyltransferase-/- mice develop chronic hepatitis and glycogen storage disease in the liver. Hepatology (Baltimore, Md.). 2007 Nov;46(5):1413–25.
217. Daxinger L, Whitelaw E. Transgenerational epigenetic inheritance: more questions than answers. Genome research. 2010 Dec;20(12):1623–8.
218. Pembrey ME, Bygren LO, Kaati G, Edvinsson S, Northstone K, Sjöström M, et al. Sex-specific, male-line transgenerational responses in humans. European journal of human genetics : EJHG. 2006 Feb;14(2):159–66.
219. Fall C. Maternal nutrition: effects on health in the next generation. The Indian journal of medical research. 2009 Nov;130(5):593–9.
220. Chen P-Y, Feng S, Joo JWJ, Jacobsen SE, Pellegrini M. A comparative analysis of DNA methylation across human embryonic stem cell lines. Genome biology. 2011 Jan;12(7):R62.
221. Gertz J, Varley KE, Reddy TE, Bowling KM, Pauli F, Parker SL, et al. Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. PLoS genetics. 2011 Aug;7(8):e1002228.
222. Saied MH, Marzec J, Khalid S, Smith P, Down TA, Rakyan VK, et al. Genome wide analysis of acute myeloid leukemia reveal leukemia specific methylome and subtype specific hypomethylation of repeats. PloS one. 2012 Jan;7(3):e33213.
223. Feber A, Wilson GA, Zhang L, Presneau N, Idowu B, Down TA, et al. Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors. Genome research. 2011 Apr;21(4):515–24.
224. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nature reviews. Genetics. 2011 Aug;12(8):529–41.
225. Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature. 2011 May 19;473(7347):398–402.
226. Ramagopalan S V, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, Burrell A, et al. A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. Genome research. 2010 Oct;20(10):1352–60.
227. Shu W, Chen H, Bo X, Wang S. Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. Nucleic acids research. 2011 Sep 1;39(17):7428–43.
228. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57–74.
229. Sandovici I, Smith NH, Nitert MD, Ackers-Johnson M, Uribe-Lewis S, Ito Y, et al. Maternal diet and aging alter the epigenetic control of a promoter-enhancer interaction at the Hnf4a gene in rat pancreatic islets. Proceedings of the National Academy of Sciences of the United States of America. 2011 Mar 29;108(13):5449–54.
230. Ferland-McCollough D, Fernandez-Twinn DS, Cannell IG, David H, Warner M, Vaag AA, et al. Programming of adipose tissue miR-483-3p and GDF-3 expression by maternal diet in type 2 diabetes. Cell death and differentiation. Macmillan Publishers Limited; 2012 Jan 6;19(6):1003–12.
231. Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome research. 2010 Jul;20(7):883–9.
232. Jiang X, Yan J, West AA, Perry CA, Malysheva O V, Devapatla S, et al. Maternal choline intake alters the epigenetic state of fetal cortisol-regulating genes in humans. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2012 Aug;26(8):3563–74.
233. Hamadani JD, Tofail F, Nermell B, Gardner R, Shiraji S, Bottai M, et al. Critical windows of exposure for arsenic-associated impairment of cognitive function in pre-school girls and boys: a

population-based cohort study. International journal of epidemiology. 2011 Dec;40(6):1593–604.
234. Wilcox AJ. On the importance--and the unimportance-- of birthweight. International Journal of Epidemiology. 2001 Dec 1;30(6):1233–41.
235. Ong KK, Ahmed ML, Emmett PM, Preece MA, Dunger DB. Association between postnatal catch-up growth and obesity in childhood: prospective cohort study. BMJ (Clinical research ed.). 2000 Apr 8;320(7240):967–71.
236. Karlsson B, Knutsson A, Lindahl B. Is there an association between shift work and having a metabolic syndrome? Results from a population based study of 27,485 people. Occupational and environmental medicine. 2001 Nov;58(11):747–52.
237. Buxton OM, Cain SW, O'Connor SP, Porter JH, Duffy JF, Wang W, et al. Adverse metabolic consequences in humans of prolonged sleep restriction combined with circadian disruption. Science translational medicine. 2012 Apr 11;4(129):129ra43.
238. De Kreutzenberg SV, Ceolotto G, Papparella I, Bortoluzzi A, Semplicini A, Dalla Man C, et al. Downregulation of the longevity-associated protein sirtuin 1 in insulin resistance and metabolic syndrome: potential biochemical mechanisms. Diabetes. 2010 Apr;59(4):1006–15.
239. Sun C, Zhang F, Ge X, Yan T, Chen X, Shi X, et al. SIRT1 improves insulin sensitivity under insulin-resistant conditions by repressing PTP1B. Cell metabolism. 2007 Oct;6(4):307–19.
240. Caton PW, Nayuni NK, Kieswich J, Khan NQ, Yaqoob MM, Corder R. Metformin suppresses hepatic gluconeogenesis through induction of SIRT1 and GCN5. The Journal of endocrinology. 2010 Apr;205(1):97–106.
241. Cantó C, Gerhart-Hines Z, Feige JN, Lagouge M, Noriega L, Milne JC, et al. AMPK regulates energy expenditure by modulating NAD+ metabolism and SIRT1 activity. Nature. 2009 Apr 23;458(7241):1056–60.
242. Chauhan G, Spurgeon CJ, Tabassum R, Bhaskar S, Kulkarni SR, Mahajan A, et al. Impact of common variants of PPARG, KCNJ11, TCF7L2, SLC30A8, HHEX, CDKN2A, IGF2BP2, and CDKAL1 on the risk of type 2 diabetes in 5,164 Indians. Diabetes. 2010 Aug 1;59(8):2068–74.
243. Yajnik CS, Janipalli CS, Bhaskar S, Kulkarni SR, Freathy RM, Prakash S, et al. FTO gene variants are strongly associated with type 2 diabetes in South Asian Indians. Diabetologia. 2009 Feb;52(2):247–52.
244. Chauhan G, Kaur I, Tabassum R, Dwivedi OP, Ghosh S, Tandon N, et al. Common variants of homocysteine metabolism pathway genes and risk of type 2 diabetes and related traits in Indians. Experimental diabetes research. 2012 Jan;2012:960318.
245. Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, et al. Genetic evidence on the origins of Indian caste populations. Genome research. 2001 Jun;11(6):994–1004.
246. Veena SR, Krishnaveni G V, Fall CH. Newborn size and body composition as predictors of insulin resistance and diabetes in the parents: Parthenon Birth Cohort Study, Mysore, India. Diabetes care. 2012 Sep;35(9):1884–90.
247. Carone BR, Fauquier L, Habib N, Shea JM, Hart CE, Li R, et al. Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. Cell. Elsevier Inc.; 2010 Dec 23;143(7):1084–96.
248. Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, et al. The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. Molecular cell. 2012 Dec 28;48(6):849–62.
249. Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, et al. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature. 2010 Feb 25;463(7284):1101–5.
250. Guibert S, Forné T, Weber M. Global profiling of DNA methylation erasure in mouse primordial germ cells. Genome research. 2012 Apr;22(4):633–41.
251. Song S. Does famine influence sex ratio at birth? Evidence from the 1959-1961 Great Leap Forward Famine in China. Proceedings. Biological sciences / The Royal Society. 2012 Jul 22;279(1739):2883–90.

Publications and Presentations Arising From This Work

Bell CG, **Finer S** **(*Joint first author*)**, Lindgren CM, Wilson GA, Rakyan VK, Teschendorff AE, Akan P, Stupka E, Down TA, Prokopenko I, Morison IM, Mill J, Pidsley R; International Type 2 Diabetes 1q Consortium, Deloukas P, Frayling TM, Hattersley AT,McCarthy MI, Beck S, Hitman GA (2010). Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS One* 18;5(11)

Finer S, Holland M, Nanty L, Rakyan V (2011). The Hunt for the Epiallele. *Environmental and Molecular Mutagenesis*; 52(1)1-11 (Review)

Finer S, Iqbal S, Mathews C, Smart M, Cravioto A, Alam D, Hitman G. Epigenetic variants are detectable in young adult offspring exposed to famine in early development. Keystone Symposium: Nutrition, Epigenetics and Human Disease (February 2013). Accepted for oral presentation

Finer S, Mathews C, Smart M, Holland M, Rakyan V, Hitman GA (2012). A model of intrauterine hyperglycaemia provides evidence for fetal programming of glucose tolerance and a platform for epigenomic studies in mouse and human. Diabetes UK Annual Professional Conference, poster presentation. ***Shortlisted for Lilly Basic Science Prize***.

Finer S, Mathews C, Holland M, Rakyan V, Hitman GA (2011). A model of intrauterine hyperglycaemia provides evidence for fetal programming of glucose tolerance and a platform for epigenomic studies in mouse and human. Wellcome Trust meeting, Epigenomics of Common Diseases, poster presentation.

Finer S, Rakyan V, Hitman G (2011). Epigenomic studies in Type 2 diabetes. Rank Nutrition Symposium, oral presentation.

Finer S, Holland M, Rakyan V, Hitman G (2010). A model of intrauterine hyperglycaemia provides evidence for foetal programming of glucose tolerance and a platform for epigenomic studies in mouse and human. 5th Conference of Epidemiological Longitudonal Studies in Europe, Cyprus. Oral presentation.

Finer S, Bell C, Lindgren C, Wilson G, Rakyan V, Prokopenko I, Teschendorff A, Down T, Heap G, Plagnol V, van Heel D, Akan P, Deloukas P, Hattersley A, International T2D 1q Consortium, Hitman GA, Beck S, McCarthy M.(2010) The *FTO* obesity and Type 2 Diabetes susceptibility haplotype is associated with SNP-dependent DNA methylation. Diabetes UK APC. Oral presentation.

Finer S, Bell C, Lindgren C, Wilson G, Rakyan V, Prokopenko I, Teschendorff A, Down T, Heap G, Plagnol V, van Heel D, Mill J, Pidsley R, Akan P, Deloukas P, Hattersley A, International T2D 1q Consortium, Hitman G, Beck S, McCarthy M.(2010) Targeted epigenomic studies of Type 2 Diabetes susceptibility haplotypes. Society for Endocrinology BES 2010. Oral presentation