Bioinformatic Approaches to Analysing RRBS Data in the Context of Pre-eclampsia

Hester Emily Roberts
Department of Pathology
University of Otago

A thesis submitted for the degree of

Doctor of Philosophy
02 March 2015
“There is a theory which states that if ever anyone discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable. There is another theory which states that this has already happened.”

- Douglas Adams, The Restaurant at the End of the Universe
Abstract

The placenta is a crucial element of human life. It sustains and enables the development from a small group of cells to a whole organism. The placenta is a unique organ that fulfils multiple roles, from delivery of oxygen and nutrients, to removal of waste products. During embryo implantation, the trophoblast cells on the outside of the blastula invade the maternal endometrium and contact maternal spiral arteries. These trophoblast cells extensively remodel the maternal spiral arteries, establishing a blood supply to the growing embryo. On occasion, elements of this process malfunction, resulting in placental pathology. Pre-eclampsia is a serious and unfortunately common disease that arises only during pregnancy. The main clinical hallmarks of pre-eclampsia are the new onset of maternal hypertension, and proteinuria. These symptoms arise only in the second half of pregnancy, but the pathogenesis lies much earlier in pregnancy. Pre-eclamptic placentas display reduced invasion of the trophoblast, and restriction of maternal spiral arteries, in turn restricting blood flow to the fetus, and causing a hypoxic state.

Epigenetic modifications in the placenta are distinct from those in any somatic tissue. Hypomethylation across the genome, and in particular at promoter regions, imprinted genes, and repetitive regions within the genome of placental cells appears to play a crucial role in the development of the placenta. Therefore, I have hypothesised that alterations to the methylation of placental cells may contribute to aberrant placentation. This thesis demonstrates the first use of Reduced Representation Bisulphite Sequencing to investigate genome-wide DNA methylation in pre-eclampsia/toxaemia. In the course of this investigation, significant progress has been achieved in our understanding of data from Reduced Representation Bisulphite Sequencing, contributing to the bioinformatic and biostatistic pipeline for the analysis of this data. A large number of CpG sites have been identified in
a matched cohort of pre-eclamptic placentas to show differential methylation when compared to a cohort of control placentas. The majority of these sites display hypomethylation in pre-eclamptic placentas, and globally, CpG sites from pre-eclamptic placentas show 2% less methylation than their control counterparts.

In addition, an investigation has been undertaken into the methylation status of two genes that have shown evidence of differential methylation. This particular study was undertaken in human blood to establish a baseline of normal methylation in a normal, accessible, and well phenotypes tissue. These results will contribute towards our understanding of DNA methylation in both normal and diseased states, with particular emphasis on the role that methylation plays in placental development.
Acknowledgements

There are a large number of people to whom I owe thanks. Without them, this PhD would never have been completed. Firstly, to my supervisors, Professor Ian Morison, Dr. Erin Macaulay, and Dr. Noelyn Hung. You have given me much in terms of your time, your inspiration, and your wisdom. Without Erin in particular this project would never have happened, and her indefatigable positive attitude and commitment are truly an inspiration. Special thanks go to the Morison Lab group, and in particular Rob Weeks and Jackie Ludgate for their invaluable technical expertise, and Dr. Helena McGrath-Cohen for the amazing laughs. Dr. Euan Rodger and Dr. Aniruddha Chatterjee did an amazing amount of groundwork to get RRBS analysis to the point where I could start using it, and always had useful inputs. Ingrid Knarston and Suzan Momani helped me a lot with all of the Sequenom work and gave me a great opportunity to both teach and learn. To the rest of the Morison lab members, both past and current I owe a great thanks for making work an enjoyable place to be - Abdul Alsaleh, Gwenn LeMee, Tim Foster, Fabian Isensee, and the members of the West Wing lab.

To Dr. Matthew Parry, thank you for your explanations of the world of statistics. I didn’t always understand it the first time, but you always found a way to make it fit into this biologist’s brain. Tanya Flynn gave me admirable tuition in using R, and made it much less scary.

I owe a great thanks to all of the women who donated their placentas to the OPUSS collection, I can only hope that the work I have done can go some-way towards returning that amazing gift. Likewise, all of the midwives and clinicians involved in that collection, particularly Dr. Tania Slatter and Dr. Celia Devenish.

For the funding I have received, I would like to thank firstly the University of Otago for the Master’s Scholarship that subsequently became a PhD
Scholarship when I decided I wasn’t ready to move on from my project just yet. Gravida: the National Research Centre for Growth and Development funded this research into the epigenetics of the placenta, and provided opportunities to interact with other amazing scientists in the field of early life growth and development. The wonderful people in the Department of Pathology were very supportive when things didn’t go to plan, and were there when I needed it, with expertise, resources, and a hug if need be. The International Federation of Placenta Associates and Genetics Otago kindly gave me funding to attend a conference on the placenta in Canada, for which I am extremely grateful.

On a personal note, I must thank my friends for keeping me (relatively!) sane throughout. In particular Tessa and Aleisha, you have given me true friendship for which I am exceedingly grateful. To my flatmates, for putting up with my eccentricities, thank you. To the members of the Horsfield Lab who generously welcomed me into their office in the final stages of my PhD, thank you. To my brother Tom, I’m sorry but I won’t be doing another PhD, so you’ll just have to do your own one to keep up the family tradition. Finally to my wonderful parents go the biggest thanks of all. You fostered curiosity in a 3 year old me who constantly asked why, and hasn’t stopped since. You have been so supportive and kind, and set a wonderful example, not just as parents, but as people whom I am honoured to call family.
Abbreviations

.csv ................ comma separated values text file format
A .................. Adenine
AB ................. Antibiotic
ACR ............... Albumin/Creatinine Ratio
ALT ............... Alanine Transaminase
ASM .............. Allele-Specific Methylation
BAM .............. Compressed binary SAM file
BH ............... Benjamini-Hochberg adjustment for multiple testing
BISMA ........... Bisulfite Sequencing DNA Methylation Analysis
BLAST ........... Basic Local Alignment Search Tool
BMI .............. Body Mass Index
bp ............... base pairs
BP ............... Blood Pressure
BSA .............. Bovine Serum Albumin
C .................. Cytosine
C/S ................ Caesarian Section
cffDNA ........... cell-free fetal DNA
CGI .............. CpG Island
CHG ................ Cytosine followed by non-guanine base, followed by guanine
CHH .............. Cytosine followed by two non-guanine bases
chr .............. chromosome
CI ............... Confidence Interval
COBRA .......... Combined Bisulphite Restriction Analysis
CpG .............. Cytosine phosphate Guanine dinucleotide
CRH ............... Corticotrophin Releasing Hormone
CV .................. Chorionic Villus
DMAP ............... Differential Methylation Analysis Package
DMR ............... Differentially Methylated Region
DMSO ............... Dimethyl Sulfoxide
DNA/RNA ........... Deoxy-/Ribonucleic Acid
dNTP ............... deoxyribose nucleotide triphosphate
EDTA ............... Ethylenediaminetetraacetic
EM C/S ............. Emergency Cesarian Section
ePCR ............... Electronic PCR
EST ............... Expressed Sequence Tag
FDR ............... False Discovery Rate
G .................. Guanine
GC .................. Guanine and Cytosine
GGTP ............... Gamma-glutamyl transpeptidase
GOI ............... Gain of Imprinting
GWAS ............. Genome-Wide Association Scan
H&E ............... Haematoxylin and eosin histology stain
HOMER ............ Hypergeometric Optimisation of Motif EnRichment
HS ............... High Sensitivity
ICM ............... Inner Cell Mass
ICR ............... Imprint Control Region
IDDM ............. Insulin-Dependent Diabetes Mellitus
IOL ............... Induction of Labour
IUGR ............. Intrauterine Growth Restriction
J Biomed Biotechnol  Journal of Biomedicine and Biotechnology

LB ................. Luria-Bertani medium

LFT ................. Liver Function Test

LOI ................. Loss of Imprinting

MALDI-TOF ...... Matrix-Assisted Laser Desorption/Ionisation Time-Of-Flight

MeDIP ............. Methylated DNA immunoprecipitation

MgSO$_4$ .......... Magnesium Sulphate

MQH$_2$O .......... MilliQ purified water

NAR ................. Nucleic Acids Research

NCD ................. Non-Communicable Disease

ncRNA ............. non-coding RNA

NVD/SVD .......... Normal/Spontaneous Vaginal Delivery

NZGL ............... New Zealand Genomics Limited

OPuS ............... Otago Placental Study

PBS ................. Phosphate Buffered Saline

PCR ................. Polymerase Chain Reaction

PE ................. Pre-eclampsia

PET ................. Pre-eclampsia/Toxaemia

PGC ................. Primordial Germ Cell

RFLP ............... Restriction Fragment Length Polymorphism

RRBS ............... Reduced Representation Bisulphite Sequencing

RSB ................. Resuspension Buffer

Rx ................. Medical Prescription

SAM ................. Sequence Alignment/Map file format

SEM ................. Standard Error of the Mean
SLIM .......... Sliding Linear Model
SNP .......... Single Nucleotide Polymorphism
SOMANZ ....... Society of Obstetric Medicine of Australia and New Zealand
SRS .......... Silver-Russell Syndrome
STL .......... Stop Ligation Mix
T ............. Thymine
TBE .......... Tris/Borate/EDTA buffer
TGS .......... Thyroid Genetics Study
TP ............ Tenney-Parker changes
TpG .......... Thymine phosphate Guanine dinucleotide
TSC .......... Trophoblastic Stem Cell
TSG .......... Tumour Suppressor Gene
TSS .......... Transcription Start Site
U ............. Uracil
UC .......... Umbilical Cord
UPD .......... Uniparental Disomy
UTR .......... Untranslated Region
UV .......... Ultraviolet light
VSM .......... Vasculo-syncytial Membrane
VUE .......... Villitus of Unknown Etiology
Xi .......... Inactive X chromosome
Y .......... Pyrimidine (Cytosine or Thymine)
## Contents

1 **Background** ................................................................. 1
1.1 The Placenta .............................................................. 1
  1.1.1 Placental Structure and Function .......................... 2
  1.1.2 Placental Development ...................................... 2
  1.1.3 Placental Plasticity ............................................ 5
1.2 Epigenetics .............................................................. 5
  1.2.1 DNA Methylation .............................................. 6
  1.2.2 Other Mechanisms of Epigenetic Control ................. 8
1.3 Imprinting ............................................................... 9
  1.3.1 Theories for Existence of Imprinting ..................... 10
  1.3.2 Primary Imprinting Control ................................. 12
  1.3.3 Secondary Imprints - Promoter Regulation ............ 14
  1.3.4 Allele Specific Methylation and Parent-of-Origin
          Effects .......................................................... 15
  1.3.5 Methods for Identifying Differentially Methylated
          Regions ......................................................... 16
  1.3.6 Dysregulation of Genomic Imprinting ................... 22
1.4 Pre-eclampsia/Toxaemia ............................................. 23
  1.4.1 Molecular Mechanisms of Pre-eclampsia/Toxaemia .... 24
  1.4.2 Genetics of PET ............................................... 26
  1.4.3 Epigenetics of PET ........................................... 27
1.5 Candidate Genes for Differential Methylation ................. 29
  1.5.1 IGFBP1 .......................................................... 30
  1.5.2 ZNF264 .......................................................... 32
1.6 Aims ..................................................................... 34
  1.6.1 Hypotheses ....................................................... 35
2 Materials and Methods

2.1 Sourcing of Samples

2.1.1 Ethical Approval

2.1.2 Collection of Blood Samples

2.1.3 Processing of Placental Samples

2.2 DNA Extraction

2.2.1 Peripheral Blood

2.2.2 Placental Tissue

2.2.3 Quantification of DNA

2.3 Genotyping of Blood Samples

2.3.1 Genotyping via SNP Analysis

2.3.2 Sanger Sequencing

2.4 Methylation of Candidate Imprinted Genes

2.4.1 Bisulphite Conversion

2.4.2 PCR Amplification of Bisulphite Treated DNA

2.4.3 Cloning of PCR Products for Methylation Analysis

2.4.4 Analysis of Clones

2.4.5 Methylation Analysis of Clones

2.5 Collection and Phenotyping of Placental Samples

2.5.1 Cohort Recruitment

2.5.2 Ethical Approval

2.5.3 Placental Pathology

2.6 Pre-eclampsia Characterisation

2.6.1 Blood Pressure and Proteinuria

2.6.2 Histology

2.7 Matching of Controls

2.7.1 Gestation

2.7.2 Fetal Sex

2.7.3 Maternal Smoking

2.7.4 Maternal BMI

2.7.5 Additional Cohort Characterisation

2.8 TruSeq Reduced Representation Bisulphite Sequencing Library Preparation

2.8.1 MspI Digestion
2.8.2 End Repair ................................................. 68
2.8.3 Adenylation of 3’ Ends .................................. 68
2.8.4 Adaptor Ligation ....................................... 69
2.8.5 AMPure Bead Purification ................................. 69
2.8.6 First Round of Size Selection ............................ 70
2.8.7 Bisulphite Conversion of RRBS Libraries .............. 72
2.8.8 Semi-Quantitative PCR .................................. 72
2.8.9 Large-Scale PCR Amplification of Libraries ............ 74
2.8.10 Round Two of Size Selection .......................... 74
2.8.11 Preparation for Multiplexing and Sequencing .......... 75
2.9 Sequentom MassArray® EpiTYPER® Methylation Validation ............................................. 77
2.9.1 Selection of Candidate Genes from RRBS for Sequentom .... 78
2.9.2 Candidate Gene Primer Design ............................ 78
2.9.3 Bisulphite Conversion of Placental DNA ................. 79
2.9.4 PCR Optimisation .......................................... 80
2.9.5 PCR Conditions ........................................... 82
2.9.6 Sequentom MALDI-TOF Reaction ......................... 82
2.9.7 Sequentom Data Analysis .................................. 83

3 Epigenetic Analysis of Two Putative Imprinted Genes - IGFBPI and ZNF264 84
3.1 Sample Information .......................................... 85
3.1.1 DNA Extraction ........................................... 85
3.2 Genotyping for SNPs ......................................... 87
3.2.1 Restriction Fragment Length Polymorphism Analysis .. 90
3.2.2 Genotyping by Sequencing ................................. 90
3.3 PCR of Bisulphite-Converted Informative Samples ......... 93
3.4 Cloning ....................................................... 95
3.4.1 Screening of Clones for Correct Inserts ................. 95
3.4.2 Plasmid Extraction and Sequencing ....................... 96
3.5 Analysis of Clones .......................................... 97
3.5.1 Allele-Specific Methylation ............................... 99
3.5.2 Parent-of-Origin Effects ................................. 109
3.6 Summary of Results ....................................... 114
### Computational Techniques and Tools for RRBS Analysis

#### 4.1 Methods for Data Processing
- **4.1.1 FastQC**
- **4.1.2 Fastx Toolkit**
- **4.1.3 Mapping to Reference Genome**
- **4.1.4 Extraction of Methylation Values**

#### 4.2 Statistical Testing of Individual CpGs
- **4.2.1 Regression and Modelling of Data**
- **4.2.2 Logistic Regression**
- **4.2.3 MethylKit**

#### 4.3 Identifying Candidate Differentially Methylated Sites
- **4.3.1 Calculation of Methylation Differences**
- **4.3.2 Annotation of Differences**

#### 4.4 Summary of Analyses

### Results of RRBS Analysis

#### 5.1 Library Composition

#### 5.2 Testing for Statistically Significant CpG Sites
- **5.2.1 MethylKit**
- **5.2.2 Quasibinomial**
- **5.2.3 Random Effects Model**

#### 5.3 Differentially Methylated Sites

#### 5.4 Significantly Different Sites
- **5.4.1 MethylKit**
- **5.4.2 Quasibinomial**
- **5.4.3 Random Effects Model**

#### 5.5 Selecting Candidates
- **5.5.1 Examining Top Hits In Detail**

#### 5.6 Validation of RRBS
- **5.6.1 Sequenom MassArray®Epityper®**

### Discussion

#### 6.1 Overview of Main Findings
- **6.1.1 Pre-eclamptic Placentas Are Consistently Globally Hypomethylated**
List of Figures

1.1 Implantation and Development of the Placenta from the Early Trophoblast .............................................. 3
1.2 Invasion of the trophoblast during early gestation. ................. 4
1.3 Reprogramming and Maintenance of Imprint Control Regions (ICR) in mammals. ........................................ 13
1.4 The current view of promoter methylation ............................... 14
1.5 Products Generated From Bisulfite Conversion and PCR ............ 17
1.6 RRBS Library Preparation Workflow ................................. 20
1.7 Molecular Construction of RRBS Libraries ............................ 21
1.8 Remodelling of spiral arteries in pregnancy .......................... 25
1.9 Splice variants of ZNF264 ............................................. 33
1.10 Hypotheses ............................................................ 35
2.1 Placenta Cohort Matching .............................................. 65
2.2 Modes of Delivery in Placenta Cohort .................................. 66
3.1 Amplification of genomic DNA for genotyping purposes .......... 88
3.2 IGFBP1 and ZNF264 Gene Regions ...................................... 89
3.3 RFLP analysis of samples X9015-X9030 .............................. 91
3.4 Genotyping via Sequencing ............................................. 91
3.5 PCRs of Bisulphite converted DNA .................................... 95
3.6 EcoRI Digest of Plasmids ................................................ 96
3.7 PCR screen of E.coli clones ............................................. 97
3.8 Calculation of Methylation ............................................. 99
3.9 Allele-Specific Methylation in the Promoters of IGFBP1 and ZNF264 101
3.10 Allele-Specific Site-by-Site Methylation in the Promoters of *IGFBP1* and *ZNF264* ........................................... 102
3.11 Allele-Specific Methylation in Individual Samples for the Promoter of *IGFBP1* ...................................................... 103
3.12 Site-by-Site Analysis of Individuals Informative for the *IGFBP1* Promoter ................................................................. 104
3.13 Allele-Specific Methylation in Individual Samples for the Promoter of *ZNF264* ............................................................. 105
3.14 Site-by-Site Analysis of Individuals Informative for the *ZNF264* Promoter ................................................................. 107
3.15 Methylation in Exon 1 of *ZNF264* and *IGFBP1* ............... 108
3.16 Methylation in Exon 1 of *IGFBP1* and *ZNF264* ............... 110
3.17 Informative Pedigrees for the Promoter of *IGFBP1* .......... 110
3.18 Mean Amplicon Methylation and Mean CpG Site Methylation in the Promoter of *IGFBP1* - Parent-of-Origin ............... 111
3.19 Parent-of-Origin Specific Methylation in Informative Individuals ............................... 113
3.20 Parent-of-Origin, Site-by-Site Analysis of Individuals Informative for the *IGFBP1* Promoter ........................................... 114

4.1 Pipeline for Analysis of RRBS data .......................................... 118
4.2 FastQC Per Base Quality ..................................................... 121
4.3 FastQC Per Sequence Quality ............................................. 122
4.4 FastQC Per Base Sequence Content ..................................... 123
4.5 FastQC Per Base GC Content ............................................. 124
4.6 FastQC Per Sequence GC Content ....................................... 125
4.7 FastQC Sequence Length Distribution ................................ 126
4.8 Overrepresented sequences ............................................. 126
4.9 Quality Scoring of Individual Reads ................................... 128
4.10 PCR Products from a Bisulphite Conversion Reaction .......... 131
4.11 Statistical Decisions for Calculating Significant Differences .... 138

5.1 Methylation of Placentas in RRBS Libraries ......................... 150
5.2 Significance Values for CpG Sites from MethylKit Analysis .... 152
5.3 Significance Values for CpG Sites From Quasibinomial Modelling Analysis .................. 154
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4</td>
<td>Significance Values for CpG Sites from Random Effects Modelling Analysis</td>
<td>156</td>
</tr>
<tr>
<td>5.5</td>
<td>Difference in Methylation Between Cases and Controls, Weighted by Coverage</td>
<td>158</td>
</tr>
<tr>
<td>5.6</td>
<td>Difference in Methylation Between Cases and Controls, Arithmetic Means</td>
<td>158</td>
</tr>
<tr>
<td>5.7</td>
<td>MethylKit Significantly Different Sites</td>
<td>160</td>
</tr>
<tr>
<td>5.8</td>
<td>Quasibinomial Significantly Different Sites</td>
<td>162</td>
</tr>
<tr>
<td>5.9</td>
<td>Random Effects Model Significantly Different Sites</td>
<td>164</td>
</tr>
<tr>
<td>5.10</td>
<td>Overlapping Sites Between All Three Approaches</td>
<td>166</td>
</tr>
<tr>
<td>5.11</td>
<td>Top 25 Candidate Sites (Part 1)</td>
<td>167</td>
</tr>
<tr>
<td>5.12</td>
<td>Top 25 Candidate Sites (Part 2)</td>
<td>168</td>
</tr>
<tr>
<td>5.13</td>
<td>Top 25 Candidate Sites (Part 2)</td>
<td>169</td>
</tr>
<tr>
<td>5.14</td>
<td>Sequenom® Validation of RRBS Methylation (Part 1)</td>
<td>173</td>
</tr>
<tr>
<td>5.15</td>
<td>Sequenom® Validation of RRBS Methylation (Part 2)</td>
<td>174</td>
</tr>
<tr>
<td>5.16</td>
<td>Sequenom® Validation of RRBS Methylation (Part 3)</td>
<td>175</td>
</tr>
<tr>
<td>C.1</td>
<td>Number of Reads and Coarse Methylation in RRBS Libraries</td>
<td>203</td>
</tr>
</tbody>
</table>
**List of Tables**

1.1 Methods for Studying Methylation ........................................ 19

2.1 SNPs used for genotyping samples ...................................... 40

2.2 Primer sequences for genotyping PCRs ................................. 41

2.3 PCR reaction mixes for genotyping PCRs ............................... 41

2.4 PCR Conditions for genotyping .......................................... 42

2.5 Restriction digest conditions for genotyping via RFLP ............. 44

2.6 BigDye® reaction mix conditions ....................................... 45

2.7 BigDye® Cycling Conditions ............................................. 45

2.8 Primers for amplification of bisulphite converted DNA .......... 48

2.9 PCR mixes for bisulphite converted DNA .............................. 49

2.10 Thermocycler conditions for PCR of bisulphite converted DNA . 49

2.11 PCR reaction for screening of *E.coli* clones for inserts .......... 51

2.12 EcoRI digestion of plasmids to screen for correct inserts ....... 52

2.13 Clinical Characteristics of Pre-eclamptic Placentas ................ 56

2.14 Gross and Histological Pathology of Placental Samples ........... 58

2.15 Matched Samples Used In This Study ................................. 62

2.16 Examplar Placental Sample ............................................ 67

2.17 MspI restriction endonuclease digest ................................ 67

2.18 Semi-quantitative PCR reaction (2X) ................................. 73

2.19 Cycling condition for Semi-quantitative PCRs ..................... 73

2.20 Large-Scale PCR Reaction ............................................. 74

2.21 Sequenom PCR amplicons ............................................. 81

2.22 Sequenom PCR reactions ............................................. 82

2.23 Cycling Conditions for Sequenom PCRs ............................... 83
3.1 List of Samples and Concentration of DNA Extracted From Whole Blood 86
3.2 SNP Genotyping . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
3.3 Informative Samples . . . . . . . . . . . . . . . . . . . . . . . . . . 94
3.4 Clones generated for the promoter regions of IGFBP1 and ZNF264 . 98

5.1 Summary of Library Content . . . . . . . . . . . . . . . . . . . . . 148
5.2 HOMER Annotation of Top 25 Differentially Methylated CpG Sites . 171
Chapter 1

Background

This thesis represents the coming together of multiple different elements regarding the study of epigenetics, with an emphasis on DNA methylation in the placenta. As we strive to understand the complexity of life, epigenetics has become an increasingly important element in elucidating how the environment interacts with the genome to produce different cell types, developmental trajectories, and disease states. To understand what happens when epigenetic regulation becomes disrupted, we must first discover what the 'normal' state is. To this end, one portion of this thesis is an investigation of the normal methylation state of two genes which have shown evidence of variable regulation with DNA methylation. The remainder, and bulk of this thesis is concerned with investigating DNA methylation on a wider scale in a particular disease state. Pre-eclampsia is a placental disorder, the cause of which is still largely unknown. This project undertakes a genome-wide analysis of DNA methylation in a cohort of pre-eclamptic and control placentas. The method used for this study (RRBS) is still in development with respect to its analysis. As such, considerable care has been taken here to perform appropriate analyses, and a large portion of this thesis is devoted to these analysis methods.

1.1 The Placenta

No one would be here today without the organ that made our early lives possible: the placenta. It facilitates the transfer of nutrients to the fetus, the removal of waste products, and the mediation of gas exchange. The placenta acts as an immunological
barrier, protecting and disguising the growing fetus from the mother’s immune system. Via a network of hormonal and small molecule signalling, the placenta regulates fetal growth and polices the maternal influence. The placenta acts as a biological life support machine to every developing fetus. In addition to these functions, it also allows the physical attachment of the embryo to the maternal uterine wall. Placental signalling also plays a crucial role in the initiation of labour. Indeed, without the placental secretion of Corticotrophin Releasing Hormone (CRH) and Progesterone, the positive feedback loop required for the initiation and maintenance of labour would be severely affected (Denda et al., 2011; Vrachnis et al., 2012).

1.1.1 Placental Structure and Function

The placenta is a highly versatile and adaptable organ that mediates the fetal-maternal interaction. Most mammals share the basic hallmarks of placentation, that is, the exchange of fetal and maternal elements via a chorioallantoic placenta. The placenta of eutherian mammals is the distinguishing feature that sets these organisms apart from other viviparous species, and it is characterised by a unique extra-embryonic cell lineage. The trophoblast is a cell type of the placenta that is derived from the polar trophectoderm and it plays a key role in the interaction between the maternal and fetal components in pregnancy (Lefebvre, 2012).

The placentas of humans and ‘higher’ apes such as gorillas, orang-utans and chimpanzees are distinct from those of other mammals, most notably in their invasiveness. This invasiveness results in the maternal blood being in direct connection with the placental membrane. This characteristic lends its name to these types of placenta: haemochorial. The fact that these types of placenta also appear in rabbits and some rodents have led some to suggest that the haemochorial placenta is the ancestral state for eutherian mammals. This viewpoint is also supported by phylogenetic analyses into the mammalian placenta (Wildman, 2011; Wildman et al., 2006).

1.1.2 Placental Development

Early polarity in the embryo dictates which cells will become the trophoblast and thence the placenta (Edwards & Hansis, 2005), and which cells will form the embryo. The cells that will form the germ layers and thus the new embryo aggregate into an Inner Cell Mass (ICM). The cells that form the outer layer of the blastocyst
and surround the ICM and blastocoele become the trophoblast, and will further develop into the syncytiotrophoblast and cytotrophoblast after gastrulation has occurred (Figure 1.1). Implantation of the embryo begins early after conception. Between 6 and 12 days post-fertilisation, the blastocyst attaches to the myometrial wall, and the trophoblast begins to proliferate and invade into the maternal tissue. This process will eventually lead to the development of the placenta, and the remodelling of the mother’s blood vessels to ensure a sufficient blood supply to the fetus (Cross et al., 1994).

![Diagram showing implantation and development of the placenta.](image)

**Figure 1.1:** Implantation and development of the placenta from the early trophoblast invading the myometrium at 5-6 days to 3 weeks gestation.

### 1.1.2.1 Implantation and Trophoblast Invasion

During normal pregnancy, the blastocyst adheres to the uterine endometrium and makes direct contact with the maternal blood supply. Immediately after implantation, trophodermal cells forming the outermost epithelial layer of the blastocyst give rise to multiple trophoblast cell types through fusion, proliferative, and invasive pathways (Cross et al., 1994; Hamilton & Boyd, 1960). At the site of implantation, cell fusion generates the primitive syncytium, which may represent the earliest invasive cell type. Cytotrophoblasts (CTBs) are then produced by the trophoderm, and begin to form primary villi by proliferation and invasion through the primitive syncytium (Knöfler & Pollheimer, 2013). Throughout pregnancy, these primary villi transform into secondary and tertiary villi that are characterised by invasion of extraembryonic mesenchymal cells, villous branching, and vascularization. These villi will form the fetomaternal interface throughout pregnancy. After contact with the maternal blood supply is established, nutrients and oxygen cross the epithelial layers of these villi enabling fetal growth and development. The multinucleated syncytiotrophoblast covers the surface
of the villi, and is in direct contact with maternal blood. As such, it plays an important part in mediating nutrient and gas exchange. As a terminally differentiated cell type, this cell population is sustained by the continuing differentiation and fusion of the underlying cytotrophoblast layer (Aplin, 2010).

Oxygen perfusion plays an important role in the differentiation process that leads to trophoblast invasion of the uterus. The early stages of placental development take place in a relatively hypoxic environment that favours cytotrophoblast proliferation rather than differentiation along the invasive pathway (Red-Horse et al., 2004). It is thought that this may be to protect against reactive oxygen species that can alter the epigenetic regulation of genes (Patterson et al., 2012). Interestingly, the incidence of the placental disease pre-eclampsia/toxaemia (PE or PET) is increased at high altitudes (Palmer et al., 1999), providing corroboration to the hypothesis that decreased placental perfusion is one of the mechanisms behind the development of pre-eclamptic signs. Pre-eclampsia is a major focus of this project, and is discussed in more detail in Section 1.4.

Figure 1.2: Invasion of the trophoblast during early gestation. Oxygen tension plays an important role in guiding the differentiation process that leads to trophoblast invasion of the uterus. Adapted with permission from Red-Horse et al. (2004)

The trophoblast cell population rapidly increases in number compared with the embryonic lineages. As development continues, cytotrophoblasts invade the uterine wall (Figure 1.2A) and plug the maternal vessels, a process that helps maintain a state
of physiological hypoxia (Figure 1.2B). Cytotrophoblasts migrate farther up arteries than veins. By 10 to 12 weeks of human pregnancy, blood flow to the intervillous space begins. As the endovascular component of cytrophoblast invasion progresses, the cells migrate along the lumina of spiral arterioles, replacing the maternal endothelial lining. Cytotrophoblasts also replace the smooth muscle walls of these vessels. In normal pregnancy the process whereby placental cells remodel uterine arterioles involves the decidual and inner third of the myometrial portions of these vessels. As a result, the diameter of the arterioles expands to accommodate the dramatic increase in blood flow that is needed to support rapid fetal growth later in pregnancy (Figure 1.2C). It is likely that failed invasion of the maternal blood vessels leads, in some cases, to abortion, whereas an inability to invade to the appropriate depth is associated with pre-eclampsia and a subset of pregnancies in which the growth of the fetus is restricted (Red-Horse et al., 2004).

1.1.3 Placental Plasticity

In comparison to many other organs, the placenta is a highly plastic structure. It is capable of adapting to and surviving in varying conditions, both environmental and genetic. The relative viability of conditions such as confined placental mosaicism for chromosomal imbalance or placental trisomy demonstrate that the placenta is able to adapt to adverse genetic conditions. The plasticity of the placenta allows it to function and support life in sub-optimal conditions. However, with this plasticity comes a fitness cost (Chevin et al., 2010). A small reduction in placental function early in development may mean that there are larger downstream effects. For example, due to a reduction in the transfer of nutrients from the mother to the fetus, there can be alterations to the epigenetic programming of the developing brain, predisposing the offspring to later life diseases such as obesity, diabetes, and schizophrenia (Cooney et al., 2002; Keverne & Curley, 2008; Schulz, 2010).

1.2 Epigenetics

Over recent decades, it has become increasingly apparent that the genetic sequence is not the only major player when it comes to understanding the processes of growth, development and disease. Epigenetic factors allow the genome to interact, not only
with itself in a regulatory manner, but also with the environment. While Waddington’s original use of the term ‘epigenetics’ was to describe the developmental processes by which genotypes give rise to phenotypes (Hall, 1992; Peters, 2009), the definition of the word has shifted. Today, the more commonly accepted definition of the word may be closer to one offered by Russo et al. (1996), which states that epigenetics is the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence. These changes in gene function can be brought about at a chromosomal level, a chromatin level and at a DNA/RNA level, and provide precise regulation of gene expression (Bird, 2007). The molecular basis of mammalian epigenetics is largely attributable to three main mechanisms: DNA methylation, histone modification and non-coding RNAs. Uncovering the variation in the epigenome will make possible the identification of both rare and common variations that are associated with gene behaviour and disease susceptibility, akin to the impact that uncovering genomic variation has had.

1.2.1 DNA Methylation

DNA methylation generally refers to the addition of a methyl group (CH$_3$) to a cytosine base. In mammals, the addition of this methyl group to CpGs at key sites in the genome can play an important role in the regulation of gene expression, altering the way DNA interacts with itself and the environment. These methylation marks are heritable through cell divisions, and vary between cells of different lineages, altering the potential function of cells, and aiding in the development along different pathways. In this way, cells that contain the same DNA sequence can display many different phenotypes. DNA methyltransferases attach methyl groups generated by the one carbon metabolism pathway to cytosine residues in the genome to form 5-methylcytosine. This occurs predominantly when the cytosine is located next to a guanine in the DNA sequence, and the two are linked by the phosphate backbone (CpG). In mammals, there are three DNA methyltransferases; Dnmt1, Dnmt3a, and Dnmt3b that are known to be essential for both embryonic development, and de novo methylation of CpG sites. DNMT1 is mainly involved in maintaining DNA methylation after DNA replication,
while DNMT3a and DNMT3b are largely responsible for establishing de novo methylation marks (He et al., 2011; Peters, 2009). This CpG methylation is a key regulator of transcription during development and differentiation of cells, and plays a significant role in ensuring the correct genes are turned off at the correct time. In particular, dense CpG methylation has the ability to effectively silence transcription via two mechanisms (Gaudet et al., 2003). Firstly, methylated DNA is able to directly prevent the binding of transcription factors required for gene expression (Watt & Molloy, 1988). Additionally, methylated DNA can recruit methyl-binding domain proteins and histone modifying enzymes such as histone deacetylase to confer histone modifications and chromatin structural changes that also silence expression (He et al., 2011; Jones et al., 1998b). The best understood example of this type of silencing is X inactivation in mammals. Since females receive two copies of the X chromosome (while males receive one copy and one, much smaller Y chromosome), genes on the ‘extra’ X chromosome in females are inactivated as a form of gene dosage compensation. During inactivation, gene promoters on one copy of the X chromosome will become highly methylated at gene promoters due to the action of Dnmt3a/Dnmt3b (Reik & Walter, 2001), inducing the formation of facultative heterochromatin, which is visible in the cell as a Barr body (Chadwick & Willard, 2003). Interestingly, the active X chromosome (Xa) actually shows a global level of methylation that is twice as high as the inactive X chromosome (Xi), but the localisation of the methylation is divergent. The Xa shows specific hypomethylation at promoters, allowing transcription, and hypermethylation at gene bodies; the Xi chromosome methylation is concentrated at CpG islands (Hellman & Chess, 2007).

1.2.1.1 CpG Islands

Vertebrate genomes are comparatively CpG poor, a result of the mutagenic ability of methylcytosine to spontaneously deaminate to TpG (Bird, 1980; Deaton & Bird, 2011). Interspersed in the genome however, are regions that are relatively CpG rich, termed
CpG islands (CGI). About 70% of known gene promoters are associated with CGIs (Saxonov et al., 2006). In the majority of cells and for the majority of genes, these promoter-associated CGIs are unmethylated, allowing transcription of the downstream gene. To control for tissue-specific and time-specific expression, CGIs can become methylated when expression of a gene is not needed. As a result of this mechanism of gene regulation, CGI promoter methylation can be a useful marker for gene expression, particularly in a tissue-specific context (Ghosh et al., 2010; Song et al., 2005).

1.2.2 Other Mechanisms of Epigenetic Control

1.2.2.1 Histone Modification

Histones are protein complexes around which DNA is wound to form nucleosomes, and form the basic unit for chromatin. When the chromatin is densely packed, it is referred to as heterochromatin, and transcription is largely blocked. The portions of the genome that are open for transcription are more loosely bound around the histone proteins, allowing access for transcription factors and polymerase enzymes (Kornberg, 1974). Each nucleosome in eukaryotes is comprised of four histone proteins, H2A, H2B, H3 and H4. Post-translational modifications to these proteins regulate the chromatin structure and thereby modulate gene activity. Acetylation, methylation, ubiquitinylation, or phosphorylation of the histone N-terminus can all have differing effects on transcription (Jenuwein, 2001). For example, acetylation of the lysine at position 27 of the histone 3 tail (H3K27) is associated with gene activity and the euchromatin state, as are H3K4me3 and H3K9me1. On the other hand, di- or tri-methylation of H3K9 has been found to be associated with transcriptional repression (Rosenfeld et al., 2009). DNA methylation is known to recruit histone-modifying enzymes, showing that the two epigenetic mechanisms are co-dependent in regulating gene expression (He et al., 2011; Henckel et al., 2009).
1.2.2.2 Non-Coding RNAs

Changes to chromatin architecture and DNA methylation are commonly regulated by RNA signalling using long non-coding RNAs (ncRNAs). These ncRNAs are still not fully understood, but it has been suggested that they play a functional role in modifying chromatin at specific genomic location by acting as guides for other proteins (Caley et al., 2010). For example, cryptic antisense ncRNA transcripts (transcripts that are degraded quickly) have been seen to play a role in epigenetic silencing of retrotransposon activity in trans by aiding deacetylation and methylation of histones (Berretta et al., 2008). In addition, ncRNAs are essential for the maintenance and modulation of many imprinted genes. The imprinted gene, Igf2r, has an ncRNA antisense transcript, Air, that is expressed from intron 2 of the paternal allele. This antisense transcript plays an active role in silencing of not only the paternal copy of Igf2r, but also Slc22a2 and Slc22a3 (Amaral & Mattick, 2008).

1.3 Imprinting

The umbrella term imprinting covers multiple aspects. This term is canonically used to describe the situation in which a gene is expressed monoallelically in a parent-of-origin specific manner. This is generally brought about by epigenetic instructions that originate in the parental germ cells. Imprinting is particularly important in the embryonic development of mammals as it provides an important mechanism for dosage control of key genes. However, it is sometimes used more broadly to describe tissue- or temporal-specific expression of a single allele. It has been seen in a number of pathological states that the timing and location of monoallelic expression is fundamental to normal function (Reik & Walter, 2001). In the 1990s, it was found that DNA methylation of the cytosine residue at certain CpG dinucleotides, particularly in the promoter regions of genes, is a significant factor in the molecular mechanism of imprinting (Li et al., 1993). Methylation marks in the egg and sperm differ, and these marks are passed on to the zy-
gote. Which allele of a gene is expressed is then determined by the inheritance of these epigenetic marks. Our understanding of imprinting was galvanised by the observation that the maternal and paternal genomes are not functionally equivalent. That is, having two copies of either the maternal or paternal genome is not enough to create a viable organism. So what is it that makes the copy of your genome that you inherit from your mother different to that you inherit from your father? And why is it important?

1.3.1 Theories for Existence of Imprinting

1.3.1.1 The Parental Conflict Hypothesis

Most of the currently known imprinted genes are involved in either fetal development, placental development or nurturing behaviour (Piedrahita, 2011). As this system of gene regulation exists only in placental mammals and marsupials, but not in the prototheric monotremes, it is believed that it may have evolved along with the unique need to protect the mother from the fetus, and the fetus from the maternal environment (Kim et al., 2001). An additional reason behind the existence of imprinted genes has been proposed in the form of the ‘parental conflict hypothesis’. This theory states that the mother, and therefore the maternal genome copy, has an imperative to conserve her resources to ensure she is able to produce subsequent offspring. The father’s genome however, has a need to maximise the survival and growth of his offspring so as to outcompete other potential fathers. Studies in both mice and humans have offered evidence to support this theory.

In mice, parthenogenetic embryos (diploid, digynic) show intrauterine growth restriction (IUGR), and an extremely underdeveloped placenta. In contrast, androgenetic embryos (diploid, diandric) show almost the reverse phenotype, with barely any embryonic development but a large, proliferative placenta (McGrath & Solter, 1984). The parallel disorders in humans, ovarian teratomas and complete hydatidiform moles, show a similar phenotype (Tabano et al., 2010; Yuen et al., 2011a). Triploid studies have also been performed in both mice and humans, in which either the mater-
nal or the paternal contribution is overrepresented. In humans, triploidy occurs in 2-3% of pregnancies. These pregnancies are never viable, and rarely survive to term, commonly ending in miscarriage (McFadden & Kalousek, 1991). A diandric triploid (two paternal genomes, one maternal) usually presents with a normal-sized or slightly growth-restricted fetus and a large, cystic placenta. The digynic triploid (two maternal genomes, one paternal genome) phenotype is characterised by an extremely growth restricted fetus and a small undeveloped placenta (McFadden & Kalousek, 1991; Yuen et al., 2011a). Specific imprinted genes have been found which reinforce the idea that maternally expressed genes restrict fetal growth and paternally expressed genes promote fetal growth. Perhaps the best-understood genes to illustrate the theory are \textit{Igf2} and \textit{Igf2r}, which has already been mentioned in the previous section. \textit{Igf2} is a paternally expressed insulin-like growth factor, and \textit{Igf2r} is a maternally expressed receptor that mediates degradation of IGF2 by the lysozymes.

1.3.1.2 Non-conflict Theories

Although a commonly cited theory behind the evolution of imprinting in mammals is the parental conflict hypothesis, it is certainly not the only hypothesis available. Spencer & Clark (2014) provide an excellent summary of the non-conflict based ideas surrounding the existence of imprinting, particularly in mammals. Some imprinting may have arisen as a genome defense against transposable elements and retroviruses, as borne out by the observation that many of these stretches of ‘foreign’ DNA are silenced by DNA methylation (McDonald et al., 2005). However, this theory is less able to explain the parental asymmetry that is seen in imprinting. Theories such as maternal-fetal co-adaptation (Wolf & Hager, 2006), and ‘mother knows best’ (Keverne & Curley, 2008) would account for the overrepresentation of imprinting in the placenta, and allow the mother to co-ordinate fetal development with maternal resources, but do not adequately account for the expression of the paternal allele in many cases of placental imprinting (Spencer & Clark, 2014). No one theory seems able to account for all im-
stances of genomic imprinting, supporting the idea of a mixed model in which different imprinted genes may have different origins.

1.3.2 Primary Imprinting Control

Known imprinted genes are clustered reciprocally in discrete regions of the genome, with maternally silenced and paternally silenced genes being located near each other. These regions are controlled over long distances by Imprint Control Regions (ICRs). Centres of imprint control are differentially marked in the germline through DNA methylation, and subsequently developed and maintained throughout fetal development and later life (Radford et al., 2011). The epigenetic programming of gametes begins after primordial germ cells (PGCs) have been specified at embryonic day 7.5 (in mice), and continues through the migration of the PGCs to the genital ridge (Radford et al., 2011). The PGCs are initially highly methylated, yet as they migrate, they undergo a wave of genome-wide demethylation (Bestor, 2000; Dean et al., 2003; Reik & Walter, 2001; Reik et al., 2003b). This demethylation event has the effect of resetting the methylation imprints in the genome. The imprints are then re-established in the developing gametes, with the sperm re-methylation taking place almost immediately, and being complete shortly after birth (Li et al., 2004). The re-methylation wave in the oocyte doesn’t take place until maturation of the gamete, shortly before ovulation (Bourc’his & Proudhon, 2008; Lucifero et al., 2004).

After fertilisation, but before implantation, the parental genomes also undergo a sweeping demethylation event (Delaval & Feil, 2004; Jaenisch & Bird, 2003). Imprinted genes, however, resist this global wave of demethylation. De novo methylation at non-imprinted genes and secondary imprinting sites takes place after implantation. De novo re-methylation of the developing embryo after implantation is highly susceptible to the maternal environment (Nye et al., 2014; Osborne-Majnik et al., 2013; Perera & Herbstman, 2011). The maternal supply of one-carbon groups, and in particular folate, is essential to the normal methylation of the fetal genome. There is some
limited evidence to suggest that the availability of such nutrients around the time of conception and implantation may affect the re-methylation processes (Kwong et al., 2006; Oliver et al., 2005; Radford et al., 2011).

In the mouse, the X-chromosome is not randomly inactivated in extra-embryonic lineages, unlike in the embryo, which can be mosaic for X-inactivation (Okae et al., 2012). This suggests that the extra-embryonic lineages may be derived from a single stem cell. This in turn leads to the question of whether in the placenta, the mechanism for maintaining imprinted expression differs from that in the embryo. The expression may not depend on promoter methylation, but this would require an in depth investigation of epigenetics and function before any conclusions could be drawn.

Figure 1.3: Reprogramming and Maintenance of Imprint Control Regions (ICR) in mammals. Reproduced with permissions from (Delaval & Feil, 2004).
1.3.3 Secondary Imprints - Promoter Regulation

As already mentioned, somatic, or secondary differentially methylated regions (DMRs) are found in the promoter region of some genes. These DMRs are thought to reinforce imprinted gene expression, as they obtain their methylation post-fertilisation (Dean et al., 2003; Reik & Walter, 2001), and require the gametic imprinting control region (ICR) to be present (Radford et al., 2011). Methylation can then spread from the ICR into the promoter to form a DMR (Reik & Walter, 2001). DMRs can also be tissue- and temporal-specific, which has been demonstrated by Yuen et al. (2011a) and Liang et al. (2011). DMRs are characterised not only by differential methylation marks on the two alleles, but also different histone modifications (Radford et al., 2011). It is these secondary differential methylation marks in the promoters of genes that have strong associations with gene-expression alterations between tissues, and during development, though the epigenetic signature at the ICR will be maintained throughout development (Henckel & Arnaud, 2010).

![Figure 1.4](image.png)

**Figure 1.4:** The current view of promoter methylation. Absence of methylation at the promoter allows transcription factors to bind and recruit polymerases, resulting in transcript production. Methylation at these locations prevents the binding of transcription factors.
1.3.4 Allele Specific Methylation and Parent-of-Origin Effects

One complication of identifying imprinted and differentially expressed regions is the existence of polymorphisms that can alter methylation and expression. For example, the gene ANO1 shows maternal specific expression only in some term placentas (Okae et al., 2012). Allele-specific methylation (ASM) refers to the situation whereby a specific allele correlates with how the gene is methylated. Because epigenetic signatures are established in the germ-line, it has been assumed that monoallelic methylation is largely due to imprinting. However, an increasing number of authors are reporting non-imprinted autosomal ASM, where a correlation between DNA methylation and gene expression is also seen (Kerkel et al., 2008; Yamada et al., 2004; Zhang et al., 2009). Recently, allele-specific methylation has been documented to be widespread through the human genome, and that ASM may represent an important epigenetic pathway that connects genetic polymorphisms with phenotypic variability (Zhang et al., 2009).

Parent-of-origin effects are similar in effect to allele-specific methylation, but the cause is instead which parent you inherit the allele from that affects how it is methylated and/or expressed, independent of gene sequence. Parent-of-origin effects are similar in mechanism to imprinting, but instead of switching off one allele completely, it may instead alter the rate or efficiency of transcription of one allele, especially if the methylation differs at key binding sites for transcription factors and histones (Boumber et al., 2008; Seoighe et al., 2006). It is important to note that allele-specific methylation and parent-of-origin effects are not necessarily examples of true imprinting. Denser methylation, or methylation at key sites may decrease the efficiency of transcription without turning it off, or alter how and when it is expressed in certain tissues.

The ability to discriminate allele-specific expression of the two alleles of genes can reveal differences in epigenetic control. The two alleles are in theory affected by the same transcription factors, yet may differ in cis-acting control elements (Bjornsson et al., 2008; Shoemaker et al., 2010).
1.3.5 Methods for Identifying Differentially Methylated Regions

Many approaches have been taken for identifying DMRs, both in an imprinted gene context and in a tissue-specific expression context. A summary of the commonly used methods for studying DNA methylation can be found in Table 1.1. Initially, methylation patterns were analysed from a candidate gene perspective. This approach remains effective for investigating, in detail, the behaviours of individual genes and understanding specific mechanisms regulation. Two of the main techniques used over the past two decades to identify differential methylation at specific loci are Combined Bisulphite Restriction Analysis and bisulphite sequencing. The first, Combined Bisulphite Restriction Analysis (CoBRA), makes use of the ability of certain restriction enzymes such as MspI and HpaII to cut differentially at CpG containing sites. The digest is performed on DNA that has been bisulphite converted and PCR amplified, and the ability of the enzyme to digest the CpG site is dependent on whether methylation at the CpG has caused the cut site to be retained or lost (Waalwijk & Flavell, 1978; Xiong & Laird, 1997). This technique has largely fallen out of use in a research environment, but is still sometimes used in a clinical context for such things as the diagnosis of Beckwith-Weidemann Syndrome (Hattori et al., 2009; Priolo et al., 2008). The second, and possibly more in-depth technique is to use bisulphite sequencing. In this technique, bisulphite conversion is used to convert unmethylated cytosines to uracils. The desired product is then amplified via PCR. The PCR product can then be sequenced directly to give an average methylation across that locus, or can be cloned into an *E. coli* vector. Sanger sequencing of the cloned product allows differentiation between both strands of both alleles, as well as identification of methylated and non-methylated CpG sites (Frommer et al., 1992). If intermediate methylation is due to unequal methylation levels (either in single cells, or heterogeneous methylation in a cell population), then the methylation levels on adjacent CpG sites should be highly correlated via linkage disequilibrium. Shoemaker et al. (2010) developed an algorithm for SNP calling from bisulfite sequence reads that enabled them to see that heterozygous SNPs allow for
discernment between methylation haplotypes.

Genome-wide techniques have now become very popular for identifying candidate genes in many aspects of genetic and epigenetic research. Techniques such as promoter and CpG microarrays allow the interrogation of a large number of CpG islands and promoter regions across a diverse set of cell types (Schumacher et al., 2006). With this method, unmethylated and methylated DNA are enriched using a series of treatments with multiple methylation sensitive restriction enzymes, and interrogated on microarrays (Yan et al., 2009). Current chips such as the Infinium HumanMethylation450 BeadChip array (Illumina, Inc., San Diego, CA USA) are able to interrogate over 485,000 CpG sites in 99% of RefSeq genes, in sites distributed across the promoter, 5’UTR, first exon, gene body, and 3’UTR. Immunoprecipitation has also been adapted to methylation analysis, with the development of an antibody against 5-methylcytosine (Reynaud et al., 1992). This antibody allows for enrichment and immunoprecipitation of methylated DNA sequences (MeDIP). The immunoprecipitated DNA can then be labelled and hybridised to a microarray (Weber et al., 2005).
The mouse and human imprintomes of placenta has used transcriptome analysis (RNA-seq) to identify differential expression, and imprinted genes have been identified based on a statistical analysis of allelic bias for expressed SNPs (Barbaux et al., 2012). In this way many putatively imprinted genes have been identified. However, some genes, such as Ano1 and Gab1 in the mouse, while they show imprinted-type parent-of-origin expression, do not show allele-specific promoter methylation, suggesting another method of regulation (Okae et al., 2012).

1.3.5.1 RRBS

Most methylation analysis done on placentas to date has consisted of microarrays and targeted candidate-gene approaches. While microarrays are a relatively cheap and simple way of generating data from a wide array of different genes and different samples, they are somewhat limited in their scope. Microarrays are limited to a small number of sites per gene (averaging 17 sites per gene at time of publication), and do not cover all CpG rich regions of the genome. If we are to gain a better understanding of the role of methylation in gene regulation, it is important that we look at the epigenome on a much larger scale.

Reduced Representation Bisulphite Sequencing (RRBS) is a technique that has recently been developed for looking at the methylome with much greater resolution (Meissner et al., 2005). This technique relies upon the generation of bisulphite sequencing libraries that enrich for CpG-dense regions by combining methylation-insensitive restriction digestion with size selection (Smith et al., 2009). The libraries can then be sequenced on a high-throughput sequencing machine such as the Illumina HiSeq 2000. Array based techniques are limited to only 2 or 3 CpG sites per region interrogated, but the libraries sequenced by RRBS can comprise many more CpG sites per gene (depending upon CpG density). Because there are hundreds of unique methylomes to map within one individual and interindividual variation is likely to be significant, international coordination using techniques such as RRBS is needed to create a full profile of
Table 1.1: A selection of the commonly used methods for studying methylation, with the relative genome coverage, unit of analysis, estimated cost per sample, and some of the potential sources of bias for each method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Genome Coverage</th>
<th>CpG sites</th>
<th>CpG Islands</th>
<th>Cost per Sample (USD)*</th>
<th>Resolution</th>
<th>Potential Sources of bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGBS</td>
<td>Whole Genome</td>
<td>30 million</td>
<td>27,769</td>
<td>$10,000</td>
<td>CpG site</td>
<td>Bisulphite conversion, PCR amplification</td>
</tr>
<tr>
<td>RRBS</td>
<td>Reduced Genome</td>
<td>6 million</td>
<td>24,633</td>
<td>$800</td>
<td>CpG site</td>
<td>Bisulphite conversion, Restriction Digest, PCR amplification</td>
</tr>
<tr>
<td>450k Microarray</td>
<td>Targeted</td>
<td>485,512</td>
<td>19,755</td>
<td>$300-400</td>
<td>CpG site</td>
<td>Bisulphite conversion, Batch effects</td>
</tr>
<tr>
<td>MeDIP-seq</td>
<td>Whole Genome</td>
<td>26-28 million</td>
<td>26,936</td>
<td>$2,000</td>
<td>&gt;100</td>
<td>PCR amplification, Relative Measurement, Restriction Digest, Antibody quality</td>
</tr>
<tr>
<td>CoBRA</td>
<td>Targeted</td>
<td>Amplicon dependent</td>
<td>NA</td>
<td>$1</td>
<td>Cut-site dependent</td>
<td>Enzyme quality, Cut-site dependent Semi-quantitative</td>
</tr>
<tr>
<td>Sequenom</td>
<td>Targeted</td>
<td>Amplicon dependent</td>
<td>NA</td>
<td>$2500/plate</td>
<td>CpG site</td>
<td>Bisulphite conversion, PCR amplification</td>
</tr>
<tr>
<td>Pyrosequencing</td>
<td>Targeted</td>
<td>Amplicon dependent</td>
<td>NA</td>
<td>$1.50/assay</td>
<td>CpG site</td>
<td>PCR amplification, Repeat-poor</td>
</tr>
<tr>
<td>Bisulphite</td>
<td>Targeted</td>
<td>Amplicon dependent</td>
<td>NA</td>
<td>$6.50</td>
<td>CpG site</td>
<td>Bisulphite conversion, PCR amplification</td>
</tr>
<tr>
<td>Sequencing</td>
<td>Targeted</td>
<td>Amplicon dependent</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NB: The Infinium 450K BeadChip became available in late 2011/early 2012, after this project had begun. Prior to this, a much smaller Infinium®Human Methylation 27K BeadChip was available.

WGBS - Whole Genome Bisulphite Sequencing; RRBS - Reduced Representation Bisulphite Sequencing; MeDIP - Methylated DNA immunoprecipitation sequencing.

*Costs are estimated based upon personal communications
normal and abnormal inter- and intra-individual methylation (Meissner et al., 2005).

The RRBS method as developed by Meissner et al. (2005) uses the methylation-insensitive restriction enzyme MspI to cut the whole genome at CCGG sites, creating a variably fragmented DNA library. Size selection on these fragments then allows one to enrich for regions of the genome that are particularly rich in CpG sites. By selecting fragments between 40 and 220 bp, one selects only $\sim 2\%$ of the genome, but this contains 80% of CpG islands, and 3 million CpG sites. Next-generation sequencing performed upon these libraries provides a huge amount of data. The primary limitation to this method is that the tools for analysing the data generated by RRBS are still in their infancy. Variations in the length of fragments stemming from the MspI digestion, read count from the sequencing protocol, and the asymmetric dinucleotide nature of CpGs all contribute unique bioinformatic challenges to the analysis of RRBS data.

Another key problem is the unit of methylation. While it is accepted that the majority of methylation occurs at CpG dinucleotides (although there is a growing body of evidence regarding methylation at other nucleotides), it is not yet fully understood how, and if, these sites of methylation act in concert. The current literature catalogues any number of individual CpG sites that show differential methylation between tis-
Figure 1.7: Molecular Construction of RRBS Libraries. Genomic DNA is extracted from tissue, then cleaved using the MspI restriction enzyme. Overhangs generated by the digestion are filled in, and an A-tail attached. Adaptors are then ligated to each end of each fragment, and a size-selection step carried out. Bisuphite treatment converts unmethylated Cytosine residues to Uracils. PCR is used to amplify the libraries, and a second round of size selection is performed to remove excess adaptors before sequencing.
sues, diseases, and developmental time points. Not every one of these differences will be functional however. Some may be down-stream effects of other mechanisms, and some may be purely stochastic. However, there is also a focus on identifying global patterns in methylation. Cancers for examples are commonly classed as globally hypo- or hyper-methylated (Kulis & Esteller, 2010; Kulis et al., 2012). The most common statement made about methylation in the placenta is that it is a globally hypomethylated tissue (Schroeder et al., 2013) - although new work, some of it by our research group, is starting to show regions of specific hyper-methylation in the placenta. Methylation sites can act together to prevent the binding of transcription factors, and perhaps the most evident example of this is the inactivated X chromosome in females. Methylation blankets much of the silenced chromosome, recruiting histone modifications in their turn to wrap most of the chromosome up so tightly that it is more or less transcriptionally silent (Csankovszki et al., 2001). We need to better understand how methylation acts together and separately in order to truly understand the mechanism of action and its role in functional biology.

1.3.6 Dysregulation of Genomic Imprinting

Given that differential methylation is so crucial to regulating gene expression, a pathogenic state can arise when the methylation itself is disrupted or altered. The role of epigenetics in cancer provides a good illustration of this phenomenon. Tumour cells exhibit a globally hypomethylated epigenetic profile, but DNA methylation in promoter CGIs is often unusually high. In mouse models that have a hypomorphic copy of Dnmt1, hypomethylation has been shown to contribute to the genomic instability seen in cancer cell types, and to aid tumorigenesis (Gaudet et al., 2003). In addition to this global hypomethylation, localised hypermethylation is observed at many CGIs, and this over-methylation is understood to contribute to the silencing of many tumour suppressor genes (TSGs) (Deaton & Bird, 2011; He et al., 2011). Whether the cancer-specific methylation of TSGs is a cause of gene silencing, or a consequence of
it remains to be seen.

A number of imprinting disorders have also been described that occur when the normal methylation status of imprinted genes is disturbed. Some of these, such as Beckwith-Wiedemann Syndrome, and its imprinting parallel, Silver-Russell Syndrome, have been linked to either hypo- or hyper-methylation of one or more imprint control regions located on chromosome 11. The 11p15 region is quite densely imprinted, with two domains, the IGF2/H19 domain, and the KCNQ1OT1/CDKN1C domain, each under the control of an ICR (Demars et al., 2011). These regions are highly conserved in mammalian species, and provide an excellent ‘model’ for the study of imprinting (Shmela & Gicquel, 2013).

Dysregulation of genomic imprinting can also play an important role in the epigenetic programming and development of the placenta. In mice (Okae et al., 2012) loss of the imprinted gene Phlda2 causes overgrowth of the placenta but not the fetus (which are actually smaller). Conversely, elevated Phlda2 is associated with IUGR and low birth-weight in humans (although the imprinted status of PHLDA2 in humans is in some doubt) (Lefebvre, 2012; Morison et al., 2001).

1.4 Pre-eclampsia/Toxaemia

Pre-eclampsia/Toxaemia is a placental disorder affecting 4-8% of all pregnancies. It can, if untreated, cause adverse outcomes for both mother and child, including miscarriage, premature labour, seizures and death. The recurrence risk of pre-eclampsia more than doubles with each pre-eclamptic pregnancy a woman undergoes (Hernandez-Diaz et al., 2009). As well as acute illness, there is a growing body of evidence regarding the long-term effects of pre-eclampsia on mother and child, contributing to later metabolic disease, particularly cardiovascular disease (Díaz Martínez et al., 2011; Jones et al., 1998a; Mongraw-Chaffin et al., 2010). As such, it is a disease that is a foremost concern of maternal and perinatal health. It seems surprising then that, despite the work put in by many researchers and clinicians, we have still not come to a clear definition
of the true nature of the disease, and indeed are unable to detect it until 20 weeks of gestation, more than halfway through pregnancy. Roberts & Post (2008) define Pre-eclampsia/Toxaemia (henceforth referred to as PET) as abnormal implantation resulting in placental ischaemia. Shallow or incomplete invasion of the trophoblast into maternal tissues, and inadequate physiological conversion of the spiral arteries needed to perfuse the placenta are both believed to back the basic biology of placental ischaemia - but the root cause is still largely unknown. Clinically, PET is defined as a blood pressure at or above 140 mmHg systolic, and/or diastolic blood pressure greater than or equal to 90 mmHg, in conjunction with proteinuria (for a more comprehensive list of clinically relevant symptoms, see Chapter 2.5). Intrauterine growth restriction is often superimposed upon PET, leading to higher rates of pre-term birth and caesarian section (SOMANZ, 2014).

1.4.1 Molecular Mechanisms of Pre-eclampsia/Toxaemia

In PET pregnancies, the invasion of the spiral arteries by the cytotrophoblast is limited. The arteries remain small in diameter and retain their muscular covering, resulting in restriction of blood flow to the placenta (Figure 1.8) (Redman, 1990; Zhou et al., 1997). This ischaemia is characteristic of PET, and is thought by some to be the cause of many of the symptoms of PET. The hypoxia extends to the syncytiotrophoblast, as indicated by increased syncytial shedding and apoptosis in pre-eclampsia (Austgulen et al., 2004). Counterintuitively, smoking seems to counteract this, as it has been shown to slow syncytiotrophoblast proliferation and apoptosis (Marana et al., 1998). However, smoking also causes maternal hypoxaemia. sFLT1 and Endoglin expression is raised in response to placental ischaemia (Roberts & Post, 2008). sFLT1 binds to circulating factors, depleting their availability to FLT1 binding, which is essential for correct endothelial function. Fetal genes are predicted to raise maternal blood pressure in order to enhance the uteroplacental blood flow, whereas maternal genes act contrary to them. If fetal and maternal genes are not regulated correctly, this could contribute to
placental dysfunction.

**Figure 1.8:** Remodelling of spiral arteries in pregnancy. **A.** During implantation, cytotrophoblasts invade and remodel the smooth muscle of the maternal spiral arteries, relaxing the blood vessels and thus providing the developing fetus with an adequate blood supply. **B.** In pre-eclampsia, this invasion is reduced, and spiral arteries are inadequately remodelled, restricting the blood supply.

A growing number of specific molecules and pathways have been implicated as having a role in the development or maintenance of PET, and much work has been done in the mouse which, although it differs in placentation, has advantages as a model organism. As such, a limited number of mouse models of PET exist, or it would be fairer to say that mouse models exist for some elements of PET. Loss of *Igf2* leads to an IUGR phenotype that affects both the embryo and the placenta (DeChiara *et al.*, 1990). Further experiments on this model using a placenta-specific promoter determined that it is the loss of *Igf2* expression specifically in the placenta that causes the fetal IUGR phenotype by reducing the placental surface area, and thus the perfusion of the placenta by maternal blood (Sibley *et al.*, 2004). The loss of a related gene, *Igf1*, causes a reduction in fetal size, but not placental size (Constancia *et al.*, 2002). *Cdkn1c* knockout mice display an overgrowth phenotype at E15.5. This growth trajectory is unable to be sustained however, a phenomenon which is a classic indicator of
placental insufficiency, as the fetus first tries to overcompensate, and then the mother is unable to meet the increased demands of a large fetus (Lefebvre, 2012; Tunster et al., 2011). Growth hormone plays a central role in pregnancy, but expression of the whole Human Growth Hormone/Chorionic Somatomammotropin (hGH/CSH) gene cluster is generally reduced in PET, significantly so for GH2-1, GH2-2, and CSH1-2 (Mannik et al., 2012).

1.4.2 Genetics of PET

The genetics of pre-eclampsia remain unclear. Female offspring of a PET pregnancy have a higher risk of developing PET themselves and linkage studies have suggested a multigene disorder, possibly involving genes with penetrance less than 50% (Williams & Broughton Pipkin, 2011). Candidate genes that have been identified in the mother and fetus include *AGT*, *FS*, *MTHFR*, and *PAPPA2* (Sheppard & Khalil, 2010). *PAPPA2* encodes a placental-specific pappalysin. These are proteases that cleave IGFBPs, resulting in local activation of IGF signalling pathways that are important regulators of fetal growth and development (Callan & Milne, 2009). There is an association between preeclampsia and associated HELLP syndrome in mothers with a child with Beckwith-Wiedemann syndrome and a mutation on the maternal CDKN1C allele (Romanelli et al., 2009). *STOX1* is another gene which has been associated with pre-eclampsia in multiple studies, (Lachmeijer et al., 2001; van Dijk et al., 2005) and suggested that epigenetic regulation of alleles was involved, but later studies dispute the finding of monoallelic expression (Iglesias-Platas et al., 2007; Kivinen et al., 2007). Much research is still being done on the role of *STOX1* in placental development due to the observation that *STOX1* overexpression in mouse trophoblasts confers a pre-eclampsia-like phenotype (Doridot et al., 2014).

The paternal contribution is not insignificant, with the relative risk of PET in a woman whose partner had already fathered a PET pregnancy in another woman being 1.8, regardless of previous PET (Williams & Broughton Pipkin, 2011). Renin genes from the
father may contribute to increasing the risk of hypertension, along with angiotensin II genes from the mother (Lie et al., 1998; Takimoto et al., 1996). In addition to this, an increased period of sexual co-habitation with the father reduces the risk of pre-eclampsia, introducing the hypothesis that aspects of the immune response could have significant impacts upon the risk of pre-eclampsia (Ros et al., 1998; Trogstad et al., 2001). A polymorphism in angiotensin is associated with severe PET, and significant linkages have been noted on chromosomes 2p13, 2p25, 9p13, and to a lesser extent at 2q, 9p, 10q, 11q, and 22q (Moses et al., 2000). Placental hypoxia as induced by poor cytotrophoblast invasion and inadequate remodelling of maternal arteries causes the production of biologically active factors such as growth factor inhibitors, anti-angiogenic proteins, inflammatory cytokines, reactive oxygen species, hypoxia-inducible factors, and antibodies to vascular angiotensin II receptor. These factors can all affect the production and activity of various vascular mediators in the endothelium, smooth muscle and extracellular matrix, leading to severe vasoconstriction and hypertension, one of the key manifestations of pre-eclampsia (Sheppard & Khalil, 2010). Many more genes have been identified as potential potential risk factors for pre-eclampsia, however, monozygotic twin studies have shown low concordance, meaning that a single gene is unlikely to be the root cause of pre-eclampsia in most cases. The complexity and low concordance of the disease between twins adds weight to the the concept of an environmental contribution to risk. A review by Jebbink et al. identified 178 genes, miRNAs and proteins in the literature that have previously been reported to be associated with pre-eclampsia.

1.4.3 Epigenetics of PET

1.4.3.1 Role of Imprinting in the Placenta

Imprinting is believed to be crucial for placental function. Interestingly, the placenta contains the highest number of expressed imprinted genes, which supports a unique placental function for imprinted genes that is distinct from functions in normal so-
matic tissue. Part of the reason for this abundance of expressed imprinted genes may be due to the lower levels of methylation in the placenta, in particular at repetitive elements (Coan et al., 2005; Macaulay et al., 2011; Reik et al., 2003a; Wang et al., 2013). In mice, about a quarter of imprinted gene expression is specific to the placenta or extra-embryonic tissues, which lends weight to the idea that the genomic imprinting is an essential mechanism in placental development and function in viviparous species (Okae et al., 2012). On the other hand however, while somatic imprinted genes are highly conserved (Monk et al., 2006a), there is a lack of conserved imprinting between human and mouse placenta. While this may reflect a paucity of knowledge about imprinting in the human placenta, it may also be a reflection of evolutionary differences in reproductive strategies between mice and humans. The maternal and paternal chromosomes are, though paired, not functionally equivalent, and imprinting is initially determined in the germ line. According to the parental conflict hypothesis of imprinted genes, paternally expressed genes are involved in the promotion of fetal growth, while maternally expressed genes are expected to restrict fetal growth and control maternal energy expenditure (Piedrahita, 2011). This sort of relationship is expected to be absent in oviparous species, as the window of maternal input is much shorter, and the period before maternal-fetal RNA transition is extended. In oviparous species, maternal mRNAs regulate fetal development until the maternal to zygotic transition (Baroux et al., 2008). Imprinted genes expressed in the placenta may play a similar role, controlling nutrient transfer, while fetal-expressed genes may regulate growth rate and determine nutritional demand (Lefebvre, 2012; Li, 2010).

The plasticity of the placenta is demonstrated by the ability of the placenta to sustain cellular defects that would be lethal if present in the fetal component. A more ‘relaxed’ control of DNA methylation in the placenta may mean that it is able to sustain growth in more adverse cellular conditions; which may contribute to the pathophysiology of placental insufficiency, one of the key factors behind pre-eclampsia (Yuen & Robinson, 2011). In fact, some epipolymorphisms have already been identified as increasing the risk of pre-eclampsia (Chelbi et al., 2007; Yuen et al., 2009).
Many putative imprinted genes have been identified in the placenta, but have later been disproved. Genes that are highly expressed in the maternal decidua can be falsely identified as showing placenta-specific imprinted expression due to contamination by maternal blood and decidua. Therefore, many placental-specific maternally expressed genes are of doubtful status, as the probability of maternal contamination in placental tissues is so high (Adibi et al., 2009; Lamb et al., 2012; Proudhon & Bourc’his, 2010). This contamination problem can be surmounted by the use of trophoblast stem cell (TSC) cultures to confirm imprinted gene expression in the trophoblast lineage. Techniques for deriving and culturing TSCs are still being established, but these should prove a powerful tool in the future (Lefebvre, 2012). Embryo transfer is an alternative way of combatting maternal contamination in imprinting studies that is more useful in model organisms. Transferring embryos between genetically dissimilar animals (e.g. different mouse strains) allows for the teasing out of false positives. Okae et al. (2012) used this method to refute evidence for 12 putatively maternally expressed genes in the mouse placenta.

1.5 Candidate Genes for Differential Methylation

One aspect of this project is to exploit the existing methylation data from numerous tissues to identify genes that are candidate DMRs, and investigate those DMRs initially in human blood, which is an accessible tissue that has a well phenotypes range of cell types. Regions that show differential methylation between alleles may have important roles in development, and understanding the role of DMRs in a normal tissue allows for better understanding of their role in specialised tissues like the placenta. In order to select candidate differentially methylated regions (DMRs), this study took advantage of the observation that these regions a) contain a high density of CpG islands and b) have a tendency to show 50% methylation overall. The regions of the genome with the highest density of CpGs tend to be located within 1.5 kb upstream of genes (i.e. promoter regions). DMRs are usually densely methylated on the repressed allele, and
unmethylated on the expressed allele. Using this knowledge, it is possible to look at the array data from other studies to identify candidate DMRs. Martín-Subero et al. (2009) published a comprehensive list of methylation in 767 genes as a part of their study of haematological neoplasms. Along with their neoplastic samples, they also included a number of normal haematological cell types as controls. They used the GoldenGate Methylation Cancer Panel I, which is an array enriched for oncogenes, TSGs, previously reported differentially methylated genes, signalling pathway genes, imprinted genes, DNA repair, cell cycle, metastasis, differentiation and apoptosis (Martín-Subero et al., 2009). From this dataset, it was possible to extract candidate regions of differential methylation for further investigation based on promoter methylation and previous reports of differential methylation in humans and other mammals (Morison et al., 2001).

1.5.1 IGFBP1

1.5.1.1 Function and Expression of IGFBP1

Insulin-like growth factor binding protein 1 (IGFBP1) is expressed by the liver, decidua, kidney and amniotic fluid and then secreted into the plasma, where it circulates the body, binding to IGF-I and IGF-II. Binding of this protein prolongs the half-life of the IGFs and alters their interaction with cell surface receptors, either promoting or inhibiting the growth promoting effect of IGFs. The protein also contains a thyroglobulin type-1 domain, thought to be involved in proteolytic degradation (UniProt). It has been seen to promote cell migration in cultures. The gene itself comprises 4 exons spanning a 5 kb region. At least 3 alternate transcriptional splice variants have been observed, encoding different isoforms. These different isoforms may have different functions in different tissues. Two of the variants have only been seen in placental tissues (Ace-View, NCBI Thierry-Mieg & Thierry-Mieg (2006)). In the rat, Igfbp1 mRNA is abundant from implantation through the proliferative phase, but drops off sharply late in gestation (Shynlova et al., 2007). Expression of the IGFBP1 gene is similarly very im-
important in the human placenta (Zhou & Bondy, 1992). Specifically, altered expression of IGFBP1 has been linked to IUGR and pre-eclampsia (Crossey et al., 2002; Koutsaki et al., 2011; Stone et al., 2003). Excess concentrations of IGFBP1 in maternal and fetal serum can contribute to IUGR (Fowler et al., 2000).

1.5.1.1 Evidence For Differential Methylation of IGFBP1
IGFBP1 exhibits intermediate levels of methylation in both the promoter and exon 1 sites interrogated across normal haematopoietic tissues in the Martín-Subero et al. (2009) array data, averaging 46% methylation across 31 normal hematopoietic tissues at both locations. This pattern of methylation suggests that one allele may be more methylated than the other. Seoighe et al. (2006) used similar data from Expressed Sequence Tags (ESTs) and SNPs to model potential causes of unequal representation of alleles in expression data and IGFBP1 fitted the imprinting model as well or better than known imprinted genes such as H19, DLK1 and SNURF. The orthologue in the mouse is located near a densely imprinted region on the proximal arm of chromosome 11 containing Grb10, the human orthologue of which is also suspected to be maternally imprinted. The gene is located in a region of chromosome 7 that has been linked to Silver-Russell Syndrome, a rare form of dwarfism. IUGR disorders such as Silver-Russell Syndrome act as support for the parental conflict hypothesis; maternally expressed genes are involved in conservation of resources for mother and a duplication of maternally expressed genes leads to an IUGR phenotype such as is seen in SRS (Tabano et al., 2010; Yuen & Robinson, 2011). Based on a microarray-based methylation study of placental triploidies, Yuen et al., also identified IGFBP1 as a putative DMR (Yan et al., 2009; Yuen & Robinson, 2011).

1.5.1.2 Silver-Russell Syndrome
SRS is an intrauterine growth restriction disorder characterised by a small body size at all stages of development, poor postnatal growth, craniofacial abnormalities and body asymmetry. The prevalence is around 1 in 100,000 among individuals with no fam-
ily history of the disorder. The main known cause of Silver-Russell Syndrome is hypomethylation due to a familial or de novo imprinting defect at 11p15. The syndrome is not inherited in a simple manner, as discordance in has been seen among twins and triplets, where only one of the children displays the Silver-Russell Syndrome phenotype. Abnormalities of chromosomes 7, 8, 15, 17, and 18 have been associated with Silver-Russell Syndrome, but in around 10% of cases not related to 11p15, the patient exhibits maternal uniparental disomy (UPD) of chromosome 7. UPD describes the chromosomal abnormality that occurs when both copies of the chromosome are inherited from one parent. Additionally, several cases involving a maternally inherited duplication of 7p11.2-p13 in children with characteristics of SRS indicate a dosage effect in this region (Monk et al., 2002). This region also contains GRB10, a known imprinted gene. If a link can be established between the methylation status of IGFBP1 and Silver-Russell Syndrome this will provide further evidence for the role of aberrant methylation in the development of this and other intrauterine growth disorders.

1.5.2 ZNF264

1.5.2.1 Function and Expression of ZNF264

ZNF264 is a Kruppel-type zinc-finger gene that encodes a protein with two Kruppel-associated box motifs, and a 13-unit zinc finger domain (Kim et al., 2001). Zinc finger proteins are able to bind nucleic acids, and therefore tend to be involved in regulation of transcription. Little else is currently known about the specific function of this gene and its protein (UniProt Consortium, 2015). ZNF264 is expressed ubiquitously across most cell types, with highest expression of mRNA in the testis. It consists of four exons, spanning a 30 kb genomic region, and has multiple splice variants, including one that is most commonly found in the placenta (Figure 1.9)(AceView, NCBI Thierry-Mieg & Thierry-Mieg (2006)).
Figure 1.9: Known protein coding variants of ZNF264. The transcript found most commonly in placenta is marked with an arrow.

1.5.2.1.1 Evidence For Differential Methylation of ZNF264

ZNF264 shows intermediate methylation (60-70% across cell types) in the promoter region interrogated by Martín-Subero et al. (2009). The exon 1 site interrogated showed low levels of methylation. The mouse orthologue Zfp264 shows paternal allele-specific expression in the neonatal and adult brain. In the mouse, the neighbouring gene Zim3 is transcribed primarily from the maternal allele, showing expression in adult testis whilst Zfp264 is paternally expressed (Kim et al., 2001), indicative of the reciprocal grouping of imprinted genes. The expression patterns of ZIM3 and ZNF264, and the rest of the PEG3/USP29 domain are largely conserved between mouse, cow and human (Huang & Kim, 2009). Both of these genes have become pseudogenes in the mouse, however, and no longer encode functioning zinc finger proteins. ZNF264 was identified in a study to identify age-dependent DMRs as showing markedly different methylation between fetal and adult samples, particularly in the brain. A 50% erosion of methylation was observed over time for CpGs associated with ZNF264 from ~50% in fetal tissues to ~5% in adult tissues (Yuen et al., 2011b).
1.6 Aims

This project aims to investigate the epigenetic plasticity of the placenta on a genome-wide level and increase our understanding of the epigenetics behind pre-eclampsia. Examining methylation patterns can reveal differences in epigenetic control from a tissue-, time- and parental-specific perspective. This project also aims to exploit that ability to aid in our understanding of differential methylation, and to identify novel DMRs.

1. To analyse the two potential DMRs *IGFBP1* and *ZNF264*
   
a. In the context of a normal tissue (Whole Blood) in a normal cohort
   
b. In the context of both normal and Pre-Eclamptic Placentas

2. To generate an Epigenetic Profile of the Placenta using RRBS

3. To develop an analytical tool for RRBS data

4. To Determine Whether Pre-eclampsia Shows a Distinct Epigenetic Landscape
   
a. To identify differential methylation between normal and dysfunctional (Pre-Eclamptic) placentas

   i. Methylation of specific candidate genes will then be quantified by Sequenom and confirmed in a validation cohort
1.6.1 Hypotheses

1. That IGFBP1 and ZNF264 display differential methylation in an allele-dependent manner

2. That pre-eclamptic placentas show an epigenetic profile distinct from that of normal placentas

3. That epigenetic differences between the human placenta and somatic tissues underlie the functional requirements of the placenta for optimal growth and development

4. That epigenetic changes in imprinted genes/DMRs are observed in pre-eclampsia/toxaemia

---

**Figure 1.10:** A diagrammatic representation of how the hypotheses addressed in this thesis are systematically and thematically linked. LOI - Loss of Imprinting, GOI - Gain of Imprinting, PET - Pre-eclampsia/Toxaemia
Chapter 2

Materials and Methods

From a methodical perspective, there are two distinct components to the research described in this thesis- the ‘wet lab’, or molecular biology based methods; and the ‘dry lab’ or bioinformatics and biostatistics computer based methods. This chapter will cover the former, or laboratory based techniques. A wide range of techniques were utilised, including PCR, cloning, and sequencing for analysis of imprinting and allele-specific methylation. RRBS library preparation involves a multi-day, multi-step protocol that incorporates many different tools, and the use of Sequenom MassArray® for validation of candidate differentially methylated sites is also described.

2.1 Sourcing of Samples

2.1.1 Ethical Approval

Blood samples were collected and used under approval from the University of Otago Human Ethics Committee for the project “Allele Specific Methylation and Parent of Origin Effects - Discovering Novel Imprinted Genes.” Human Ethics Committee. 28 April 2011. Reference Code 11/092. Approval was given to collect a small (10 mL) blood sample from a peripheral vein, and to request blood samples from the parents of the probands.
The placental tissue in this study is from the Otago Placenta Study and was previously collected under the approval of the Lower South Regional Ethics Committee (project key: LRS10/09/38). More details regarding the placental collection can be found in Section 2.5.

2.1.2 Collection of Blood Samples

A total of 46 blood samples from individuals of New Zealand/European descent were collected from three sources:

- 17 whole blood samples obtained via venipuncture. Blood was collected into BD Vacutainers® spray-coated with K$_2$EDTA.
- The Merriman Lab, Department of Biochemistry, University of Otago provided the remainder of the samples. These samples were collected as part of two previous genetic-based family studies. The individuals concerned provided their consent for their DNA to be used in other approved studies.
  - 7 samples of DNA from two families, from the New Zealand Insulin-Dependent Diabetes Cohort
  - 21 samples from 7 families from the Thyroid Genetics Study

2.1.3 Processing of Placental Samples

2.1.3.1 Dissecting Placental Samples

Placental samples were collected and processed by Dr. Noelyn Hung and Dr. Tania Slatter. Samples were taken as 3 x 5 x 1 cm cross-sections, which were put into cassettes and stored at -80 °C. For this study, a small sub-section (3 x 0.5 x 0.5 cm) was taken from the original frozen sample and stored separately at -80 °C. Further processing was performed only on this sub-section of the sample.

A 0.5 x 0.5 cm piece of tissue was sliced from the sub-section using a razor blade. To gain a more representative cell population for DNA extraction, this smaller
portion was taken from the centre of the original cross-section. As these samples are placental tissue, they are suffused with maternal blood. To minimize contamination from this source, samples were washed in a petri dish in PBS for 5 min. The tissue was gently disrupted with sterile tweezers to remove as much maternal and fetal blood as possible. Samples were rinsed again in PBS, then drained of PBS and transferred into a 2 mL microcentrifuge tube for DNA extraction. A small number of samples had been preserved in formalin prior to freezing. This was used to preserve histology for other analyses. For these samples, an additional wash with PBS was performed to remove excess formalin. While an attempt to construct libraries from these formalin-fixed samples was made, the DNA proved too fragmented to reliably construct RRBS libraries from.

2.2 DNA Extraction

2.2.1 Peripheral Blood

DNA was extracted from whole blood samples using a Qiagen QIAamp® DNA Mini and Blood Mini Kit. The spin protocol for DNA Purification from Blood or Body Fluids was utilized. To begin, 20 µL of proteinase K was added to 200 µL of whole blood to lyse both the cells, and the protein component of the blood sample. Then 200 µL of buffer AL was added to the sample and incubated at 56 °C for 10 min followed by the addition of 200 µL of 100% ethanol was then added to the solution. The resulting solution was then applied to a QIAamp Mini spin column in a 2 mL collection tube and centrifuged at 6000 x g for 1 min. The filtrate was discarded while the DNA remained bound to the filter. For the first wash step 500 µL of Buffer AW1 was then added and centrifuged at 6000 x g for 1 min before discarding the filtrate. A second wash step, using Buffer AW2, was then performed and the filtrate again discarded. Pure DNA was then eluted using 200 µL of Buffer AE. Eluted DNA was stored at 4 °C, and the remainder of the blood samples stored at -20 °C in case
further extractions needed to be undertaken.

2.2.2 Placental Tissue

DNA extractions were performed using the QIAamp DNA Mini Kit (Qiagen Cat. #51306) with some small modifications to the manufacturer's protocol. All reagents used at this step were supplied with the kit, with the exception of the Proteinase K, which we found to have a lower efficacy compared to Proteinase K prepared in our laboratory from stock supplied by Invitrogen (Life Technologies Cat #25530-015). The tissue was weighed for each sample to determine the input, which averaged 80 mg. Next, 180 µL of Buffer ATL was added to the microcentrifuge tube containing the sample, followed by 25 µL of Proteinase K at 10 mg/mL. The resulting mixture was vortexed, and then placed on a 56 °C Eppendorf thermomix at 500 rpm overnight. After lysis was complete, tubes were briefly centrifuged, and 200 µL of Buffer AL added. The sample was vortexed, then incubated at 70 °C for 10 min. Following this, 200 µL of 100% ethanol was added to the sample and mixed by pulse-vortex for 15 secs. The resulting solution was applied to a QIAamp Mini spin column and centrifuged at 16,000 x g for 1 min. The tube containing the filtrate was discarded, and the spin column transferred to a clean 2 mL collection tube. DNA was then split into two aliquots, each of 200 µL, to serve as a working sample, and a stock sample. Finally, 150 ng of whole genomic DNA was run on a 2% agarose gel to check integrity and degree of RNA contamination. DNA was stored at 4 °C.

2.2.3 Quantification of DNA

The concentration and quality of the DNA was quantified by applying 1 µL to a NanoDrop ND-1000 Spectrophotometer
2.3 Genotyping of Blood Samples

2.3.1 Genotyping via SNP Analysis

SNP analysis was used to identify informative individuals. SNPs in both the Promoter and Exon 1 regions of *IGFBP1* and *ZNF264* were identified using the UCSC Genome browser (Fujita *et al.*, 2011) and validated for use in this population using dbSNP build 132 (Sherry *et al.*, 2001) (http://www.ncbi.nlm.nih.gov/snp). Informative individuals were defined as those who were heterozygous at one or more of the SNP sites listed in Table 2.1. Using heterozygous individuals enables us to distinguish the two alleles, and analyse their methylation patterns according to both SNP allele and parent-of-origin for each copy. As two regions of the CpG Island for each gene were being interrogated, informative SNPs for each region had to be identified, to differentiate between the alleles in each sample. Genotyping was undertaken to identify individuals that were heterozygous for the SNPs shown in Table 2.1.

Table 2.1: SNPs used for genotyping samples.

<table>
<thead>
<tr>
<th></th>
<th>IGFBP1</th>
<th>ZNF264</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter SNP</td>
<td>rs1065780 (H=0.492)*</td>
<td>G/A</td>
</tr>
<tr>
<td>Exon 1 SNP</td>
<td>rs9658194 (H=0.302)*</td>
<td>A/C</td>
</tr>
<tr>
<td></td>
<td>[rs58278481** (H=0.295)]</td>
<td></td>
</tr>
</tbody>
</table>

* Heterozygosity data obtained from dbSNP.

** This SNP was originally interrogated, but found to be non-informative in this sample group, so rs7250511 was used thereafter.

2.3.1.1 Primer Design

Primers for genotyping were designed to span the SNP and the surrounding CpG sites to be included for methylation analysis. The primers were designed using Primer3, an open source primer design software available at http://frodo.wi.mit.edu/primer3/
(Skaletsky & J., 2000) and manufactured by Integrated DNA Technologies (IDT, Leu-
van, Belgium).

Table 2.2: Primer sequences for genotyping PCRs

<table>
<thead>
<tr>
<th>Gene Target</th>
<th>Amplicon Size</th>
<th>Primer Name</th>
<th>ID</th>
<th>Tm (°C)</th>
<th>Primer Sequence 5-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGFBP1 Promoter</td>
<td>702</td>
<td>gIGFBP1_P_F1</td>
<td>4269</td>
<td>56</td>
<td>AGA GAA AAC CTT TGC ATT TGC T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gIGFBP1_P_R1</td>
<td>4270</td>
<td></td>
<td>CTT GCA CCA GGA GGT TAA TGA T</td>
</tr>
<tr>
<td>IGFBP1 Exon 1</td>
<td>650</td>
<td>gIGFBP1F.2</td>
<td>4192</td>
<td>56</td>
<td>GGT ACT GCT CCT GCT GAC TG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gIGFBP1R.2</td>
<td>4193</td>
<td></td>
<td>GCC TCC CTC TAC TCT TGC TG</td>
</tr>
<tr>
<td>ZNF264 Promoter</td>
<td>636</td>
<td>gZNF264_P_F1</td>
<td>4273</td>
<td>56</td>
<td>GTC CAA AGA AGG CTC AAA AGA G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gZNF264_P_R1</td>
<td>4274</td>
<td></td>
<td>ATT CCT ATC CTA CTC CCC TGC T</td>
</tr>
<tr>
<td>ZNF264 Exon 1</td>
<td>829</td>
<td>gZNF264F</td>
<td>4189</td>
<td>54</td>
<td>GCA GGG GAG TAG GAT AGG AAT C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gZNF264R.3</td>
<td>4195</td>
<td></td>
<td>GAT GGC TAG AAC ATC CAG GAA A</td>
</tr>
</tbody>
</table>

2.3.1.2 PCR Conditions

The following 20 µL standard PCR mixes were used for genotyping PCRs unless otherwise stated:

Table 2.3: PCR reaction mixes for genotyping PCRs. For negative controls, MilliQ H₂O was sub-
stituted for DNA

<table>
<thead>
<tr>
<th>PCR reaction buffer, 10x</th>
<th>2 µL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MgCl₂ (25 mM)</td>
<td>1.2 µL</td>
</tr>
<tr>
<td>dNTPs (2.5 mmol/L)</td>
<td>1 µL</td>
</tr>
<tr>
<td>Forward Primer (10 µM)</td>
<td>1 µL</td>
</tr>
<tr>
<td>Reverse Primer (10 µM)</td>
<td>1 µL</td>
</tr>
<tr>
<td>DMSO</td>
<td>2 µL</td>
</tr>
<tr>
<td>FastStart Taq DNA Polymerase, 5 U/µL</td>
<td>0.2 µL</td>
</tr>
<tr>
<td>MilliQ H₂O</td>
<td>10.6 µL</td>
</tr>
<tr>
<td>Template genomic DNA (concentration varies)</td>
<td>1 µL</td>
</tr>
</tbody>
</table>
2.3.1.2.1 Optimisation of PCR
All PCRs were optimised for annealing temperature on a BioRad DNAEngine® Peltier Thermal Cycler using a temperature gradient protocol. In addition, PCRs were optimised for magnesium chloride conditions when appropriate.

2.3.1.2.2 PCR Thermocycling Conditions
The standard cycling conditions for PCR were as follows in Table 2.4 (with variations in annealing temperature for each separate reaction).

Table 2.4: PCR Conditions for genotyping

<table>
<thead>
<tr>
<th>Time</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 min</td>
<td>95°C</td>
</tr>
<tr>
<td>30 s</td>
<td>95°C</td>
</tr>
<tr>
<td>30 s</td>
<td>56°C</td>
</tr>
<tr>
<td>45 s - 3min</td>
<td>72°C</td>
</tr>
<tr>
<td>7 min</td>
<td>72°C</td>
</tr>
<tr>
<td>∞</td>
<td>4°C</td>
</tr>
</tbody>
</table>

2.3.1.3 Visualisation of PCR Products
PCR products were electrophoresed on 2% agarose gels. The gels were prepared by dissolving 2% agarose LE in 1x TAE Buffer containing ethidium bromide (10 mg/mL). Then, 10 µL of PCR product was combined with 2 µL of xylene cyanol loading dye, and loaded onto the gel. In addition, 8 µL of 1 Kb Plus DNA Ladder™ was loaded onto each gel as a size marker. Gels of 12-48 wells were run for up to 40 min at 100V in 1x TAE Buffer containing ethidium bromide (10 mg/mL). The products were then visualised and photographed under UV light using the BioRad GelDoc™ XR+ Transilluminator with QuantityOne v. 4.6.5 software.
2.3.1.4 Genotyping via RFLP Analysis

Initially, for the exon 1 regions of both *IGFBP1* and *ZNF264*, genotyping was attempted using restriction enzyme digest to distinguish the two alleles. SNP Cutter (*Zhang et al.*, 2005) was used to identify restriction enzymes that would preferentially cut one allele at the SNP site.

*IGFBP1*

BsrDI (NEB, Catalogue #R0574L). This enzyme should cut when the rs9658194 SNP allele of *IGFBP1* is C, but not when it is A. Therefore, a genotype of CC should yield fragments of 504 and 146 bp; AA should give one fragment of 650 bp; and AC should give three fragments of 650, 504 and 146 bp. The cut site is asymmetric at the following location:

5’ ...GCAATG NN↓... 3’
3’ ...CGTTAC↑NN ... 5’

*ZNF264*

Bsu36I (NEB, Catalogue #R0524L). This enzyme should cut when the rs58278481 SNP allele of *ZNF264* is C, but not when it is G. Therefore, a genotype of GG should yield fragments of 652 and 177 bp; CC should give three fragments of 516, 177, and 136 bp; and GC should give fragments of sizes 652, 177, and 136 bp. The cut site is asymmetric at the following location:

5’ ...CC↓TNA GG... 3’
3’ ...GG ANT↑CC... 5’
2.3.1.4.1 Reaction Mix for Restriction Enzyme Digest

Table 2.5: Restriction digest conditions for genotyping via RFLP

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10x Reaction Buffer</td>
<td>2 µL</td>
<td></td>
</tr>
<tr>
<td>(Buffer 3 for Bsu36I/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buffer 2 for BsrDI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bovine Serum Albumin (BSA)</td>
<td>0.2 µL</td>
<td></td>
</tr>
<tr>
<td>MilliQ H₂O</td>
<td>9.3 µL</td>
<td></td>
</tr>
<tr>
<td>Restriction Enzyme</td>
<td>0.5 µL</td>
<td></td>
</tr>
<tr>
<td>PCR Product</td>
<td>8 µL</td>
<td></td>
</tr>
</tbody>
</table>

Incubation was for 12-24 hours. Temperature conditions differed for the two enzymes:
Bsu36I - 37°C
BsrDI - 65°C
Digested PCR products were visualised on a 2% agarose gel run for 45 min at 100V.

2.3.2 Sanger Sequencing

Sanger sequencing was used to produce sequence chromatograms for genotyping purposes, and for analysing clones of bisulphite converted DNA. Initially, genotyping was performed on PCR products to identify those samples that were heterozygous at one or more of the SNPs tested (Table 2.1).

2.3.2.1 Purification of PCR Products for Sequencing

After the PCR products were visualised via gel electrophoresis, the remainder (10 µL) of the PCR product was purified using the Zymo DNA Clean & Concentrator™-5 according to the recommended protocol. Briefly, 5 volumes of DNA Binding Buffer were added to 1 volume of PCR product. The mixture was then loaded onto a Zymo-Spin Column (and 2 mL collection tube). The tube was centrifuged at 14000 rpm for 30 s and the flow through discarded. Next, 200 µL of Wash Buffer was added to the column and again centrifuged at 14000 rpm for 30 sec. This wash was repeated. The
column was then transferred to a clean 1.5 mL microcentrifuge tube and eluted in 10 µL MilliQ H₂O. The elution was then reapplied to the column to increase the yield. The concentration and quality of the DNA was quantified by applying 1 µL to a NanoDrop ND-1000 Spectrophotometer.

### 2.3.2.1.1 BigDye® Sequencing Reaction

A standard 10 µL reaction was as follows in Table 2.6. The sequencing reaction was thermally cycled using a BioRad DNAEngine® Peltier Thermal Cycler as laid out in Table 2.7.

**Table 2.6: BigDye® reaction mix conditions**

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigDye® v3.1 (or v3.0 dGTP)</td>
<td>1 µL</td>
</tr>
<tr>
<td>5x Sequencing Buffer</td>
<td>0.8 µL</td>
</tr>
<tr>
<td>Primer (1 pmol/µL)</td>
<td>2 µL</td>
</tr>
<tr>
<td>Template sequence</td>
<td>X µL (according to concentration)*</td>
</tr>
<tr>
<td>MilliQ H₂O</td>
<td>To 10 µL</td>
</tr>
</tbody>
</table>

* 1 ng/100 bp/ 10 µL reaction for single-stranded PCR products. 250 ng/ 10 µL reaction for double-stranded Plasmid DNA.

**Table 2.7: BigDye® Cycling Conditions**

<table>
<thead>
<tr>
<th>Step</th>
<th>Conditions</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denaturation</td>
<td>96°C for 1 min</td>
<td>x 1 Cycle</td>
</tr>
<tr>
<td>Thermal Cycling</td>
<td>96°C for 30s</td>
<td>x 25 Cycles</td>
</tr>
<tr>
<td></td>
<td>50°C for 15s</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60°C for 4 min</td>
<td></td>
</tr>
</tbody>
</table>

### 2.3.2.1.2 Ethanol/Sodium Acetate Precipitation Protocol

An ethanol/sodium acetate precipitation protocol was used to further purify the sequencing products. For each 10 µL reaction 31.25 µL 100% ethanol, 1.5 µL 3M NaOAc (pH 4.6), and 7.25 µL MQ H₂O was added, and the resultant mixture incubated in the
dark at room temperature for 30 min. the tubes were then spun in an Eppendorf 5810R centrifuge at 3000 rpm for 45 min at room temperature. The supernatant was removed without disturbing the DNA pellet by blotting onto a paper towel and then gently spinning the tubes upside down at 500 rpm for 1 min. The remaining pellet was washed by adding 150 µL of 70% Ethanol to the tubes, and the mixture spun at 3000 rpm for a further 10 min. The supernatant was again removed by blotting and gentle spinning upside down. Samples were air dried in the dark for 20 min. Samples were then stored at 4°C in the dark until sent for sequencing. Samples were sent to the Genetic Analysis Services (Department of Anatomy, University of Otago, Dunedin) for sequencing on the ABI 3730 xl DNA Analyser.

2.3.2.2 Preliminary Sequence Analysis

Sequence reads were returned electronically in the form of chromatograms, which were viewed on 4Peaks (Griekspoor, 2005) and Geneious v.5.5 and 6.0 (Kearse et al., 2012). Samples that were heterozygous at any of the SNPs were identified as those possessing two peaks at the SNP site.

2.4 Methylation of Candidate Imprinted Genes

Methylation analysis was performed only on those samples identified as being heterozygous for one or more of the SNPs. Methylation analysis was performed by the sequencing of cloned PCR products from bisulphite-converted DNA. Bisulphite conversion is based on the ability of bisulphite to deaminate cytosines to uracils. For methycytosines, however, the methyl group prevents this conversion from occurring. When sequenced, any converted uracil residues will be read as thymine, while methylated cytosines will still be read as cytosines (Frommer et al., 1992). Regions of interest can be analysed for methylation patterns by PCR of bisulphite converted DNA followed by sequencing. Bisulphite conversion causes problems with PCR amplification and thus imposes limitations on the size of the amplicons. The bisulphite treatment has a
tendency to degrade and fragment the original genomic DNA (Tusnady et al., 2005). Secondly, the inherent GC rich nature of the regions investigated results in homopolymeric runs of uracils that cause difficulty for the DNA Polymerase (Grunau et al., 2001). For the above stated reasons, PCR amplification of bisulphite treated DNA is commonly limited to amplicons of fewer than 350 bp (Tusnady et al., 2005). To identify allele-specific (or parent-of-origin) methylation, the PCR products must be cloned before sequencing.

2.4.1 Bisulphite Conversion

The EZ DNA Methylation™ Kit from Zymo was used for bisulphite conversion. Conversion was performed following the manufacturer’s protocol, with a starting DNA volume of 15 µL. The DNA was first diluted with 5 µL of M-Dilution Buffer, and made up to 50 µL with MQ H₂O. This diluted DNA was then incubated at 37°C for 15 min. To the above solution was added 100 µL of CT Conversion Reagent, and the resultant mix was incubated overnight (12-16 hours) in the dark at 50°C. After incubation, the sample was loaded, along with 400 µL of M-Binding Buffer, onto a Zymo-Spin™ IC Column (in a 2 mL collection tube). The sample was centrifuged for 30 s at 12000 rpm and the flow-through discarded. Then 100 µL of M-Wash Buffer was added to the column and centrifuged. The flow-through was again discarded. 200 µL of M-Desulphonation Buffer was added to the column and incubated at room temperature for 15-20 min. After incubation, the column was centrifuged and the flow-through discarded. Another two washes were performed, with 200 µL of M-Wash Buffer each time. After the second centrifugation, the column was placed into a clean 1.5 mL microcentrifuge tube and 10 µL of M-Elution Buffer applied to the filter. The column was allowed to stand for 1 min at room temp and then centrifuged for 30 s at full speed. The eluate was reapplied to the filter and centrifuged again to optimise the yield. Bisulphite converted DNA was stored at -30°C for up to 6 months.

Nanodrop analysis was performed to ensure that the rate of attrition from bisulphite
conversion did not significantly affect the DNA concentration.

2.4.2 PCR Amplification of Bisulphite Treated DNA

After bisulphite conversion, PCR amplification is necessary to generate enough product for the cloning reaction, and to visually quantify the length and quality of the bisulphite converted PCR product via gel electrophoresis.

2.4.2.1 Design of Primers for Bisulphite Converted DNA

Primers were designed with the aid of openly available online tools, namely Meth-Primer (Li & Dahiy, 2002), and BiSearch (Tusnady et al., 2005). However, both of these tools were limited in their function, particularly when attempting to design across a CpG-dense rich region. As a result, some primers were adjusted by hand. Primer 4204 for Exon 1 of *IGFBP1* in particular had to be designed to include a CpG site; this is denoted as Y in the oligonucleotide to allow for the possibility that this CpG may or may not be converted. The ePCR function of BiSearch, and the methBLAST online tool (Pattyn et al., 2006) were both used to validate the primers, and check for sequence specificity.

| Table 2.8: Primers for amplification of bisulphite converted DNA |
|----------------------|------------------|-----------------|------|------------------|
| Gene Target          | Amplicon Size    | Primer Name     | ID   | Tm (°C)          | Primer Sequence 5’-3’                           |
| *IGFBP1* Promoter    | 329              | mIGFBP1_P_F1    | 4267 | 51               | TTA TTT GGG GTA TTG TTT TTT G                  |
|                      |                  | mIGFBP1_P_R1    | 4268 |                 | ATA AAT TTA TTT TAC TAA TAC ACC CA             |
| *IGFBP1* Exon 1      | 317              | mIGFBP1F3       | 4204 | 49               | GTA ATT TTT GTA YGT TTT TAT T                  |
|                      |                  | mIGFBP1Rb       | 4183 |                 | TAT ACC CTA AAA CAT CTT CCT TCT T              |
| *ZNF264* Promoter    | 328              | mZNF264_P_F1    | 4271 | 51               | TTT GTG TTT TTT AAG ATG ATG ATT                |
|                      |                  | mZNF264_P_R1    | 4272 |                 | AAT ATA ATT TCC TCT AAT TAT TCA AC             |
| *ZNF264* Exon 1      | 295              | mZNF264F        | 4184 | 49               | GGG GTT TTT TTT TAG TTT TG                     |
|                      |                  | mZNF264Ra       | 4185 |                 | CCA AAA AAA ACA TCA TTC ACT CC                 |
2.4.2.2 PCR Conditions

The following 20 µL standard PCR mixes (Table 2.9) were used for PCRs on bisulphite converted DNA unless otherwise stated. For negative controls, MilliQ H₂O was substituted for DNA template. For the ZNF264 Exon 1 primer pair (4184 and 4185), 2 µL of DMSO was added (H₂O reduced to 9.6 µL)

<table>
<thead>
<tr>
<th>PCR reaction buffer, 10x</th>
<th>2 µL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MgCl₂ (25 mM)</td>
<td>1.2 µL</td>
</tr>
<tr>
<td>dNTPs (2.5 mmol/L)</td>
<td>1 µL</td>
</tr>
<tr>
<td>Forward Primer (10 µM)</td>
<td>1 µL</td>
</tr>
<tr>
<td>Reverse Primer (10 µM)</td>
<td>1 µL</td>
</tr>
<tr>
<td>FastStart Taq DNA Polymerase, 5 U/µL</td>
<td>0.2 µL</td>
</tr>
<tr>
<td>MilliQ H₂O</td>
<td>11.6 µL</td>
</tr>
<tr>
<td>Template bisulphite converted DNA</td>
<td>2 µL</td>
</tr>
</tbody>
</table>

The standard cycling conditions for PCR were as follows in Table 2.10

<table>
<thead>
<tr>
<th>Time</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denaturation/Activation</td>
<td>4 min 95°C</td>
</tr>
<tr>
<td>Denaturation</td>
<td>30 s 95°C</td>
</tr>
<tr>
<td>Annealing</td>
<td>30 s 51°C for Promoter PCRs</td>
</tr>
<tr>
<td></td>
<td>49 °C for Exon 1 PCRs</td>
</tr>
<tr>
<td>Elongation</td>
<td>45 s - 3 min 72 °C</td>
</tr>
<tr>
<td>Extension</td>
<td>7 min 72 °C</td>
</tr>
<tr>
<td>Cooling</td>
<td>∞ 4 °C</td>
</tr>
</tbody>
</table>
2.4.3 Cloning of PCR Products for Methylation Analysis

Before cloning, the PCR product was purified to remove any trace of primer-dimer using the DNA Clean & Concentrator™-5 kit described in Section 2.3.2.1.

2.4.3.1 Cloning (Topo TA Kit)

Cloning of the bisulphite converted DNA products was performed using the TOPO TA Cloning® Kit for Sequencing (Invitrogen) and OneShot® Mach1™T1R Chemically Competent E. coli cells. 2 µL of the purified PCR product was mixed with 0.5 µL of the plasmid vector pCR® 4-TOPO® and 0.5 µL of Salt Solution (1.2 M NaCl and 0.06 M MgCl2). This mixture was incubated for 5 min at room temperature. From the above mix, 1 µL was added to 25 µL of MACH1 competent cells. The cells were incubated on ice for 5 min. The cells were heat-shocked for 30 s in a 42°C waterbath, then immediately put on ice. For the initial incubation 125 µL of S.O.C Medium was added to the cells, which were then shaken horizontally at 37°C for 1 hour. Cells were spread onto pre-warmed (37°C) LB Agar plates containing either 100 µg/mL Ampicillin (for early experiments) or 50 µg/mL Kanamycin (all later experiments). Two plates, one of 100 µL and one of 50 µL, were spread for each transformation. The plates were incubated at 37°C overnight.
2.4.3.2 Screening Transformants for Inserts

Screening was performed on the overnight colonies by one of two methods. Initially, screening was performed after isolation of the plasmid by restriction digest, but as this was rather inefficient in terms of time and yield, a switch was made to analysing transformants by PCR.

2.4.3.2.1 Screening by PCR

This protocol utilises the primer sites located either side of the insertion site to amplify the insert and thus know whether the size of the insert present is the same as the insert desired. A sterile toothpick was used to select a colony, and the cells resuspended in the solution outlined in Table 2.11

<table>
<thead>
<tr>
<th>Table 2.11: PCR reaction for screening of E.coli clones for inserts</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR reaction buffer, 10x</td>
</tr>
<tr>
<td>MgCl2 (25 mM)</td>
</tr>
<tr>
<td>dNTPs (2.5 mmol/L)</td>
</tr>
<tr>
<td>M13 Forward Primer (10 µM)</td>
</tr>
<tr>
<td>M13 Reverse Primer (10 µM)</td>
</tr>
<tr>
<td>FastStart Taq DNA Polymerase, 5 U/µL</td>
</tr>
<tr>
<td>MilliQ H$_2$O</td>
</tr>
</tbody>
</table>

A patch plate was also made in order to preserve the colonies for further analysis. This was incubated overnight at 37°C. The PCR was performed on a thermocycler using the cycles outlined in Table 2.10, with an annealing temperature of 50°C, and the results visualised by gel electrophoresis as outlined in Section 2.3.1.3. Those samples with the correct size amplimer (insert size plus 167 bp from the plasmid) were identified for plasmid extraction and further analysis.
2.4.3.2 Screening by EcoRI digestion

Restriction digest with EcoRI can also be used to screen the colonies for the correct insert after miniprep. The enzyme cuts on either side of the insert, and the results can be visualised by gel electrophoresis. The digestion mix is as follows in Table 2.12 and the digestion mix was incubated overnight at 37°C.

Table 2.12: EcoRI digestion of plasmids to screen for correct inserts

<table>
<thead>
<tr>
<th>Volume (µL)</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Plasmid DNA</td>
</tr>
<tr>
<td>2</td>
<td>10 x EcoRI Buffer</td>
</tr>
<tr>
<td>13</td>
<td>MQ H₂O</td>
</tr>
<tr>
<td>0.5</td>
<td>EcoRI Restriction Enzyme</td>
</tr>
<tr>
<td>20.5</td>
<td>Total</td>
</tr>
</tbody>
</table>

2.4.3.3 Plasmid Miniprep

Once colonies with the correct insert had been identified, they were selected using a toothpick and resuspended in 1 mL of LB Broth + Kanamycin (or Ampicillin) in a 2 mL microcentrifuge tube. These cells were then incubated at 37°C overnight with horizontal shaking. Before beginning the miniprep protocol, the overnight cultures were centrifuged at 12000 rpm for 5 min to pellet the cells. The supernatant was discarded, and the cells resuspended in 600 µL of LB Broth + antibiotic. The resuspended cells were grown for an additional 1-2 hours at 37 °C on a horizontal shaker. Plasmid minipreps were performed using the Zyppy™ Plasmid Miniprep Kit, with the protocol as according to the manufacturer. To lyse the cells, 100 µL of Lysis Buffer was added to the 600 µL of bacterial culture to lyse the cells, and mixed thoroughly. to stop the lysis reaction, 450 µL of cold Neutralisation Buffer was then added. The solution was then centrifuged at 12000 rpm for 4 min to pellet the cell debris. The supernatant was transferred to a Zymo-Spin IIN column (and 2 mL collection tube), and spun at
12000 rpm for 15 s. The flow-through was discarded. 200 µL of Endo-Wash Buffer was applied to the column and centrifuged for 15 s. 400 µL of Zyppy Wash Buffer was added to the column, and this was also spun through (12000 rpm for 30 s). The column was then transferred to a clean 1.5 mL microcentrifuge tube. 30 µL of Zyppy Elution Buffer was then applied directly to the matrix, and allowed to stand at room temperature for at least 1 min. Centrifugation for 15 s at 12000 rpm was used to elute the DNA. The eluent was then reapplied to the column and centrifugation repeated to give a high concentration of pure plasmid DNA. The samples were nanodropped to determine the concentration of plasmid DNA. The DNA was stored at 4 °C until either digested with EcoRI for screening (see below) or sent for sequencing.

2.4.4 Analysis of Clones

Clones that were found to have the correct insert length were then sequenced for methylation analysis. The plasmid DNA was either prepared for sequencing in the lab by the protocol shown in Section 2.3.2.1; or 1 µL of M13 Forward Primer (4004) was added to 150 ng of plasmid DNA was sent to the Genetic Analysis Service, and technicians there performed the sequencing reaction. Where possible, at least 12 clones for each sample of bisulphite-converted DNA were sequenced.

2.4.5 Methylation Analysis of Clones

The chromatogram sequence traces returned by the Genetic Analysis Service were checked for accuracy, and additional base calling was performed using Geneious Software (Kearse et al., 2012). The plasmid sequence was trimmed off, and the inserts aligned according to which sample they were derived from. The methylation patterns across clones for each sample, and each gene were analysed using a variety of online methylation analysis tools. BISMA (Rohde et al., 2010) is able to process a .fasta file of sequences compiled in Geneious, and compare and align it with the original (un-converted) sequence. It will then output the file in a graphical format, and perform an
analysis of percentage methylation at each site, and across each clone. These outputs were used to create a binary matrix of methylation at each CpG site, for each clone, and for each amplicon.

2.5 Collection and Phenotyping of Placental Samples

Accurate phenotyping of samples is crucial to any biological study. In human studies, due to the complicated and sensitive nature of the tissue, it can be difficult to obtain all of the information that would be desirable. Efforts were made in this study to match samples based upon factors that are most likely to influence the methylation of the DNA. Additional phenotype data will allow both more accurate matching based upon factors known to affect the risk of pre-eclampsia, such as ethnicity, previous history of pre-eclampsia, and clinical severity. The better the match between case and control, the more accurately the source of the difference between them can be identified. Collection of placentas was performed by Drs Noelyn Hung, Tania Slatter, and Celia Devenish, with the assistance of clinical staff and midwives. Gross dissections were performed by Drs Noelyn Hung and Tania Slatter. The samples used in this study were selected by myself and Dr Erin Macaulay from the larger OPuS placental cohort according to pre-eclampsia pathology, and matched control samples were identified from the same cohort.

2.5.1 Cohort Recruitment

All of our placental samples were obtained with consent from New Zealand donors as part of the OPuS (Otago Placental Study) collection. Placentas were processed within 24 hours of delivery, following refrigeration. Clinical data collected included maternal smoking history, maternal medical history, maternal first trimester body weight and height, index pregnancy parity and gravidity, gestational age, birth weight, and any obstetric and/or neonatal complications. For this study into the epigenetics of pre-
eclampsia/toxaemia (PET), placentas from individuals within the OPuS cohort with pre-eclampsia were selected. Suitable matched controls were identified from amongst the remainder of the OPuS collection based on gestation, followed by fetal sex, maternal smoking, and maternal BMI - all factors that may affect DNA methylation. For further information on the matching criteria, see Section 2.7.

2.5.2 Ethical Approval

Ethical approval was obtained from the New Zealand Health and Disability Lower South Regional Ethics Committee and from the Multiregional Ethics Committee (Reference Number LRS/09/09/038). All of the women provided written informed consent.

2.5.3 Placental Pathology

A trained examiner, blinded to the clinical history, assessed fresh or refrigerated placentas by standard pathologic means. Eight representative transmural sections within 50 mm of the umbilical insertion were dissected per placenta. H&E sections were prepared according to standard procedures, and examined by a perinatal pathologist (Slatter et al., 2014). After gross pathological assessment of placentas, three of the transmural sections were frozen at -70°C and later used for DNA extraction.

2.6 Pre-eclampsia Characterisation

The SOMANZ (Society of Obstetric Medicine of Australia and New Zealand) guidelines state that for a woman to be diagnosed as pre-eclamptic, she must exhibit new hypertension after the 20th week of gestation, and also have one or more of the following indications:

- Renal involvement, usually manifesting as significant proteinuria measured on a dipstick and/or by a spot protein/creatinine ratio reading. Other indications may include raised serum or plasma creatinine, and oliguria.
• Liver involvement as indicated by raised serum transaminases and/or severe epigastric or upper right quadrant pain
• Haematological involvement such as thrombocytopenia, hemolysis, or disseminated intravascular coagulation
• Neurological involvement, which becomes increasingly severe when progression to eclampsia occurs and seizures occur. Hyperreflexia, severe headaches, persistent visual disturbances and stroke are also significant in the diagnosis of Pre-eclampsia/eclampsia
• Pulmonary edema
• Fetal Growth Restriction
• Placental abruption

All Pre-eclamptic samples within our cohort were clinically diagnosed in accordance with these guidelines (Table 2.13), and samples from women with hypertension, but without the accompanying other symptoms were excluded for the purposes of this discovery study.

<table>
<thead>
<tr>
<th>OPuS Study #</th>
<th>Booking BP</th>
<th>Onset of PET/weeks</th>
<th>Proteinuria</th>
<th>LFT</th>
<th>Max BP</th>
<th>PE Criteria Complete</th>
<th>Rx</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>normal</td>
<td>27</td>
<td>yes</td>
<td>nil</td>
<td>130/100</td>
<td>yes</td>
<td>MgSO4, clexane, aspirin throughout pregnancy</td>
</tr>
<tr>
<td>199</td>
<td>normal</td>
<td>31</td>
<td>yes</td>
<td>nil</td>
<td>raised</td>
<td>yes</td>
<td>labetalol</td>
</tr>
<tr>
<td>423</td>
<td>normal</td>
<td>31</td>
<td>yes</td>
<td>nil</td>
<td>190/115</td>
<td>yes</td>
<td>MgSO4</td>
</tr>
<tr>
<td>424</td>
<td>normal</td>
<td>33</td>
<td>yes</td>
<td>GGTP raised</td>
<td>raised</td>
<td>yes</td>
<td>MgSO4</td>
</tr>
<tr>
<td>364</td>
<td>normal</td>
<td>32</td>
<td>yes</td>
<td>nil</td>
<td>190/115</td>
<td>yes</td>
<td>nifedipine and labetalol</td>
</tr>
<tr>
<td>556</td>
<td>normal</td>
<td>late</td>
<td>yes</td>
<td>raised</td>
<td>yes</td>
<td>labetalol and domperidone</td>
<td>MgSO4 and Labetalol</td>
</tr>
<tr>
<td>526</td>
<td>normal</td>
<td>late</td>
<td>yes</td>
<td>nil</td>
<td>raised</td>
<td>yes</td>
<td>MgSO4 and Labetalol</td>
</tr>
<tr>
<td>234</td>
<td>normal</td>
<td>36</td>
<td>yes</td>
<td>ACR500</td>
<td>160/88</td>
<td>yes</td>
<td>labetalol and MgSO4</td>
</tr>
<tr>
<td>110</td>
<td>normal</td>
<td>late</td>
<td>yes</td>
<td>nil</td>
<td>raised</td>
<td>yes</td>
<td>not known</td>
</tr>
<tr>
<td>105</td>
<td>normal</td>
<td>36</td>
<td>normal</td>
<td>ALT raised, deranged LFT</td>
<td>154/94</td>
<td>yes</td>
<td>labetalol?, AB in labour</td>
</tr>
<tr>
<td>233</td>
<td>normal</td>
<td>39</td>
<td>yes</td>
<td>nil</td>
<td>160/100</td>
<td>yes</td>
<td>labetalol and MgSO4</td>
</tr>
<tr>
<td>70</td>
<td>normal</td>
<td>39</td>
<td>yes</td>
<td>nil</td>
<td>150/110</td>
<td>yes</td>
<td>nil</td>
</tr>
<tr>
<td>170</td>
<td>normal</td>
<td>39</td>
<td>yes</td>
<td>nil</td>
<td>150/90</td>
<td>yes</td>
<td>MgSO4</td>
</tr>
</tbody>
</table>

LFT = Liver Function Tests, BP = Blood Pressure, Rx = Medications
2.6.1 Blood Pressure and Proteinuria

Raised blood pressure and protein in the urine are the two major indicators of pre-eclampsia when observed together. As such, blood pressure readings and dipstick urine tests are performed (usually by a midwife) throughout pregnancy. As pre-eclampsia requires the onset of new hypertension (or a significant worsening of pre-existing hypertension), the monitoring of a woman’s blood pressure early in pregnancy is important to establish a baseline against which to compare later readings. Likewise, pre-existing proteinuria could also complicate diagnosis, but this is much less common, and usually associated with pre-existing conditions.

2.6.2 Histology

Histology was performed by Dr. Noelyn Hung on all placental samples included in this study. She performed the histology blinded to the clinical diagnosis, but not to gestation, as this greatly influences the cellular composition of the placenta. Tenney-Parker changes characteristic of pre-eclampsia were noted in (9/14) of the clinically diagnosed pre-eclamptic samples, and none of these changes were seen in any of the controls (0/14). Tenney-Parker changes are a histological formation that are indicative of placental pathology, and are classically observed on nearly all terminal villi in pre-eclamptic placentas (Tenney & Parker, 1940). In these formations, nuclei in the syncytiotrophoblast are aggregated into structures called syncytial knots. The nuclei within syncytial knots display highly condensed chromatin, and are thought to be a mechanism by which non-functional nuclei can be sequestered (Fogarty et al., 2013).
Table 2.14: Gross and histological pathology of placental samples (Assessment performed by Dr. Noelyn Hung).

<table>
<thead>
<tr>
<th>OPuS Study #</th>
<th>Gestation</th>
<th>Placenta Weight</th>
<th>Breadth</th>
<th>Width</th>
<th>Thickness</th>
<th>Cord Length</th>
<th>Gross Morphology Placental Bed</th>
<th>Microscopic Pathology Summary</th>
<th>Pathology Summary</th>
<th>Baby birth weight (g)</th>
<th>Fetal/Placental Weight Ratio</th>
<th>Placental Function Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>30.3</td>
<td>395</td>
<td>110</td>
<td>100</td>
<td>15</td>
<td>15</td>
<td>Normal</td>
<td>Villi very congested with fetal blood, but otherwise normal for 30 weeks. No inflammation.</td>
<td>Normal</td>
<td>1275</td>
<td>0.31</td>
<td>Increased</td>
</tr>
<tr>
<td>47</td>
<td>30.3</td>
<td>395</td>
<td>140</td>
<td>140</td>
<td>15</td>
<td>42</td>
<td>Marginal UC insertion</td>
<td>Villi very congested with fetal blood, but otherwise normal for 30 weeks. No inflammation.</td>
<td>Marginal Insertion</td>
<td>1475</td>
<td>0.27</td>
<td>increased</td>
</tr>
<tr>
<td>195</td>
<td>31.1</td>
<td>239</td>
<td>110</td>
<td>160</td>
<td>13</td>
<td>16</td>
<td>Marginal UC insertion</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Marginal Insertion</td>
<td>1555</td>
<td>0.15</td>
<td>reduced</td>
</tr>
<tr>
<td>159</td>
<td>33.4</td>
<td>405</td>
<td>160</td>
<td>160</td>
<td>21</td>
<td>31</td>
<td>Normal</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Normal</td>
<td>2000</td>
<td>0.20</td>
<td>Satisfactory</td>
</tr>
<tr>
<td>140</td>
<td>35.2</td>
<td>538</td>
<td>200</td>
<td>210</td>
<td>18</td>
<td>25</td>
<td>Normal</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Normal</td>
<td>3120</td>
<td>0.17</td>
<td>Reduced</td>
</tr>
<tr>
<td>121</td>
<td>36.1</td>
<td>504</td>
<td>180</td>
<td>210</td>
<td>18</td>
<td>60</td>
<td>Normal</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Normal</td>
<td>2880</td>
<td>0.18</td>
<td>Satisfactory</td>
</tr>
<tr>
<td>134</td>
<td>36.2</td>
<td>563</td>
<td>190</td>
<td>220</td>
<td>20</td>
<td>58</td>
<td>Normal</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Satisfactory</td>
<td>2795</td>
<td>0.20</td>
<td>Satisfactory</td>
</tr>
<tr>
<td>122</td>
<td>36.6</td>
<td>526</td>
<td>180</td>
<td>180</td>
<td>18</td>
<td>52</td>
<td>Normal</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Reduced</td>
<td>3490</td>
<td>0.15</td>
<td>Reduced</td>
</tr>
<tr>
<td>520</td>
<td>37.2</td>
<td>460</td>
<td>13</td>
<td>21</td>
<td>22</td>
<td>45</td>
<td>Normal</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Reduced</td>
<td>3400</td>
<td>0.14</td>
<td>Reduced</td>
</tr>
<tr>
<td>478</td>
<td>38.1</td>
<td>464</td>
<td>19.5</td>
<td>23</td>
<td>18</td>
<td>29</td>
<td>Normal</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Reduced</td>
<td>3100</td>
<td>0.15</td>
<td>Reduced</td>
</tr>
<tr>
<td>410</td>
<td>39.1</td>
<td>452</td>
<td>21</td>
<td>17.5</td>
<td>12</td>
<td>46</td>
<td>Normal</td>
<td>Large infarction, good vsm Infarction</td>
<td>Infarction</td>
<td>3965</td>
<td>0.11</td>
<td>Reduced</td>
</tr>
<tr>
<td>32</td>
<td>39.3</td>
<td>370</td>
<td>120</td>
<td>180</td>
<td>15</td>
<td>19</td>
<td>Marginal UC insertion</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Marginal Insertion</td>
<td>2870</td>
<td>0.13</td>
<td>Reduced</td>
</tr>
<tr>
<td>173</td>
<td>39.4</td>
<td>526</td>
<td>190</td>
<td>200</td>
<td>12</td>
<td>47</td>
<td>Normal</td>
<td>Normal: No inflammation, vsm normal. Transmural section.</td>
<td>Reduced</td>
<td>3620</td>
<td>0.15</td>
<td>Reduced</td>
</tr>
</tbody>
</table>

UC = Umbilical cord, vsm = vasculo-syncytial membrane, PE = pre-eclampsia

Continued on next page
**Continued from previous page.**

**Table 4.2:** Gross and histological pathology of placental samples (Assessment performed by Dr. Noelyn Hung).

<table>
<thead>
<tr>
<th>OPUS Study #</th>
<th>Gestation</th>
<th>Placenta Weight</th>
<th>Breadth</th>
<th>Width</th>
<th>Thickness</th>
<th>Cord Length</th>
<th>Gross Morphology</th>
<th>Microscopic</th>
<th>Pathology Summary</th>
<th>Baby birth weight (g)</th>
<th>Fetal/Placental Weight Ratio</th>
<th>Placental Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>416</td>
<td>40.2</td>
<td>590</td>
<td>26</td>
<td>18</td>
<td>16</td>
<td>110</td>
<td>Normal</td>
<td>Normal: No inflammation, vsm normal. Transmural section</td>
<td>Normal</td>
<td>4040</td>
<td>0.15</td>
<td>Reduced</td>
</tr>
<tr>
<td>68</td>
<td>29.6</td>
<td>189</td>
<td>160</td>
<td>130</td>
<td>31</td>
<td>110</td>
<td>Infarction</td>
<td>Tenney-Parker change, maybe some atherosis</td>
<td>PE changes marked</td>
<td>1325</td>
<td>0.14</td>
<td>Reduced</td>
</tr>
<tr>
<td>199</td>
<td>31.1</td>
<td>218</td>
<td>115</td>
<td>170</td>
<td>15</td>
<td>22</td>
<td>Infarction</td>
<td>Hypertension, inflammation, no inflammation</td>
<td>PE changes</td>
<td>1270</td>
<td>0.17</td>
<td>Reduced</td>
</tr>
<tr>
<td>423</td>
<td>31.3</td>
<td>336</td>
<td>18</td>
<td>19</td>
<td>15</td>
<td>49</td>
<td>Infarction</td>
<td>Major Tenney-Parker change, very poor vsm, minor acute inflammation in decidua</td>
<td>PE changes</td>
<td>1985</td>
<td>0.17</td>
<td>Reduced</td>
</tr>
<tr>
<td>424</td>
<td>33.1</td>
<td>464</td>
<td>14</td>
<td>18</td>
<td>25</td>
<td>36</td>
<td>Normal</td>
<td>Patchy Tenney-Parker change, patchy good vsm, overall hypermature for 33 weeks</td>
<td>PE changes</td>
<td>1880</td>
<td>0.25</td>
<td>Increased</td>
</tr>
<tr>
<td>364</td>
<td>34.4</td>
<td>538</td>
<td>19.5</td>
<td>18</td>
<td>24</td>
<td>32</td>
<td>Normal</td>
<td>Mild Tenney-Parker change if any, no inflammation, no atherosis</td>
<td>PE changes</td>
<td>2795</td>
<td>0.19</td>
<td>Satisfactory</td>
</tr>
<tr>
<td>526</td>
<td>36.2</td>
<td>530</td>
<td>12.5</td>
<td>20</td>
<td>20</td>
<td>34</td>
<td>Marginal UC insertion</td>
<td>Minor chronic sclerosing villitis</td>
<td>Chronic Villitis (VUE), Marginal Insertion</td>
<td>2400</td>
<td>0.22</td>
<td>Satisfactory</td>
</tr>
<tr>
<td>234</td>
<td>36.5</td>
<td>367</td>
<td>16</td>
<td>19</td>
<td>19</td>
<td>30</td>
<td>Fibrinous deposition</td>
<td>Perivillus fibrin deposition on slide (focal but large area subchorionic), focal chorangiosis, Tenney-Parker change in places, no inflammation</td>
<td>PE changes</td>
<td>2770</td>
<td>0.13</td>
<td>Reduced</td>
</tr>
<tr>
<td>250</td>
<td>37</td>
<td>590</td>
<td>20</td>
<td>19</td>
<td>19</td>
<td>33</td>
<td>Marginal UC insertion</td>
<td>Tenney-Parker change with patchy villus membrane formation (poor for 40 weeks)</td>
<td>PE changes, large placenta for PE, Marginal Insertion</td>
<td>3900</td>
<td>0.15</td>
<td>Reduced</td>
</tr>
<tr>
<td>110</td>
<td>37.2</td>
<td>402</td>
<td>180</td>
<td>190</td>
<td>22</td>
<td>30</td>
<td>Marginal UC insertion</td>
<td>Some Tenney-Parker change, no inflammation, chunks of calcification</td>
<td>PE changes</td>
<td>2690</td>
<td>0.15</td>
<td>Reduced</td>
</tr>
</tbody>
</table>

UC = Umbilical cord, vsm = vasculo-syncytial membrane, PE = pre-eclampsia

Continued on next page
### Table 4.2: Gross and histological pathology of placental samples (Assessment performed by Dr. Noelyn Hung).

<table>
<thead>
<tr>
<th>OPuS Study #</th>
<th>Gestation</th>
<th>Placenta Weight</th>
<th>Breadth</th>
<th>Width</th>
<th>Thickness</th>
<th>Cord Length</th>
<th>Gross Morphology</th>
<th>Microscopic</th>
<th>Pathology Summary</th>
<th>Baby birth weight (g)</th>
<th>Fetal Placental Weight Ratio</th>
<th>Placental Function Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>105</td>
<td>38.2</td>
<td>618</td>
<td>180</td>
<td>170</td>
<td>na</td>
<td>21</td>
<td>Normal</td>
<td>Some Tenney-Parker change, no inflammation, chunks of calcification</td>
<td>PE changes</td>
<td>3530</td>
<td>0.18</td>
<td>satisfactory</td>
</tr>
<tr>
<td>233</td>
<td>39</td>
<td>368</td>
<td>15</td>
<td>NA</td>
<td>15</td>
<td>31</td>
<td>Infarction</td>
<td>Intervillous thrombus, infarction, poor villus formation, no inflammation</td>
<td>PE changes</td>
<td>2760</td>
<td>0.13</td>
<td>Reduced</td>
</tr>
<tr>
<td>70</td>
<td>39.3</td>
<td>462</td>
<td>180</td>
<td>150</td>
<td>25</td>
<td>32</td>
<td>Normal</td>
<td>Normal Transmural section, no inflammation, vsm normal</td>
<td>Normal</td>
<td>2585</td>
<td>0.18</td>
<td>satisfactory</td>
</tr>
<tr>
<td>170</td>
<td>39.4</td>
<td>590</td>
<td>190</td>
<td>230</td>
<td>14</td>
<td>54</td>
<td>Normal</td>
<td>Mild TP change if any, no inflammation, no atherosis</td>
<td>PE changes</td>
<td>3585</td>
<td>0.16</td>
<td>Reduced</td>
</tr>
<tr>
<td>556</td>
<td>35</td>
<td>371</td>
<td>20</td>
<td>16</td>
<td>20</td>
<td>17</td>
<td>Normal</td>
<td>Normal Transmural section, no inflammation, vsm normal</td>
<td>-</td>
<td>2260</td>
<td>0.16</td>
<td>Reduced</td>
</tr>
</tbody>
</table>

UC = Umbilical cord, vsm = vasculo-syncytial membrane, PE = pre-eclampsia, VUE = Villitis of unknown etiology.
2.7 Matching of Controls

For the purpose of epigenetic analysis, it was important to perform case-control matching based on factors that are able to influence the epigenetic profile of cells, particularly factors that play a key role in pregnancy outcomes. It was determined that, while it would be preferable to match even more thoroughly, controls would be matched as closely as possible based on four main factors. In order of importance for matching, these factors were: Pregnancy Gestation, Fetal Sex, Maternal Smoking, and Maternal BMI. These controls also came from pregnancies that displayed neither hypertension nor proteinuria.

2.7.1 Gestation

It is well known that methylation in the placenta changes over gestation (Novakovic et al., 2011). Because pre-eclampsia is a disease of early pregnancy that manifests in late pregnancy, the decision was made to use a range of gestations, to ensure that changes we saw between cases and controls were not due to the gestation of the placentas studied. In this way, we aimed to detect changes in the diseased placentas that were as a result of the disease status - and thus may give an indication as to possible causation. Alternatively, a change that is non-causative, yet associates strongly with disease state may act as a good biomarker of later PET risk for women early in their pregnancy, prior to the gestation at which current detection systems are effective. No differences in gestation were observed between cases and controls in this study (Figure 2.1B).

2.7.2 Fetal Sex

Large differences exist between female and male biology, and while many of these differences are known, and can be controlled for, the differences that exist in methylation profiles are still being elucidated. Therefore samples were matched for fetal sex.
Table 2.15: Matching conditions for samples used in this study. Weighting for matching was performed in the order of Gestation > Fetal sex > Smoking status > Maternal BMI.

<table>
<thead>
<tr>
<th>Status</th>
<th>OPU Study #</th>
<th>Gestation</th>
<th>Sex</th>
<th>Smoker</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE</td>
<td>68</td>
<td>29.6</td>
<td>M</td>
<td>No</td>
<td>28.8</td>
</tr>
<tr>
<td>normal</td>
<td>47</td>
<td>30.3</td>
<td>M</td>
<td>No</td>
<td>23</td>
</tr>
<tr>
<td>PE</td>
<td>199</td>
<td>31.1</td>
<td>M</td>
<td>No</td>
<td>25</td>
</tr>
<tr>
<td>normal</td>
<td>195</td>
<td>31.1</td>
<td>M</td>
<td>10+/day for 8 yrs</td>
<td>30.8</td>
</tr>
<tr>
<td>PE</td>
<td>423</td>
<td>31.3</td>
<td>F</td>
<td>No</td>
<td>23.4</td>
</tr>
<tr>
<td>normal</td>
<td>46</td>
<td>30.3</td>
<td>F</td>
<td>No</td>
<td>23</td>
</tr>
<tr>
<td>PE</td>
<td>424</td>
<td>33.1</td>
<td>F</td>
<td>No</td>
<td>24.6</td>
</tr>
<tr>
<td>normal</td>
<td>159</td>
<td>33.4</td>
<td>F</td>
<td>No</td>
<td>25.9</td>
</tr>
<tr>
<td>PE</td>
<td>364</td>
<td>34.4</td>
<td>F</td>
<td>Yes</td>
<td>19.5</td>
</tr>
<tr>
<td>normal</td>
<td>140</td>
<td>35.2</td>
<td>F</td>
<td>Yes</td>
<td>27.3</td>
</tr>
<tr>
<td>PE</td>
<td>556</td>
<td>35</td>
<td>F</td>
<td>Yes</td>
<td>45</td>
</tr>
<tr>
<td>normal</td>
<td>121</td>
<td>36.1</td>
<td>F</td>
<td>No</td>
<td>30.1</td>
</tr>
<tr>
<td>PE</td>
<td>526</td>
<td>36.2</td>
<td>M</td>
<td>Yes</td>
<td>20</td>
</tr>
<tr>
<td>normal</td>
<td>134</td>
<td>36.2</td>
<td>M</td>
<td>No</td>
<td>24.9</td>
</tr>
<tr>
<td>PE</td>
<td>234</td>
<td>36.5</td>
<td>M</td>
<td>Yes but quit 14 days prior</td>
<td>28.2</td>
</tr>
<tr>
<td>normal</td>
<td>122</td>
<td>36.6</td>
<td>M</td>
<td>No</td>
<td>27.7</td>
</tr>
<tr>
<td>PE</td>
<td>110</td>
<td>37.2</td>
<td>F</td>
<td>No</td>
<td>20.6</td>
</tr>
<tr>
<td>normal</td>
<td>520</td>
<td>37.2</td>
<td>F</td>
<td>No</td>
<td>21</td>
</tr>
<tr>
<td>PE</td>
<td>105</td>
<td>38.2</td>
<td>M</td>
<td>No</td>
<td>26</td>
</tr>
<tr>
<td>normal</td>
<td>478</td>
<td>38.1</td>
<td>M</td>
<td>No</td>
<td>23</td>
</tr>
<tr>
<td>PE</td>
<td>233</td>
<td>39</td>
<td>F</td>
<td>No</td>
<td>35.2</td>
</tr>
<tr>
<td>normal</td>
<td>410</td>
<td>39.1</td>
<td>F</td>
<td>No</td>
<td>33.5</td>
</tr>
<tr>
<td>PE</td>
<td>70</td>
<td>39.3</td>
<td>F</td>
<td>No</td>
<td>27.2</td>
</tr>
<tr>
<td>normal</td>
<td>32</td>
<td>39.3</td>
<td>F</td>
<td>No</td>
<td>25.4</td>
</tr>
<tr>
<td>PE</td>
<td>170</td>
<td>39.4</td>
<td>F</td>
<td>Yes</td>
<td>33.1</td>
</tr>
<tr>
<td>normal</td>
<td>173</td>
<td>39.4</td>
<td>F</td>
<td>Yes</td>
<td>26.4</td>
</tr>
<tr>
<td>PE</td>
<td>250</td>
<td>40.2</td>
<td>M</td>
<td>No</td>
<td>24</td>
</tr>
<tr>
<td>normal</td>
<td>416</td>
<td>40.2</td>
<td>M</td>
<td>No</td>
<td>19.7</td>
</tr>
</tbody>
</table>
2.7.3 Maternal Smoking

A large volume of information exists relating to the effects of maternal smoking upon fetal development. Indeed, recent work done by Noelyn Hung, Tania Slatter, Celia Devenish and collaborators (University of Otago) has shown the effect that smoking during pregnancy has on DNA damage in the placenta. In part due to the social stigma attached to smoking, and in particular to smoking during pregnancy, it can be difficult to obtain accurate information with regards to whether a woman has smoked during her pregnancy. For this tissue collection there were excellent data on smoking during pregnancy, including information on smokers who ceased smoking during their pregnancy (Slatter et al., 2014). Smoking during pregnancy has been shown to cause changes in both global and specific DNA methylation (Suter et al., 2011; Wilhelm-Benartzi et al., 2012), so it was important to match on this factor to isolate pre-eclampsia related factors from smoking related factors.

2.7.4 Maternal BMI

Maternal BMI measures were based upon a woman’s pre- or peri-conception weight, rather than including any weight changes that occurred during pregnancy. It is known that abnormal weight gain during pregnancy can greatly increase the risk of pregnancy complications. However, if, as trophoblast invasion evidence suggests, this disease is establishing in early pregnancy rather than late pregnancy, a woman’s weight around the time of conception and implantation is more likely to have a bearing on fetal development. Maternal over-nutrition and under-nutrition have both been linked to changes in fetal birth weight, increases in childhood obesity, risk for cardiovascular disease, diabetes, and other non-communicable diseases (NCDS). A wider range of BMIs were observed in women with Pre-eclamptic pregnancies, but the difference is not significant, and largely due to a single individual with a BMI above 40 (Figure 2.1C).
2.7.5 Additional Cohort Characterisation

In addition to the factors accounted for in the matching of the placental samples, a number of other clinical factors were measured to allow for future stratification, and post-hoc analyses. These factors may contribute to clinical variation in the disease, and may alter the risk of disease in pregnancy. Maternal age (Figure 2.1A), parity, and pre-existing conditions in the mother may have a significant impact upon the risk of developing pre-eclampsia, especially if the mother has had previous pre-eclamptic pregnancies. Medications the mother may have been taking during her pregnancy, or medications prescribed to treat pre-eclampsia - e.g. Magnesium sulphate or Labetalol may have an impact upon the clinical presentation or epigenetic state of the placenta. Likewise, the method of delivery may have an impact upon the epigenetics of the placenta, as can whether the pregnancy is a singleton, or multiple. As expected, comparisons between the cases and controls on matching factors did not find any significant differences in the factors (Figure 2.1). Maternal age and personalised growth centiles were also compared. Maternal age has been associated with an increased risk of pre-eclampsia (Lamminpää et al., 2012), but no differences in maternal age were observed between cases and controls in this cohort.
2.7.5.1 Method of Delivery

One of the possible confounding factors in a study of this nature is method of delivery. As expected, we saw a much higher proportion of interventions during birth in the
PET cohort than in the control cohort. The pre-term controls were largely a result of spontaneous pre-term labour, with the exception of the twin pregnancies, which are expected to deliver earlier in gestation than a singleton pregnancy. The majority of PET cases were delivered by caesarean section, emergency or elective. While the method of delivery was not being controlled for, it may become relevant for future studies. Because PET cases are more likely to require intervention, the differences in methylation may be reflecting the need for intervention during birth rather than the pre-eclampsia itself. Therefore, it should be acknowledged as a potential confounder in studies of this sort.

**Figure 2.2:** Modes of delivery in placenta cohort. All deliveries that did not require intervention were observed in control samples, while Caesarian Sections were most common in pre-eclamptic patients. NVD/SVD = Normal/Spontaneous Vaginal Delivery, C/S = Caesarian Section, EM C/S = Emergency Caesarian Section, IOL = Induction of Labour
2.8  TruSeq Reduced Representation Bisulphite Sequencing Library Preparation

The TruSeq DNA sample preparation kit (Illumina) is designed for use with intact DNA, so it was necessary to make several alterations to the recommended protocol in order to adapt it for use with bisulphite converted DNA. The following section will be an exemplum of the protocol as used for each placental DNA sample.

2.8.1  MspI Digestion

<table>
<thead>
<tr>
<th>OPuS Study #</th>
<th>Concentration (ng)</th>
<th>µL to make 2.5 µg</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>146.98</td>
<td>17.69</td>
</tr>
</tbody>
</table>

Table 2.17: MspI restriction endonuclease digest

<table>
<thead>
<tr>
<th>Sample 68</th>
<th>µL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.69</td>
<td>DNA</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>MspI</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Buffer 4</td>
<td></td>
</tr>
<tr>
<td>14.31</td>
<td>MQH₂O</td>
<td></td>
</tr>
<tr>
<td>40.00</td>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

The TruSeq protocol recommends an input of 1 µg DNA, but it was found by Chatterjee et al. (2012a) that this input was too low to allow adequate DNA recovery at the end of the library preparation process, perhaps due to the bisulphite-converted nature of the DNA. Instead, 2.5 µg was used as an initial input. MspI cuts at a CCGG site, regardless of methylation at that site.
2.8.1.1 QiaQuick Purification

Following enzymatic digest, the Qiagen QiaQuick PCR Purification kit was used to purify DNA, eluting in 60 µL of elution buffer. Five volumes of Buffer PB were added to one volume of DNA digest, and the resulting mixture applied to a QiaQuick spin column. Columns were centrifuged in a tabletop microcentrifuge for 1 min at 10,000 x g, and the flow-through discarded. The filter was washed with 750 µL of Buffer PE, and centrifuged. The flow-through was again discarded, and the column spun for an additional minute to remove traces of ethanol. The column was then placed into a clean 1.5 mL microcentrifuge tube, and 60 L of Buffer EB applied to the filter, and eluted via centrifugation after a 10-minute incubation to increase the yield.

2.8.2 End Repair

The restriction digest creates sticky ends that must be changed to blunt ends for A-tailing and adaptor ligation. The End Repair mix has 3’ to 5’ exonuclease activity to remove the 3’ overhangs and polymerase activity to fill in the 5’ overhangs. To each 60 µL elution from the previous step 40 µL of End Repair Mix was added. The resulting solution was transferred to a 0.2 mL PCR tube and incubated in a BioRad DNA Engine® Thermal Cycler at 30°C for 30 min. Following this incubation, the reaction was stopped and the samples purified using the Qiagen MinElute PCR purification kit. The protocol for this kit is the same as for the QiaQuick kit, but uses a different filter column, and allows elution in a smaller volume. Samples were eluted in 18 µL of Buffer EB.

2.8.3 Adenylation of 3’ Ends

To the 3’ end of each blunt ended fragment, an adenine residue is attached to prevent the fragments from sticking to each other. This correspond with an overhanging thymine residue on the 3’ end of the adaptors, thus ensuring fragments will bind only
to adaptors and not each other. To this end, 12.5 µL of A-tailing mix was added to 
each sample, and the solution was transferred to a 0.2 mL PCR tube and incubated in 
a Thermal Cycler at 37°C for 30 min.

2.8.4 Adaptor Ligation

Individual adaptors for each sample were carefully selected for optimal performance 
in a multiplexed sequencing run. Certain adaptors do not match well when performing 
multiplexed sequencing reaction on the Illumina HiSeq platform (4 samples/lane). At-
ttempts were made when selecting adaptors to ensure matched pairs could be placed in 
the same lane to minimise differences due to their position on the Illumina run. How-
ever, this was not always possible. To the 30 µL reaction from above was added 2.5 µL 
of Resuspension Buffer (RSB), 2.5 µL of DNA Ligase mix, and 2.5 µL of the appro-
priate Adaptor Index. The solution was then incubated at 30°C for 10 min on a thermal 
cycler. Following this incubation, 5 µL of Stop Ligation (STL) mix was added to stop 
the reaction.

2.8.5 AMPure Bead Purification

This clean-up step makes use of AMPure XP beads (Beckman Coulter), magnetic 
beads which bind DNA longer than 100 bp. This enables removal of unbound adap-
tors, and shorter templates that have not bound adaptors. AMPure beads were removed 
from storage at 4°C approximately 30 min before use to enable them to equilibrate to 
room temperature. The AMPure XP Beads were vortexed until well dispersed, then 
42.5 µL of the beads was added to each sample tube, and the entire volume gently 
pipetted up and down 10 times to thoroughly mix them. The tubes were then incubated 
at room temperature for 15 min. The tubes were then placed upon a Magnetic Stand-
96 (Life Technologies) for around 5 min, or until all the beads had aggregated at one 
point. Next, 80 µL of the supernatant was then carefully removed without disturbing 
the beads, and then discarded. Two washes, each of 200 µL of freshly prepared 80%
ethanol were performed to remove anything that is not bound to the beads. After the wash steps, as much of the ethanol as possible was removed from the tubes, and the samples were allowed to air dry at room temperature for approximately 15 min, or until the ethanol had completely evaporated. The dried pellets were then removed from the magnetic plate, resuspended in 20 µL of RS Buffer, and gently pipetted up and down to mix before incubating at room temperature for 5-10 min. After the incubation, the tubes were placed back onto the magnetic stand for about 2 min, or until the beads had magnetised to the side of the tube. Finally, 19 µL of the supernatant was transferred to a new 0.2 mL tube.

2.8.6 First Round of Size Selection

DNA fragments ranging from 40-220 bp in size were selectively retained for library preparation using a simple gel-based approach. A 3% (w/v) NuSieve™ GTG™ Agarose gel (Lonza) was prepared according to the manufacturers instructions, and allowed to set for about 45 min. The supernatant from the AMPure bead purification was mixed with 2.5 µL of xylene cyanol loading dye, and the entire 21.5 µL was loaded onto the gel, with 6.5 µL of 25 bp DNA Ladder (Invitrogen). The gel was run on a Takara system at 50 V for 90 min. A UV tabletop transilluminator was used to visualise the ladder and razor blades used to remove chunks of gel between 150-330 bp in size (40-220 bp DNA fragments + 2 x 50 bp adaptors) for each sample. Blades were changed between samples to reduce possibility of cross-contamination. Gel fragments were placed into 2 mL microcentrifuge tubes and weighed to determine gel weight.

2.8.6.1 Gel Extraction

Two different gel extraction kits were used during the preparation of libraries. Libraries that were sequenced on run 464 were extracted using the Qiaex II Gel extraction kit, and those that were sequenced on run 556 used the QIAquick Gel extraction kit (Both from Qiagen. Both protocols will be summarised here.
Qiaex II
The Qiaex II kit requires the addition of 3 volumes of Buffer QX1 to 1 volume of gel. Qiaex II beads were resuspended by vortexing, and 10 µL of beads added to each sample. Samples were incubated at 50°C on a thermomixer for 10 min to solubilise the gel and bind DNA to the beads. Tubes were then spun in a benchtop microcentrifuge, and the supernatant removed. The pellet was the washed with 500 µL of Buffer QX1, spun and the supernatant removed. Two washes each of 500 µL were then performed in the same manner. After completion of the wash steps, the pellet was allowed to air dry for 20 min, or until the pellet became a pure white. To elute DNA, 20 µL of MilliQ water was added to the pellet and thoroughly mixed. After a 10-minute incubation, samples were centrifuged for 1 minute, and the supernatant transferred to a clean tube. The elution step was then repeated, and the eluates combined.

QIAquick
The QIAquick gel extraction kit is a column-based extraction tool. Similar to the Qiaex II, 3 volumes of Buffer QG were added to 1 volume of gel. The samples were incubated for 10 min at 50°C on a thermomixer to solubilise the gel. Once the gel was completely dissolved, 1 gel volume of isopropanol was added to the sample. After mixing, the entire solution was transferred to a QIAquick spin column and collection tube. The column was then spun for 30 sec at 10,000 x g and the flow-through discarded. The filter was washed with an additional 500 µL of Buffer QG to remove traces of agarose, spun, and the flow-through discarded. To wash, 750 µL of Buffer PE was applied to the column, and allowed to sit at room temperature for 2 min before this too was spun, and the flow-through discarded. An additional spin of 1 minute served to remove traces of ethanol from the filter. The column was then transferred to a clean 1.5 mL microcentrifuge tube, and 40 µL of Buffer EB added to the column. After incubating for 10 min, the tubes were spun, and the eluate re-applied to the column, allowed to sit for 2 min, and re-spun. The final eluate was then transferred to a 0.2 mL PCR tube, ready for bisulphite conversion.
2.8.7 Bisulphite Conversion of RRBS Libraries

The Zymo EZ DNA Methylation kit was used for bisulphite conversion of DNA libraries according to the manufacturers protocol, with some slight changes. Prior to its use, the CT conversion reagent was prepared by adding 210 µL of M-Dilution Buffer and 750 µL of MilliQ water, and mixing it thoroughly for 20 min, or until the crystals had dissolved. To the 40 µL elution from the previous clean-up step was added 10 µL of M-Dilution Buffer. This DNA was then incubated at 37°C for 15 min. To the solution was added 100 µL of the pre-prepared CT Conversion Reagent, and the resultant mix is incubated on a thermal cycler overnight (16 hrs) at 50°C. After incubation, the sample was loaded, along with 400 µL of M-Binding Buffer, onto a Zymo-Spin™ IC Column (and 2 mL Collection Tube). The sample was centrifuged for 30 s at 12000 rpm and the flow-through discarded. Then 100 µL of M-Wash Buffer was added to the column and centrifuged. The flow-through was again discarded. 200 µL of M-Desulphonation Buffer was added to the column and incubated at room temperature for 15-20 min. After incubation, the column was centrifuged and the flow-through discarded. Another two washes of 200 µL of M-Wash Buffer each time were performed. After the second centrifugation, the column was placed into a clean 1.5 mL microcentrifuge tube and 20 µL of M-Elution Buffer applied to the matrix. The column was incubated for 10 min at room temperature and then centrifuged for 30 s at full speed. The eluate was reapplied to the matrix and centrifuged again to optimise the yield. Bisulphite-converted DNA was stored at -30°C.

2.8.8 Semi-Quantitative PCR

To determine the optimal cycle number for library amplification, a small test PCR was run at two different cycle numbers (Table 2.19). Primers from the TruSeq DNA kit are designed to bind to the adaptors that have been annealed to ones DNA fragments. These primers are methylated so as to be unaffected by bisulphite conversion. PCR master-mixes was set up as shown in Table 2.18:
Table 2.18: Semi-quantitative PCR reaction (2X)

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x PCR Buffer</td>
<td>2.5 µL</td>
</tr>
<tr>
<td>dNTPs (25 nM)</td>
<td>3 µL</td>
</tr>
<tr>
<td>TruSeq PCR Primer Cocktail</td>
<td>2 µL</td>
</tr>
<tr>
<td>DNA Library</td>
<td>3 µL</td>
</tr>
<tr>
<td>Pfu Turbo Taq</td>
<td>1.2 µL</td>
</tr>
<tr>
<td>MilliQ H2O</td>
<td>13.3 µL</td>
</tr>
<tr>
<td>Total</td>
<td>25 µL</td>
</tr>
</tbody>
</table>

The PCR reactions were prepared and then split, with 12.5 µL of PCR master mix aliquoted into each of 2 PCR tubes. The PCRs were then run on two independent blocks on the same BioRad thermocycler.

Table 2.19: Cycling condition for Semi-quantitative PCRs

<table>
<thead>
<tr>
<th>Time</th>
<th>Temperature</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denaturation/Activation</td>
<td>2 min</td>
<td>95°C</td>
</tr>
<tr>
<td>Denaturation</td>
<td>30 s</td>
<td>95°C</td>
</tr>
<tr>
<td>Annealing</td>
<td>30 s</td>
<td>65°C</td>
</tr>
<tr>
<td>Elongation</td>
<td>45 s</td>
<td>72°C</td>
</tr>
<tr>
<td>Extension</td>
<td>7 min</td>
<td>72°C</td>
</tr>
<tr>
<td>Cooling</td>
<td>∞</td>
<td>4°C</td>
</tr>
</tbody>
</table>

PCR products (12.5 µL + 2 µL loading dye) were loaded alongside a 25 bp ladder (6 µL) and run on a 4-20% Criterion precast TBE gradient gel (Bio-Rad) at 100 V for 100 min (1x TBE buffer). The gel was stained with SYBR green at a concentration of 1:10,000 for 30 min. Gels were then imaged, and the optimal cycle number for large-scale amplification determined. This was defined as the lowest number of cycles for which a satisfactory yield could be attained.
2.8.9 Large-Scale PCR Amplification of Libraries

Table 2.20: Large-Scale PCR Reaction

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x PCR Buffer</td>
<td>10 µL</td>
</tr>
<tr>
<td>dNTPs (25 nM)</td>
<td>12 µL</td>
</tr>
<tr>
<td>TruSeq PCR Primer Cocktail</td>
<td>8 µL</td>
</tr>
<tr>
<td>DNA Library</td>
<td>12 µL</td>
</tr>
<tr>
<td>Pfu Turbo Taq</td>
<td>4.8 µL</td>
</tr>
<tr>
<td>MilliQ H20</td>
<td>2.84 µL</td>
</tr>
<tr>
<td>Total</td>
<td>100 µL</td>
</tr>
</tbody>
</table>

12.5 µL of PCR master mix was aliquoted into PCR strip tubes (8 aliquots) - this was to keep conditions as close to that of the semi-quantitative PCR as possible. The PCR was run on the BioRad thermocycler at the pre-determined number of cycles. Post-PCR, the aliquots of PCR product were combined, and the Qiagen MinElute kit was used (as described previously) to purify DNA, eluting in 18 µL of Buffer EB.

2.8.10 Round Two of Size Selection

A 3% NuSieve gel was prepared as described for the first round of size selection. Then 2 µL of DNA loading buffer was added to the 18 µL DNA library, and it was loaded into a single well. Molecular weight marker (6 µL of 25 bp DNA ladder) was added to each of the left-most and right-most wells. The gel was run at 50 V for 80 min. Libraries of size 150-330 bp were excised as described in the first round of size-selection, and the gel fragments weighed. Libraries were extracted from the gel using either the Qiaex II, or QIAquick kits, as described previously, eluting in 20 µL.
2.8.11 Preparation for Multiplexing and Sequencing

As previously determined by Chatterjee et. al. (2012), four RRBS libraries can be multiplexed into a single lane without adversely affecting the output for each sample. However, the ability to multiplex libraries is affected by the adaptor index attached to each library. Certain adaptors work better together than others due to the unique six-nucleotide sequence of the index. Illumina uses a green laser to sequence G/T and a red laser to sequence A/C. At each cycle of the six calibration rounds, there needs to be a nucleotide in each of the green channel (G/T) and the red channel (A/C) to ensure the image is read properly. The Illumina Experiment Manager Software was used when planning the multiplexing to ensure that the bases of the index cycles were balanced.

2.8.11.1 Qubit Quantification

The concentration of each library was determined via Qubit. Briefly, a working solution was first made up consisting of 199 µL x n (where n is the number of libraries that are being quantified plus two Qubit standards) of Qubit Buffer dsDNA High Sensitivity (HS), and 1 µL x n Qubit HS reagent. To an Invitrogen clear 0.5 mL tube was then added 198 µL of working solution, and 2 µL of the library of interest. In addition to the libraries, two Qubit standards were made up, each consisting of 190 µL of working solution and 10 µL of either standard #1 or standard #2. All tubes were briefly vortexed, and then incubated at room temperature for 2 min. Following the instructions on the display screen of the Qubit, the two standards were used to calibrate the machine before each run, and the concentration of each library determined. Qubit quantifications for Illumina HiSeq run 464 were performed by NZGL.

2.8.11.2 Bioanalyser Qualification

An Agilent Bioanalyser assay was used to determine the quality of each library, and the size distribution of the DNA fragments within each library. Reagents were removed from the fridge 30 min before use. Following the instructions for the Agilent High-
Sensitivity (HS) DNA kit, chips were prepared as follows. HS DNA dye concentrate was vortexed for 10 sec and spun down. 15 µL of HS dye was then transferred into a vial of HS DNA gel matrix, and vortexed to mix thoroughly. The gel-dye mix was then transferred to a spin filter and centrifuged for 10 min at 6000 rpm. The filter was discarded, and the eluted gel-dye mix retained. The Gel-dye mix can be stored at 4°C for up to 6 weeks. If the gel-mix had been prepared previously, it was equilibrated to room temperature before proceeding. The chip was placed upon the chip priming station, with the base plate set to position C, and the adjustable clip set to the lowest position. 9 µL of gel-dye mix was then pipetted into the well marked with a G, and the chip priming station closed by pushing the plunger of the syringe down smoothly and firmly until held by the clip. After exactly 60s, the clip was released, after an additional 5 s, the plunger was pulled back to the 1 mL mark. The chip was removed from the priming station, and 9.0 µL of gel-dye pipetted into each of the remaining wells marked G. 5 µL of the green-capped HS DNA marker was then loaded into the ladder well and each of the 11 sample wells. 1 µL of the yellow-capped HS DNA ladder was loaded into the ladder well, and 1 µL of sample or marker (unused wells) into each of the sample wells. The chip was placed on the IKA vortex mixer for 60 s at 2400 rpm, and inserted into the Bioanalyzer within 2 min.

The HS-DNA assay was selected on the instrument panel of the 2100 Expert software, and the appropriate sample information entered before starting the run. After completion of the run, the Region Table utility was used to perform a ‘gating’ of the fragments between 150 and 330 bp. This provides one with the average fragment length within that gate, and is used for calculating molarity for the Illumina HiSeq.

### 2.8.11.3 Calculation of Molarity

For run 556, we provided NZGL with 10 µL of each library at a concentration of 10 nM. These libraries were then diluted to 2 nM by NZGL (10 nM allows for NZGL to perform re-runs if necessary) and then pooled. The starting concentration of each
library was calculated using the concentration determined by Qubit assay, and the average read length from Bioanalyser according to the following equation:

\[ nM = \frac{\text{concentration (ng/µL)}}{(\text{average fragment length(bp)} \times 0.00065)} \]

2.8.11.4 Dilution of Libraries

Libraries were then diluted to 10 nM using the equation:

\[ \frac{(\text{concentration in nM})}{100} = \text{volume of library needed} \]

Volume was then made up to 10 µL with MQH₂O, and the libraries stored at -80°C until sent to the sequencing facility.

2.9 Sequenom MassArray®EpiTYPER® Methylation Validation

Sequenom MassARRAY is a useful tool for targeted and highly-quantitative methylation analysis. It measures methylation of PCR products based on the molecular weight of DNA fragments using MALDI-TOF (Matrix-Assisted Laser Desorption/Ionization-Time Of Flight) mass spectrometry. Unlike sequencing based techniques, it is not liable to allele or sequencing bias, so makes an excellent validation tool for confirming results from a method like RRBS, where a genome-wide technique is used to identify a large number of candidates for further investigation. For reasons of efficiency, it is not ideal to perform targeted bisulphite re-sequencing for more than a few genes when a large number of samples need to be analysed, as when diagnostic tests are performed. Sequenom allows us to analyse a greater number of samples at little extra effort.
2.9.1 Selection of Candidate Genes from RRBS for Sequenom

The analysis pipeline underwent various permutations, and Sequenom experiments were designed based upon an earlier permutation than that used to generate the results shown in Chapter 5. Genes were selected for further validation from the RRBS analysis via a number of criteria. The first criterion was that they show a distinct difference (>15%) in methylation for at least one CpG site between pre-eclamptic and normal placentas. For a CpG site to be analysed to this end, it had to be covered by at least 8 sequence reads in at least 10 of the 15 samples for each group (cases and controls). For more details on how this analysis was performed, see the sections of Chapter 4 which pertain to the MethylKit analysis tool. After the CpGs meeting these criteria were identified, a number of other criteria were applied to select gene regions for validation by Sequenom. Gene regions were selected according to the following criteria: largest differences, multiple CpGs, multiple fragments, location in or near a promoter, location in or near a splice site, and imprinted genes. In all, 44 genes were selected for primer design.

2.9.2 Candidate Gene Primer Design

Primers for each of the validation genes were designed using the Sequenom Epidesigner tool (available at http://www.epidesigner.com/). Sequences of 1000 bp (candidate CpG site +/- 500 bp) were obtained from the UCSC genome browser, and pasted into the Epidesigner to refine primers. Amplicons had to contain, and be informative for, the CpG(s) that were deemed of interest from the RRBS analysis. Sequenom uses an enzyme digest to fragment DNA prior to analysis and the number of CpGs in the amplicon for which data can be observed depends upon the locations of cut site. The Epidesigner software often offered several possibilities of amplicons to choose from, so where possible, the amplicon with the greatest coverage past the original RRBS fragment and which included the most CpG sites was selected. For Sequenom analysis, a T7-promoter tag must be added to the reverse primer, and a 10 bp tag to the
forward primer, as shown below:

5’- AGGAAGAGAG - Forward Primer Sequence - 3’
5’- CAGTAATACGACTCACTATAGGGAGAAGGCT - Reverse Primer Sequence - 3’

2.9.3 Bisulphite Conversion of Placental DNA

As in previous sections (2.4.1 and 2.5.7), the EZ DNA MethylationTM Kit from Zymo was used for bisulphite conversion, with a few alterations. Conversion was performed following the manufacturer’s protocol, with 1 µg of DNA made up to 45 µL with MQ H2O. The DNA was then made up to 50 µL with 5 µL of M-Dilution Buffer. This DNA solution was then incubated at 37°C for 15 min. To the above solution was added 100 µL of CT Conversion Reagent, and the resultant mix was placed in a thermal cycler for a cycling incubation using Zymo’s “Alternative Cycling Protocol” of 20 cycles of 95°C for 30 seconds, then 50 °C for 15 min.

After incubation, the sample was loaded, along with 400 µL of M-Binding Buffer, onto a Zymo-SpinTM IC Column (and 2 mL collection tube). The sample was centrifuged for 30 s at 12000 rpm and the flow-through discarded. Then 100 µL of M-Wash Buffer was added to the column and centrifuged. The flow-through was again discarded. 200 µL of M-Desulphonation Buffer was added to the column and incubated at room temperature for 15-20 min. After incubation, the tube was centrifuged and the flow-through discarded. The filter was washed twice, with 200 µL of M-Wash Buffer each time. After the second centrifugation, the column was placed into a clean 1.5 mL microcentrifuge tube, and 30 µL of M-Elution Buffer applied to the filter. The column stood for 1 min at room temp and was then centrifuged for 30 s at full speed. A further 30 µL of M-Elution Buffer was applied to the filter and centrifuged again to obtain a final volume of 60 µL. Bisulphite converted DNA was stored at -30°C for up to 6 months.
2.9.4 PCR Optimisation

PCR optimisation was required before the primers could be used on the placental samples. Temperature gradient PCRs for each of the genes were carried out using control bisulphite converted DNA. Temperatures were selected which provided the strongest product band, but the least primer-dimer or multiple banding. Primer-dimer is known to interfere with the Sequenom assay and can inadvertently contribute to the amount of signal observed. Primer-dimer can also falsely increase the calibrating spectra quality, and thus decrease the accuracy of the assay overall.

Primer pairs were initially trialled at 56-67°C, and if results were ambiguous, then they were further tested at a reduced range of temperatures and with 5% DMSO. PCR products were electrophoresed on a 2% agarose gel stained with ethidium bromide to detect PCR product. The optimal annealing temperatures and primer sequences are displayed in Table 2.21.
Table 2.21: Sequenom PCR amplicons. Details for the amplicons and primers used for Sequenom validation of methylation.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Primer Name</th>
<th>Temp °</th>
<th>Amplicon (bp)</th>
<th># CpGs</th>
<th>Forward and Reverse Primers 5’-3’</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNLIPRP1</td>
<td>4580</td>
<td>60.6</td>
<td>496</td>
<td>23 (15)</td>
<td>AA AAT ATT TTT TGG TTA AGG TGA CC TAC TCC TCT AAA CCT CAA TTT CC</td>
</tr>
<tr>
<td></td>
<td>4581</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLG-AS1</td>
<td>4582</td>
<td>64.3</td>
<td>385</td>
<td>14 (9)</td>
<td>GA GGT ATT AGA TTT TTT TGG GGA TG</td>
</tr>
<tr>
<td></td>
<td>4583</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR1B</td>
<td>4586</td>
<td>60.6</td>
<td>486</td>
<td>34 (29)</td>
<td>GT TGT AGT TGG GGA ATT TGG GT CC AAT ACT AAA AAC AAA CCC CTC C</td>
</tr>
<tr>
<td></td>
<td>4587</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GREM1</td>
<td>4602</td>
<td>65</td>
<td>419</td>
<td>31 (24)</td>
<td>GT TGT TTT GGG TTT GTA GGG AA AC TCT ACC CTT ACC AAT CTC CAT CT</td>
</tr>
<tr>
<td></td>
<td>4603</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLC16A7</td>
<td>4604</td>
<td>62.8</td>
<td>469</td>
<td>4 (4)</td>
<td>TA TTT ATA GTG TGG TTT GGG TGG TT CC TAA CCT AAA TAA AAA CAA TCA AAA AA</td>
</tr>
<tr>
<td></td>
<td>4605</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNRPN</td>
<td>4606</td>
<td>55.9</td>
<td>509</td>
<td>13 (9)</td>
<td>TT TGG ATG ATG TTT TGA AAA TTA TTT ATG TC TCT CTA CAA TAA AAT CAA AAA AAA</td>
</tr>
<tr>
<td></td>
<td>4607</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>abParts</td>
<td>4608</td>
<td>61</td>
<td>460</td>
<td>12 (11)</td>
<td>GT TTA TGG TTT TGG AGT AGG TGG TT</td>
</tr>
<tr>
<td></td>
<td>4609</td>
<td></td>
<td></td>
<td></td>
<td>AC CAC AAT AAA CCC TCT AAT AAT CT</td>
</tr>
<tr>
<td>ADARB2</td>
<td>4610</td>
<td>64.9</td>
<td>425</td>
<td>14 (12)</td>
<td>TT TGG ATG ATG TTT TGA AAA TTA TTT ATG TC CTT CTA CAA TAA AAT CAA AAA AAA</td>
</tr>
<tr>
<td></td>
<td>4611</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAMSAP3</td>
<td>4612</td>
<td>63.3</td>
<td>363</td>
<td>23 (21)</td>
<td>AG GGA AGA TGG GTT TTT AGT TAT GT AA ACC TCC TCT TAA ACC AAT CTA CA</td>
</tr>
<tr>
<td></td>
<td>4613</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDH4</td>
<td>4614</td>
<td>63.3</td>
<td>233</td>
<td>24 (19)</td>
<td>TT TGG TTT TGG ATG GGG TAG TGT</td>
</tr>
<tr>
<td></td>
<td>4615</td>
<td></td>
<td></td>
<td></td>
<td>AC CCC CAA ACC TTA AAC AAA AA</td>
</tr>
<tr>
<td>CPLX1</td>
<td>4616</td>
<td>63.3</td>
<td>459</td>
<td>35 (21)</td>
<td>AT TGG TGT TTA TTT GTT GTG ATG TAT TT CT TAA CAT TAA CCT TCA TAA CCT TAC A</td>
</tr>
<tr>
<td></td>
<td>4617</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLJ26245</td>
<td>4618</td>
<td>66.1</td>
<td>371</td>
<td>15 (13)</td>
<td>GA AGT TAA GTA GGG TTT GTG TTT TTA AAT CTA GGG TTT TGG TTT GTG TTT TGG TTT</td>
</tr>
<tr>
<td></td>
<td>4619</td>
<td></td>
<td></td>
<td></td>
<td>AC CTA AAA CCT CTA ACC AAT AAT TT</td>
</tr>
<tr>
<td>GALNT9 (a)</td>
<td>4620</td>
<td>63.3</td>
<td>435</td>
<td>11 (9)</td>
<td>AT TGG TGT TTT TGG AGT AAT TTT TTA AAT CTA GGG TTT TGG TTT TTG TTT TGG TTT</td>
</tr>
<tr>
<td></td>
<td>4621</td>
<td></td>
<td></td>
<td></td>
<td>AA ACC TCC AAT AAC AAA ACA TCA AA</td>
</tr>
<tr>
<td>GALNT9 (b)</td>
<td>4622</td>
<td>61.2</td>
<td>322</td>
<td>12 (12)</td>
<td>AT TGG TAT TAG TGG TGG GTG GAA AGT AAA G</td>
</tr>
<tr>
<td></td>
<td>4626</td>
<td></td>
<td></td>
<td></td>
<td>TC CTC TAA CCC ACA AAA TAA AAA AA</td>
</tr>
<tr>
<td>GLOD5</td>
<td>4626</td>
<td>67.8</td>
<td>449</td>
<td>9 (8)</td>
<td>TT TTT ATT TGG TTT TGG ATG AGG GAG TGG TTT</td>
</tr>
<tr>
<td></td>
<td>4627</td>
<td></td>
<td></td>
<td></td>
<td>AA AAT AAT CTA CCC ACC TCA ACC TCA</td>
</tr>
<tr>
<td>GMPPA</td>
<td>4628</td>
<td>64.9</td>
<td>339</td>
<td>29 (28)</td>
<td>GG GTT TGT TTT AGG AAT TTT G</td>
</tr>
<tr>
<td></td>
<td>4629</td>
<td></td>
<td></td>
<td></td>
<td>AA TAC TAT ACT AAC CCA AAT CTA CCA ACC ACT</td>
</tr>
<tr>
<td>GNAS</td>
<td>4632</td>
<td>66.1</td>
<td>388</td>
<td>32 (25)</td>
<td>GG AGT TAT TTT TGG TAG AGT AAG G</td>
</tr>
<tr>
<td></td>
<td>4633</td>
<td></td>
<td></td>
<td></td>
<td>AA AAT ACC TCA AAA TCT ACC TCC TCT</td>
</tr>
<tr>
<td>KCNJ12</td>
<td>4636</td>
<td>66.2</td>
<td>515</td>
<td>15 (13)</td>
<td>TA GGT TTT TGG TGG GAG GTG AT</td>
</tr>
<tr>
<td></td>
<td>4637</td>
<td></td>
<td></td>
<td></td>
<td>TT ACA ATA CCC AAA ACC TTG TCC</td>
</tr>
<tr>
<td>KCNQ2</td>
<td>4638</td>
<td>67.0</td>
<td>477</td>
<td>16 (16)</td>
<td>GG AGT AAG GAA ATT GAG GTA TGG TAA CC CAC CTC CAA CTA AAA AAC TAA CT</td>
</tr>
<tr>
<td></td>
<td>4639</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OXCT1</td>
<td>4640</td>
<td>61.2</td>
<td>344</td>
<td>18 (16)</td>
<td>GG GTA TAA AAA GGA TTT TTA AGG GTT</td>
</tr>
<tr>
<td></td>
<td>4641</td>
<td></td>
<td></td>
<td></td>
<td>CC TCT AAA AAC CCA AAA CCT AAA AC</td>
</tr>
<tr>
<td>RHOBTB1</td>
<td>4646</td>
<td>61.2</td>
<td>493</td>
<td>45 (25)</td>
<td>TT AAT TTT GTA TAA AGG TAG TAT G</td>
</tr>
<tr>
<td></td>
<td>4647</td>
<td></td>
<td></td>
<td></td>
<td>AA AAC ACA AAT AAC CCA AAC TCT CTA C</td>
</tr>
</tbody>
</table>

81
2.9.5 PCR Conditions

Each PCR reaction mix was made up as a 10 µL volume so that 5 µL could be run on a gel to check for the presence of product prior to loading on the Sequenom platform. PCRs were set up as shown in Table 2.22. For negative controls, 2 µL of MilliQ H$_2$O was substituted for the DNA. Sequenom PCRs were performed on a BioRad DNAEngine® Peltier Thermal Cycler. Cycling conditions were as follows (Table 2.23). After PCR, 5 µL of each reaction was loaded onto a 2% agarose gel stained with ethidium bromide and run at 100 V for 30 min to confirm successful amplification of the correct product. The remaining 5 µL was stored at 4°C.

<table>
<thead>
<tr>
<th>Table 2.22: Sequenom PCR reactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x Hot Star Buffer</td>
</tr>
<tr>
<td>dNTPs (25 nM)</td>
</tr>
<tr>
<td>Hot Star Taq</td>
</tr>
<tr>
<td>Forward Primer</td>
</tr>
<tr>
<td>Reverse Primer</td>
</tr>
<tr>
<td>MilliQ H$_2$O</td>
</tr>
<tr>
<td>Bisulphite-converted DNA</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

2.9.6 Sequenom MALDI-TOF Reaction

Once the PCR reactions had been successfully completed, the samples were loaded into 384-well reaction plates (ABgene, Cat# AB-1384) and shipped on dry ice to the Sequenom Facility at the Liggins Institute in Auckland, NZ. Reactions were carried out by Chandrakanth Bhoothpur following the EpiTYPER® Application Protocols.
Table 2.23: Cycling Conditions for Sequenom PCRs

<table>
<thead>
<tr>
<th>Time</th>
<th>Temperature</th>
<th>Denaturation/Activation</th>
<th>Denaturation</th>
<th>Annealing</th>
<th>Elongation</th>
<th>Extension</th>
<th>Cooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 min</td>
<td>94°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 s</td>
<td>94°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x 40 Cycles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 s</td>
<td>(Optimal annealing temperature)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 min</td>
<td>72°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 min</td>
<td>72°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.9.7 Sequenom Data Analysis

Data were received from the Liggins institute in the form of .csv files containing a quantified methylation proportion for each sample at each CpG site. The EpiTYPER® software calculates this proportion based on the mass of methylated fragments compared to unmethylated fragments of DNA. Due to the chemistry of the Sequenom reaction, sometimes a fragment contains more than one CpG site, depending on where in its sequence the DNA can be cleaved. For the purpose of analysis, these sites were treated as a single site. Various analyses were performed on the data received; methylation across the amplicons, mean methylation in cases and controls, standard deviations, and confidence intervals. Particular focus was paid to the CpG site(s) that were significantly different between cases and controls in the RRBS analysis, to determine how well the differences were validated by Sequenom.
Chapter 3

Epigenetic Analysis of Two Putative Imprinted Genes - *IGFBP1* and *ZNF264*

This chapter addresses the results of aim 4a of the project: To analyse the two potential Differentially Methylated Regions *IGFBP1* and *ZNF264* in the context of a normal tissue in whole blood. Single Nucleotide Polymorphisms (SNPs) of high frequency in the general population have been identified, and used to differentiate the two alleles from each individual for two amplicons covering the promoter and exon 1 of both genes. Cloning combined with bisulphite conversion was then used to determine the methylation status of each allele. These methylation data have been analysed for two main effects: Allele-Specific-Methylation (ASM) in which the methylation state of individual CpG sites or the amplicon as a whole are linked to the specific SNP allele; and Parent-Of-Origin effects, in which the methylation state of an allele is determined by which parent it was inherited from.
3.1 Sample Information

In all, 46 samples were collected from three New Zealand-based sources:

- 17 whole blood samples from unrelated individuals were obtained via venipuncture of the median cubital vein.
- The Merriman Lab, Department of Biochemistry, University of Otago provided the remainder of the samples. These DNA samples were extracted from whole blood, and collected as part of previous genetic-based family studies. The donors provided their consent for their DNA to be used in other approved studies
  - 7 samples of DNA from two families, from the New Zealand Insulin-Dependent Diabetes Cohort (NZIDDM)
  - 21 samples from 7 families from the Thyroid Genetics Study (TGS)

All blood samples were New Zealand/European in origin

3.1.1 DNA Extraction

Extraction of DNA from blood was performed within 24 hours of collection, with the exception of sample X9019.03, which was stored at 4°C for a week prior to extraction. Concentrations of DNA were variable (Table 3.1), but as all downstream uses of the DNA would be subject to PCR, this was not of concern. The concentration of DNA samples received from the Merriman Lab were all 200 ng/µL.
**Table 3.1:** List of Samples and Concentration of DNA extracted from whole blood. The suffix .01 denotes the father’s sample, and .02 the mother’s.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Concentration (ng/µL)</th>
<th>Sample ID</th>
<th>Concentration (ng/µL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X9015</td>
<td>65.6</td>
<td>X9033.02 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9016</td>
<td>10.6</td>
<td>X9034 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9017</td>
<td>29.7</td>
<td>X9035 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9018</td>
<td>20.7</td>
<td>X9034.02 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9019</td>
<td>555.83</td>
<td>X9036 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9020</td>
<td>239.1</td>
<td>X9036.01 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9021</td>
<td>690.3</td>
<td>X9036.02 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9022</td>
<td>50.15</td>
<td>X9037 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9023</td>
<td>966.85</td>
<td>X9037.02 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9024</td>
<td>18.47</td>
<td>X9038 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9025</td>
<td>1204.58</td>
<td>X9038.02 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9026</td>
<td>13.03</td>
<td>X9039 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9027</td>
<td>27.71</td>
<td>X9039.01 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9028</td>
<td>12.94</td>
<td>X9040 (TGS)</td>
<td>200</td>
</tr>
<tr>
<td>X9029</td>
<td>34.2</td>
<td>X9041 (NZIDDM)</td>
<td>200</td>
</tr>
<tr>
<td>X9030</td>
<td>10.53</td>
<td>X9041.01 (NZIDDM)</td>
<td>200</td>
</tr>
<tr>
<td>X9031 (TGS)</td>
<td>200</td>
<td>X9042 (NZIDDM)</td>
<td>200</td>
</tr>
<tr>
<td>X9031.01 (TGS)</td>
<td>200</td>
<td>X9042.01 (NZIDDM)</td>
<td>200</td>
</tr>
<tr>
<td>X9031.02 (TGS)</td>
<td>200</td>
<td>X9042.02 (NZIDDM)</td>
<td>200</td>
</tr>
<tr>
<td>X9032 (TGS)</td>
<td>200</td>
<td>X9043 (NZIDDM)</td>
<td>200</td>
</tr>
<tr>
<td>X9032.01 (TGS)</td>
<td>200</td>
<td>X9044 (NZIDDM)</td>
<td>200</td>
</tr>
<tr>
<td>X9032.02 (TGS)</td>
<td>200</td>
<td>X9045</td>
<td>1071.99</td>
</tr>
<tr>
<td>X9033 (TGS)</td>
<td>200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TGS - Samples were obtained from the Thyroid Genetics Study
NZIDDM - Samples were obtained from the New Zealand Insulin-Dependent Diabetes Cohort
3.2 Genotyping for SNPs

The first step of this project was to optimise PCR protocols both of genomic DNA and of bisulphite converted DNA. Several primer pairs were tested for each amplimer to determine which primer set amplified most reliably. Primers were designed to SNP-containing fragments from both the promoter and exon 1 for \textit{IGFBP1} and \textit{ZNF264} (Figure 3.2). These sites were amplified for each sample (Figure 3.1). Primer dimer was observed for the Exon 1 amplicon of \textit{ZNF264} which was removed during clean up (Figure 3.1B). Non-specific binding was observed as a lighter band directly below the 650 bp band produced by the PCR amplification of Exon 1 of \textit{IGFBP1} (Figure 3.1C). These non-specific products were not observed upon later sequencing analysis of the products.
**Figure 3.1:** Successful amplification of genomic DNA for genotyping. (A) Lanes 1-17, *IGFBP1* Exon 1 - Samples X9015-X9030 + negative control (650 bp); Lanes 18-34, *ZNF264* Exon 1 - Samples X9015-X9030 + negative control (829 bp). (B) Lanes 1-15, *ZNF264* Exon 1 - Samples X9031-X9044 + negative control (829 bp). (C) Lanes 1-15, *IGFBP1* Exon 1 - Samples X9031-X9044 + negative control (650 bp). (D) Lanes 1-32, *IGFBP1* Promoter - Samples X9015-X9045 + negative control (702 bp); Lanes 33-64, *ZNF264* Promoter - Samples X9015-X9045 + negative control (636 bp). Markers are Invitrogen 1 Kb Plus DNA Ladder, arrows indicate 500bp.
**Figure 3.2**: *IGFBP1* and *ZNF264* gene context. A) *IGFBP1* B) *ZNF264*. Inset is the promoter/exon 1 region of *ZNF264* for greater clarity. Genotyping amplicons are noted in purple, Methylation-specific amplicons in red. Promoter associated CpG islands (CGI) are shown in green. The SNPs for which each individual was genotyped are shown in blue.
3.2.1 Restriction Fragment Length Polymorphism Analysis

Restriction Fragment Length Polymorphism (RFLP analysis) was trialled as a rapid method of SNP genotyping. The Exon 1 SNPs rs9658194 (IGFBP1) and rs58278481 (ZNF264) both incorporate a restriction enzyme recognition site. Although these digestions were performed upon 9 different occasions, they did not provide consistent results for genotyping. A representative example of the digest patterns that were observed is illustrated in Figure 3.3. Genotyping of rs9658194 using BsrDI showed results that were inconsistent, and did not always cut the product to the expected size. A genotype of CC should have yielded fragments of 504 and 146 bp; AA should show one fragment of 650 bp; and AC should have given three fragments of 650, 504 and 146 bp. In Figure 3.3A, only three genotypes drawn from digestion with BsrDI matched those found by sequencing. Genotyping of rs58278481 showed consistent results, but no variation between the samples (See Figure 3.3B). Digestion of ZNF264 exon 1 PCR fragments with Bsu36I suggested that samples X9015-9030 all carried the GG genotype at rs58278481 as a genotype of GG should have yielded fragments of 652 and 177 bp; CC should have given three fragments of 516, 177, and 136 bp; and GC should have given fragments of sizes 652, 177, and 136 bp. When the putative genotypes gleaned from RFLP analysis were compared to those obtained by sequencing, there was also a very poor correlation. Possible reasons for this may be non-specific enzyme activity, or the presence of some factor in the DNA sequence that inhibits enzymatic cleavage. Genotyping by sequencing was adopted as the strategy from this point onwards.

3.2.2 Genotyping by Sequencing

As an alternative strategy, the PCR products from Figure 3.1 were sent for sequencing by capillary-separation on an ABI 3730 xl DNA Analyser. The chromatograms were examined to determine the SNP genotype of each sample. All proband samples were genotyped for each of the four SNPs interrogated. The parental samples were genotyped only for the promoter region SNPs, rs1065780 (IGFBP1) and rs3810136
Figure 3.3: RFLP analysis of X9015-X9030 + water control = Lanes 1-17 (A) Digestion of *IGFBP1* Exon 1 PCR fragment with BsrDI to determine the genotype of rs9658194. (B) Digestion of *ZNF264* Exon 1 PCR fragment with Bsu36I to determine the genotype of rs58278481. Markers are Invitrogen 1 Kb Plus DNA Ladder, arrows indicate 500bp. *IGFBP1*: CC - 504 bp and 146 bp; AA - 650 bp; AC - 650, 504, and 146 bp. *ZNF264*: GG - 652 bp and 177 bp; CC - 516 bp, 177 bp and 136 bp; GC - 652 bp, 177 bp and 136 bp.

The results are outlined in Table 3.2. Heterozygous SNPs in DNA sequence can be identified in chromatograms by the presence of two bases at a single location where the two alleles differ in sequence (Figure 3.4). Of the 46 samples, 15 were heterozygotic and informative for the promoter of *IGFBP1*, and 10 for exon 1 of *IGFBP1*. Seven samples were informative for the promoter of *ZNF264*, and 16 were informative for exon 1 of *ZNF264*.

Figure 3.4: Chromatogram trace of *ZNF264* rs3810136, illustrating the difference between the homozygous GG and heterozygous GA alleles. In the heterozygous samples a double peak (R) is seen at the SNP site, evidence that there are two different alleles present in the sample.
Table 3.2: SNP genotyping. The suffix .01 denotes the father’s sample, and .02 the mother’s.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>rs9658194</th>
<th>rs1065780</th>
<th>rs7250511</th>
<th>rs3810136</th>
</tr>
</thead>
<tbody>
<tr>
<td>X9015</td>
<td>A/C</td>
<td>A/G</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9016</td>
<td>CC</td>
<td>A/G</td>
<td>TT</td>
<td>GG</td>
</tr>
<tr>
<td>X9017</td>
<td>CC</td>
<td>A/G</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9018</td>
<td>CC</td>
<td>GG</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9019</td>
<td>A/C</td>
<td>A/G</td>
<td>CC</td>
<td>A/G</td>
</tr>
<tr>
<td>X9020</td>
<td>CC</td>
<td>GG</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9021</td>
<td>CC</td>
<td>GG</td>
<td>TT</td>
<td>GG</td>
</tr>
<tr>
<td>X9022</td>
<td>A/C</td>
<td>A/G</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9023</td>
<td>CC</td>
<td>A/G</td>
<td>TT</td>
<td>A/G</td>
</tr>
<tr>
<td>X9024</td>
<td>A/C</td>
<td>GG</td>
<td>TT</td>
<td>GG</td>
</tr>
<tr>
<td>X9025</td>
<td>CC</td>
<td>A/G</td>
<td>C/T</td>
<td>A/G</td>
</tr>
<tr>
<td>X9026</td>
<td>CC</td>
<td>GG</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9027</td>
<td>CC</td>
<td>GG</td>
<td>C/T</td>
<td>A/G</td>
</tr>
<tr>
<td>X9028</td>
<td>A/C</td>
<td>GG</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9029</td>
<td>CC</td>
<td>AA</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9030</td>
<td>A/C</td>
<td>GG</td>
<td>TT</td>
<td>GG</td>
</tr>
<tr>
<td>X9031</td>
<td>A/C</td>
<td>AA</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9032</td>
<td>CC</td>
<td>GG</td>
<td>TT</td>
<td>GG</td>
</tr>
<tr>
<td>X9032.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X9033</td>
<td>A/C</td>
<td>A/G</td>
<td>CC</td>
<td></td>
</tr>
<tr>
<td>X9034</td>
<td>A/C</td>
<td>A/G</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9034.02</td>
<td>AA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X9035</td>
<td>CC</td>
<td>A/G</td>
<td>CC</td>
<td>GG</td>
</tr>
<tr>
<td>X9036</td>
<td>CC</td>
<td>GG</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9037</td>
<td>CC</td>
<td>A/G</td>
<td>TT</td>
<td>GG</td>
</tr>
<tr>
<td>X9038</td>
<td>CC</td>
<td>A/G</td>
<td>CC</td>
<td>GG</td>
</tr>
<tr>
<td>X9039</td>
<td>CC</td>
<td>A/G</td>
<td>CC</td>
<td>GG</td>
</tr>
<tr>
<td>X9039.01</td>
<td>GG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X9040</td>
<td>CC</td>
<td>A/G</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9041</td>
<td>A/C</td>
<td>A/G</td>
<td>CC</td>
<td>GG</td>
</tr>
<tr>
<td>X9042</td>
<td>CC</td>
<td>GG</td>
<td>C/T</td>
<td>GG</td>
</tr>
<tr>
<td>X9042.01</td>
<td></td>
<td></td>
<td></td>
<td>GG</td>
</tr>
<tr>
<td>X9042.02</td>
<td></td>
<td></td>
<td></td>
<td>GG</td>
</tr>
<tr>
<td>X9043</td>
<td>CC</td>
<td>GG</td>
<td></td>
<td>GG</td>
</tr>
<tr>
<td>X9044</td>
<td>CC</td>
<td>GG</td>
<td>C/T</td>
<td>A/G</td>
</tr>
<tr>
<td>X9045</td>
<td>CC</td>
<td>GG</td>
<td>CC</td>
<td>A/G</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IGFBP1</th>
<th>ZNF264</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs9658194</td>
<td>rs1065780</td>
</tr>
</tbody>
</table>


3.3 PCR of Bisulphite-Converted Informative Samples

PCR amplification of bisulphite-treated DNA is less reliable than that of genomic DNA. The bisulphite treatment can degrade and fragment the genomic DNA. Additionally, the conversion of all unmethylated non-CpG cytosine residues to thymine residues results in a comparatively low GC content for the converted strands. For these reasons, several combinations of primers were trialled for each amplimer. The annealing temperature was also reduced to 49-51°C to reduce the binding threshold for the primers. The primers were made as long as possible to counteract the low-specificity of the PCR conditions, but their length was constrained by the need to avoid incorporating CpG sites into the primers. For one primer sequence, mIFBP1F3, it was necessary to incorporate a CpG site, so the primer ordered consisted of a mixture of C and T (Y) at that site to allow for variable methylation. Those samples that were found to be heterozygotic at one or more of the SNP sites were treated with bisulphite and then amplified by PCR.
Table 3.3: Informative samples for each of the four amplimers

<table>
<thead>
<tr>
<th>IGFBP1</th>
<th>ZNF264</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Promoter</td>
<td>Exon 1</td>
</tr>
<tr>
<td>X9015</td>
<td>X9015</td>
</tr>
<tr>
<td>X9016</td>
<td>X9019</td>
</tr>
<tr>
<td>X9017</td>
<td>X9022</td>
</tr>
<tr>
<td>X9019</td>
<td>X9024</td>
</tr>
<tr>
<td>X9022</td>
<td>X9028</td>
</tr>
<tr>
<td>X9023</td>
<td>X9030</td>
</tr>
<tr>
<td>X9025</td>
<td>X9031</td>
</tr>
<tr>
<td>X9033</td>
<td>X9033</td>
</tr>
<tr>
<td>X9034</td>
<td>X9034</td>
</tr>
<tr>
<td>X9035</td>
<td>X9041</td>
</tr>
<tr>
<td>X9037</td>
<td></td>
</tr>
<tr>
<td>X9038</td>
<td></td>
</tr>
<tr>
<td>X9039</td>
<td></td>
</tr>
<tr>
<td>X9040</td>
<td></td>
</tr>
<tr>
<td>X9041</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.5: PCRs of bisulphite converted DNA. A) IGFBP1 Exon 1 (Amplicon size 317 bp). Lanes 1-9: X9015, X9019, X9022, X9024, X9028, X9030, X9033, X9034, water control. B) ZNF264 Exon 1 (Amplicon size 295 bp). Lanes 1-8: X9017, X9018, X9020, X9027, X9031, X9034, X9036, X9040 (water control not shown) C) IGFBP1 Promoter (Amplicon size 329 bp). Lanes 1-15: X9015, X9016, X9017, X9019, X9022, X9023, X9025, X9033, X9034, X9035, X9037, X9038, X9039, X9040, X9041, water control. D) ZNF264 Promoter (Amplicon size 328 bp). Lanes 1-7: X9019, X9023, X9025, X9027, X9033, X9044, X9045, water control. Ladders are Invitrogen 1 Kb Plus DNA Ladder, arrows indicate 500 bp. A few samples for ZNF264 failed to amplify.

3.4 Cloning

The purified bisulphite PCR products were cloned into Mach1 competent cells that were grown on LB agar with selection by ampicillin. In most cases, a sufficient number of colonies for downstream analysis (10-30) grew after overnight incubation.

3.4.1 Screening of Clones for Correct Inserts

Two approaches were taken to screening colonies for insertion of the correct DNA fragment. Enzyme digestion of the purified 4 kb plasmid with EcoRI (Figure 3.6) was compared to PCR screening of the unpurified colonies. Both techniques were found
to be adequate for the purpose, but the majority of screening was done by PCR for speed and ease of processing. The insert size depends on the method of screening, and the gene segment being cloned. Sizes of the amplicons can be seen in Table 2.8. Restriction digests add 18 bp (of vector) to the length of the insert, while PCR adds 167 bp (of vector) to the length of the PCR product.

Figure 3.6: A representative example of EcoRI digest of 35 purified 4 kb plasmids transfected with either the exon 1 amplicon of IGFBP1 from one individual (row 1, lanes 7-11) or the exon 1 amplicon of ZNF264 from four individuals (All remaining lanes). Of the 35 clones screened here, 14 contained the correct insert of size of 335 bp for exon 1 IGFBP1 (317 + 18 bp from vector, marked with a red asterisk), and 313 bp for exon 1 ZNF264 (295 + 18 bp from vector, marked with a blue asterisk). Ladders are Invitrogen 1 Kb Plus DNA Ladder, arrows indicate 500 bp.

3.4.2 Plasmid Extraction and Sequencing

For sequencing purposes, it is necessary to extract the plasmids only from the E.coli clones. Plasmid minipreps are designed to purify plasmids from E.coli at high concentration. For capillary sequencing, a concentration of 150 ng/5 µL is required. Initial minipreps did not yield this concentration, so an additional incubation step was introduced after the overnight incubation (for detail on the miniprep protocol see Section
2.4.3.3). Adding this incubation increased the yield two-fold on average (from \( \sim 15 \) ng/µL to \( \sim 30 \) ng/µL), enough to ensure reliably clean, analysable sequence. Sequencing was performed using the M13 reverse primer, located within the pCR\(^\circ\)4-TOPO vector. Vector sequences were trimmed from the resulting sequence files using Geneious Software (Kearse \textit{et al.}, 2012), as stated in Section 2.4.5.

### 3.5 Analysis of Clones

The methylation analysis of the clones took place in two parts - Allele-Specific Methylation, and Parent-of-Origin Specific Methylation. The promoter and exon 1 regions for both \textit{IGFBP1} and \textit{ZNF264} were analysed with respect to those two parts, with an emphasis on the promoter regions. The two different alleles for each amplicon will henceforth be referred to by the allele of the SNP polymorphism with which they are associated, i.e. \textit{IGFBP1} exon 1 has an A allele and a C allele, while the \textit{IGFBP1} pro-
Table 3.4: Clones generated from the DNA of individuals that are informative for the promoter regions of IGFBP1 and ZNF264. Allele refers to the SNP allele used to differentiate the two copies of DNA. For mean methylation across the amplicon, the average methylation across all CpGs in each clone was compared. Site-by-Site analyses calculated the average methylation across all clones for each CpG site within an amplicon.

<table>
<thead>
<tr>
<th>Gene</th>
<th>IGFBP1</th>
<th>ZNF264</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>X9015</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>X9016</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>X9017</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>X9019</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>X9022</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>X9023</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>X9025</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>X9027</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>X9033</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>X9034</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>X9035</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>X9037</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>X9038</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>X9039</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>X9040</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>X9041</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>X9045</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Totals</td>
<td>133</td>
<td>117</td>
</tr>
</tbody>
</table>
moter has an A allele and a G allele. Methylation is calculated in two ways - averaged across the amplicon, and per methylation site (Figure 3.8).

![Figure 3.8](image)

**Figure 3.8:** Calculation of Methylation for an individual at the IGFBP1 promoter. This amplicon contains a SNP polymorphism at the first CpG site (A allele). Blue circles denote unmethylated CpG sites, while red ones denote methylated sites. The presence or absence of this SNP allele affects the methylation percentage outcome.

### 3.5.1 Allele-Specific Methylation

Analyses of methylation were performed using GraphPad Prism® data analysis software. The type of analysis performed depended on the data - whether the whole amplicon was being analysed, or the individual sites within the amplicon were being examined. Initially, information on methylation at each CpG site for each clone was converted into binary data (0 = unmethylated, 1= methylated), and further analyses were carried out from here. Four analyses were carried out for each amplicon: 1) Mean
methylation across the amplicon, 2) Mean methylation for each CpG site, 3) Mean amplicon methylation for each individual, and 4) Site by site methylation for each individual.

1) For mean methylation across the amplicon, mean methylation was calculated for each clone, and then these values were segregated depending on their SNP allele. A Wilcoxon matched-pairs signed rank test was used, pairing the alleles from each individual.

2) Mean site-by-site analyses were performed on the mean methylation at each single CpG site.

3) For each individual, their mean methylation for each allele was then calculated for the whole amplicon, and these values analysed using a Wilcoxon matched-pairs signed rank test, pairing the alleles from each individual.

4) Mean methylation at each single CpG site in each individual was calculated as the proportion of clones from each individual that were methylated. Calculating standard deviations for proportions can be problematic, due to sampling bias, so the standard error of the mean (SEM) was calculated instead using Equation 3.1, where \( p \) is the sample proportion and \( n \) is the sample size.

\[
SE_p = \sqrt{\frac{p(1 - p)}{n}}
\]  

Analyses 2, 3 and 4 were performed using multiple t-tests, one per row. A row either contained all of the methylation data from every sample for a single CpG site, or all of the methylation data from a single sample for analysis 3. Correction for multiple testing was performed using the Holm-Sidak method. Results were deemed significant when they show \( p < 0.05 \) adjusted for the number of tests. The number of clones that were tested for each amplicon can be seen in Table 3.4. The number of CpG sites analysed for each amplicon varies. The promoter amplicon for IGFBP1 consists of 6 CpGs, plus an additional CpG depending on the SNP allele, which occurs at a CpG dinucleotide. The exon 1 amplicon for IGFBP1 comprises 18 CpG sites. For the two ZNF264 amplicons, 16 CpG sites were analysed in the promoter, and 19 in the first
3.5.1.1 *IGFBP1* Promoter

*IGFBP1* did not show a significant difference in mean methylation between the SNP alleles (Figure 3.9A) - however, there was a possibility that this was an artefact. The SNP used to differentiate alleles in this amplicon introduces a CpG site if the allele is a guanine. Since amplicons carrying the G allele containing an additional CpG that is not present in the the A polymorphism, an additional analysis was performed in which the polymorphic CpG site was excluded. Removing the data for the first CpG site did not change the significance between the two alleles (Figure 3.9A - columns 3 and 4), and slightly reduced the mean difference between the alleles from 3.6% to 1.8%.

![IGFBP1 Allele Specific Methylation](image1)

**Figure 3.9:** Allele-Specific Methylation in the Promoter of A) *IGFBP1* an B) ZNF264. For IGFBP1 methylation is an average across the amplicon (7CpGs) when including the SNP CpG site, and 6 CpGs when the SNP is excluded. Methylation difference approached significance for ZNF264. Methylation is an average across the amplicon (16 CpGs). Statistical analysis was a Wilcoxon matched-pairs signed rank test between mean methylation of the two alleles for each sample.
When analysed site-by-site, no significant differences between the alleles were detected (Figure 3.10A). The highest difference in methylation occurred at CpG 5, which showed a difference of 8.3%, at a p-value of 0.07.

When allele-specific methylation was examined in individual samples however, some differences were observed (Figure 3.11). Samples X9025, X9039, X9040 and X9041 all showed a significant difference between alleles when the entire amplicon was analysed, with X9025, X9039, and X9041 showing higher methylation in the G allele (17.3%, 23.0%, and 28.2% differences respectively), and X9040 higher in the A allele (30.5% difference). X9039 and X9040 are from the same family, and both inherited their A allele from their mother, so it is of note here that they do not show the same allele-specific methylation pattern. When the polymorphic CpG site is excluded, significance remained for samples X9040 and X9041, and the direction of methylation difference did not change, although the mean methylation values did change slightly for the G allele (28.2% and 23.8% differences respectively).

\textit{IGFBP1}

<table>
<thead>
<tr>
<th>Cpg Site</th>
<th>G Allele</th>
<th>A Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>0.7</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>0.4</td>
<td>0.1</td>
</tr>
</tbody>
</table>

\textbf{Figure 3.10:} Allele-specific site-by-site methylation in the promoters of A) \textit{IGFBP1} and B) \textit{ZNF264}. There was no evidence of allele-specific methylation at any site in the Promoter of \textit{IGFBP1}. Methylation is an average across all individuals. A allele n=133, G allele n=117 clones. \textit{ZNF264} did show a significant difference at CpG 3 (*p-value = 0.002). CpG 6 also approached significance (p-value = 0.041). Methylation is an average across all individuals. Statistical analysis was by multiple t-tests, with adjustment for the number of tests - 6 tests for IGFBP1 and 16 tests for ZNF264 (one per CpG site, Holm-Sidak adjusted).
Figure 3.11: Allele-specific methylation in individual samples for IGFBP1. A) Average methylation of A and G alleles in informative samples for the whole amplicon (7 CpG sites). B) Average methylation of A and G alleles in informative samples for CpGs 2-7 (excluding the SNP CpG site). Statistical analysis was performed by multiple t-tests (one per individual), with Holm-Sidak correction for multiple testing. * = p < 0.018

A site-by-site analysis of the individuals who were informative for the promoter of IGFBP1 revealed for most samples a similar pattern to that shown in Figure 3.10A, with higher methylation in CpGs 2, 3 and 7, and very little methylation at sites 4, 5 and 6 (Figure 3.12). Samples X9016, X9025, and X9039 each had a single CpG site that shows significantly higher methylation in the G allele. X9040 shows significantly higher methylation in the A allele at sites 4 and 7, with raised methylation on this allele at other sites also. These site differences contribute strongly to the differences seen in Figure 3.11.
Figure 3.12: Site-by-Site Analysis of Individuals Informative for the *IGFBP1* Promoter. Methylation percentage is displayed on the Y-axis, and CpG site within the amplicon along the X-axis. Four individuals show significance when analysed with multiple t-tests (one per site), X9016, X9025, X9039, and X9040. Sites are deemed significant when they show \( p < 0.05 \) Holm-Sidak adjusted for the number of tests (one per CpG site - 6 sites with information for both alleles).
3.5.1.2 ZNF264 Promoter

ZNF264 did not show a significant difference in mean methylation between the SNP alleles, (p=0.09) (Figure 3.9B). The mean difference between alleles was 5.23% however, so statistical significance would be unlikely to be linked to biological significance.

When analysed site-by-site, differences between the alleles were present (Figure 3.10B). The slightly higher mean methylation of the G allele seen in Figure 3.9B is coming from just a few CpG sites. CpG 3 displayed a difference in methylation of 30.6% between the two alleles. This difference was statistically significant (p=0.002, significance adjusted using the Holm-Sidak method for 16 tests). Two other CpGs also showed a difference between the alleles of greater than 10%, and one of these, CpG 6, also approached significance (p=0.041, significance using the Holm-Sidak method adjusted for 16 tests). For each of these sites, it was the G allele that showed the higher level of methylation.

![ZNF264 Promoter Allele Specific Methylation in Informative Individuals](image)

**Figure 3.13:** Allele-specific methylation in individual samples for ZNF264. Average methylation of A and G alleles in informative samples for the whole amplicon (16 CpG sites). Statistical analysis was performed by multiple t-tests (one per individual), with Holm-Sidak correction for multiple testing. Significance = p <0.05 before correction for 16 tests.
A sample-wise analysis of the informative individuals for the ZNF264 promoter did not show any significant differences (Figure 3.13), though in 5 of the 6 samples analysed, the G allele showed higher methylation than the A allele. The A allele shows a much more stable methylation pattern, varying by only 6% between individuals (17.7-23.8%), while the G allele shows a much wider range (14.0-39.0%).

Only a single site for a single sample shows a large and significant difference between alleles when a site-by-site analysis is performed on each individual samples (Figure 3.14). CpG site 3 in sample X9027 has 28.5% methylation in the A allele, and 100% methylation in the G allele - a difference of 72.5% (p=0.0006). Several other sites in X9027, as well as others in X9025 and X9033 also approached significance, with differences larger than 50% and p-values less than 0.05, but these were not significant after correction for multiple testing.
Figure 3.14: Site-by-site analysis of individuals informative for the ZNF264 promoter. Methylation percentage is displayed on the Y-axis, and CpG site within the amplicon along the X-axis. Individual X9027 shows significance at CpG site 3 when analysed with multiple t-tests (one per site, Holm-Sidak correction). Sites are deemed significant when they show $p<0.05$ before Holm-Sidak adjustment for multiple testing (16 tests per sample).
3.5.1.3 *IGFBP1* Exon 1

Methylation found in the exon 1 amplicon of *IGFBP1* was low, with little difference between the A and C allele-associated clones. The A allele sequences showed 4.5% methylation, while C allele sequences showed 2.3% methylation (Figure 3.15A). This is in contrast to the data reported by Martín-Subero *et al.* (2009), which showed a 37.3% methylation level in peripheral blood samples. The site analysed on the Illumina array (chr7:45928007) used by Martín-Subero *et al.* (2009) is in the 5’ UTR of the gene, whereas the amplicon used for cloning was at the 3’ end of the first exon, which may account for the discrepancy between datasets. A small number of clones showed methylation at between 1 and 3 sites across the amplicon (18 CpG sites). Only 4 clones were recovered that contained the A allele: this may in part account for the lack of difference between the alleles.

When analysed site-by-site, a little more was learned. CpGs 1 and 15-18 showed a small amount of methylation, and the A allele at CpG 15 showed 25% methylation, compared to 4% in the T allele, a significant difference (p=0.002).

![Figure 3.15: Methylation in Exon 1 of A) *IGFBP1*. Methylation is an average across the amplicon (19 CpGs). Statistical analysis was a t-test between mean amplicon methylation of each clone for each allele. A allele n=4, and C allele n=44 clones. B) *ZNF264*. This amplicon did not yield any informative clones (n=20 clones), but average methylation for this amplicon was <2%.](image-url)
3.5.1.4 ZNF264 Exon 1

Once the exon 1 region of ZNF264 was bisulfite converted, the alleles could no longer be differentiated, as it was a C/T polymorphism. This polymorphism would have been detectable on the reverse strand (See Figure 4.10 for an illustration of this principle), but unfortunately not enough of the clones generated were in this orientation to differentiate the alleles. The data gathered from this amplicon does however agree with Martín-Subero et al. (2009) that methylation levels in exon 1 of ZNF264 are very low in peripheral blood (Figure 3.15B). Unfortunately, due to the CpG-rich nature of the region, it was impossible to design an amplicon that overlapped the site probed by the Illumina array, but the amplicon used does lie within 300 bp of the Illumina site (chr19:57702916). Of the 20 clones from four individuals (X9015, X9018, X9020, and X9027) analysed for this 19-CpG amplicon, only 4 contained a single methylated CpG site, despite the fact that the amplicon covers an intron-exon boundary. Mean methylation for this amplicon is 1.605%, compared to Martín-Subero’s 0.069% - a non-significant difference. When analysed site-by-site, little more was learned, save that CpGs 1 and 15-18 showed a small amount of methylation. No site displayed higher than 10% methylation. Because of the very low levels of methylation observed in all of the preliminary data gathered, this amplicon was not pursued further.

3.5.2 Parent-of-Origin Effects

Only a small proportion of the samples formed informative family groups, so conclusions drawn from their analysis are perforce somewhat tentative. Four families were found to be informative for the promoter region of IGFBP1 (Figure 3.17). For ZNF264, there was a single heterozygous individual among the family group samples, but the mother’s sample could not be amplified for genotyping.
Figure 3.16: Methylation in Exon 1 of A) IGFBP1. A Allele n=4, and C allele n=44 clones. Statistical analysis was multiple t-tests between mean CpG site methylation of clones for each allele, with Holm-Sidak correction for multiple testing. CpG site 15 showed statistical significance (p=0.0025). B) ZNF264. This amplicon did not yield any informative clones, but the amplicon was almost entirely unmethylated, with small peaks at CpGs 2, 8, and 17-18. n=20 clones for both alleles combined.

Figure 3.17: Informative pedigrees for the promoter of IGFBP1. Six individuals from four families were informative for the promoter amplicon of IGFBP1. The genotype for individual X9038.02 was not able to be ascertained.
3.5.2.1 *IGFBP1* Promoter

For the four family groups that were informative for *IGFBP1*, no significant difference was observed when the whole amplicon was analysed (Figure 3.18A). The paternally inherited allele showed slightly lower methylation than the maternally inherited copy. However, it should be noted that in three of the four families it was the G allele that was paternally inherited, so the difference observed may be as a result of the same artefact observed when examining allele-specific methylation. Therefore, as with the analysis for allele-specific methylation, the same analysis was performed with the exclusion of the initial CpG site. This exclusion did not significantly change the outcome. Mean amplicon methylation was analysed using a Wilcoxon signed-rank test.

![Figure 3.18: Mean amplicon methylation and mean CpG site methylation in the promoter of IGFBP1 by Parent-of-Origin. A) Mean amplicon methylation in maternal and paternal alleles of the IGFBP1 promoter, including and excluding CpG site 1. Significance = p<0.05 by Wilcoxon signed-rank test. B) Site-by-site methylation in maternal and paternal alleles of the IGFBP1 promoter. CpG site 1 is the location of the polymorphic CpG site. Significance = p<0.05 by multiple t-tests before adjustment using the Holm-Sidak method.](image)

When analysed site-by-site however, some differences were seen (Figure 3.18B). CpG 1 cannot be treated the same as the other CpG sites in this context, as the SNP
polymorphism affects the presence of the CpG site. In these families, five of the six informative individuals inherited the A allele from their mother, meaning that only a very small proportion of the clones from the maternal allele actually contain the CpG site. CpG site 2 demonstrates a 21% reduction in methylation from 96% in the maternal allele to 75% in the paternally inherited allele (p=0.003), and CpG site 4 shows the same pattern from 22% in the maternal allele to 0% in the paternally inherited allele (p=0.0006).

No individual sample shows significant differences between the maternal and paternal alleles (Figure 3.19). Differences of up to 30% were observed, but variance was high.
Figure 3.19: Parent-of-Origin specific methylation of the promoter of IGFBP1 in informative individuals including (A) and excluding (B) the polymorphism at CpG 1. No individual sample showed a significant difference between the two parental alleles. Error bars represent a 95% confidence interval of the mean.

Two samples showed sites with significant and large differences between maternal and paternal alleles when analysed by multiple t-tests with correction for multiple testing using the Holm-Sidak method (Figure 3.20). X9039 displays 60% methylation in the paternal allele at CpG 5, but no methylation in the maternal allele (p=0.004). CpG 6 also approached significance (p=0.036). X9040 shows significantly higher methylation in the maternal allele at CpG sites 4 and 7 (p=0.003, and p=0.002 respectively), and approaches significance at CpGs 2 and 5 (p=0.018 for both). Interestingly, X9039 and X9040 are children of the same pairing, and inherited the G allele from their father, but show an inverse pattern of methylation at CpG sites 5-7.
**Figure 3.20:** Parent-of-Origin, site-by-site analysis of individuals informative for the IGFBP1 promoter. Due to the presence of the polymorphism at CpG 1, only one allele contains this site. Statistical analysis within individuals was performed via multiple t-tests, one per CpG site. Significance was defined as $p < 0.05$ prior to multiple testing correction with the Holm-Sidak method.

### 3.6 Summary of Results

In all, these results provide a clearer picture of the methylation that is occurring in the CpG islands that span the promoter and exon 1 regions of both IGFBP1 and ZNF264. Both genes demonstrate localised allele-specific differences in methylation between alleles. Both genes show increased methylation further upstream from the transcription start site, and very little methylation within the gene body, a pattern that is consistent with the canonical idea of gene regulation via DNA methylation Siegfried & Simon (2010). Some individuals demonstrate marked differences in methylation between alleles in IGFBP1 (See Figure 3.11). A possible parent-of-origin effect has been identi-
fied in the region immediately upstream of *IGFBP1* (See Figure 3.18B). *IGFBP1* may exhibit both allele-specific methylation and/or Parent-of-Origin effects that are subject to cell type, spatial, or temporal effects, and further investigation is required to elucidate the role that methylation plays in the regulation of *IGFBP1*. Differences are less pronounced in *ZNF264* but are still present in one sample, and at one CpG location, and warrant further investigation with a larger sample of individuals before the hypothesis that this gene exhibits allele-specific methylation is rejected. In addition to a larger number of individuals, efforts could be made to ensure that the number of clones derived from each allele was more balanced. For some individuals, there was an over-representation of one allele in the clones that were sequenced, leading to low statistical power in some analyses (e.g. for the promoter region of *IGFBP1*, 13 clones were found to carry the A allele, while 3 carried the G allele in sample X9023). Allelic biases may have arisen during the PCR amplification of these amplicons, or during transformation, with one allele providing some selective pressure either towards, or against itself. The most straightforward way to assess whether either of these scenarios may be occurring is to generate many more clones and sequence them. With larger sample sizes comes more statistical power, and the opportunity to ask a wider range of questions of the data.
Chapter 4

Computational Techniques and Tools for RRBS Analysis

Reduced Representation Bisulfite Sequencing is a method for extracting relatively large amounts of information on the methylation of CpGs in the genome for a relatively low input of DNA. DNA libraries are enriched for CpG sites, thus reducing the amount of uninformative DNA included in the sequencing. The technique consists of taking DNA from the desired sample, cleaving it with the methylation unspecific enzyme MspI, and performing a size selection upon the cleaved fragments. The end result is that one has enriched for further processing those fragments that represent the areas of the genome where CpG sites are relatively dense. This filtering process means that the selected fragments include 80% of the CpG islands, and around 4 million individual CpG sites, but only around 2.5% of the sequence of the whole genome. So RRBS allows us to analyse the majority of the relevant sites at relatively low cost but high coverage. DNA methylation in mammals primarily occurs at CpG dinucleotides, which are relatively uncommon through large stretches of the genome due to spontaneous deamination and inefficient repair mechanisms (Bird, 2002). The majority of the genome contains less than 1 CpG per 100 bp. Most of these CpGs tend to be methylated, regardless of the cell type. In contrast, parts of the genome contain short stretches
with 10-fold higher levels of CpGs (~1 CpG per 10 bp).

The tools available for analysis of RRBS and other genome-wide methylation techniques are still in their infancy. A number of tools, such as MethylKit, DMAP, BiSeq exist and others are currently being developed (Akalin et al., 2012; Hebestreit et al., 2013; Stockwell et al., 2014). The issue with using tools that are still in development, and are being developed for specific purposes is that often they will not have the full complement of tools that are required for a specific purpose. I came to realise, while attempting to adapt these tools to my purpose, that in many ways the existing bioinformatics tools were inadequate for a project of this scope. As such, this PhD has ended up taking a far more technical approach than the originally intended hunt for biological significance.

This chapter will cover in detail, not only the bioinformatic methods used to process and analyse RRBS data, but also a critical assessment of a number of currently available tools. As such, it includes both step-by-step instructions, as one finds in a methods chapter, and critical discussion of the individual methods. Results are presented separately in Chapter 5. Some of the tools are publicly available, open source programs and scripts, while others were developed in-house to address specific issues for which tools were not available. Techniques and tools have advanced considerably even in just the short time since the analysis portion of this project began, so therefore some of the issues discussed here may no longer be currently applicable. Nevertheless, I will discuss all aspects of the analysis to highlight the challenges faced in developing methods for comprehensively dealing with genome-wide methylation data, specifically data from Reduced Representation Bisulfite Sequencing.

In Figure 4.1, the processing that was applied to raw data is illustrated. While initially appearing complex, it shows the steps required to take raw sequencing data through to analysed data, including options for different tools.
Figure 4.1: Pipeline for analysis of RRBS data. Different colours indicate the tools used. Raw data from the Illumina Hiseq (orange) is passed through FastQC (yellow) to assess quality. Quality trimming and adaptor removal uses the FASTX toolkit and DMAP (blue) before Bismark (purple) is used to map the reads to the genome. MethylKit (green) is an R package that provides one option for extraction of methylation values, Bismark the other. MethylKit provides a reliable method for merging data from both strands of a CpG, and also has its own pipeline for analysis. Elements in red were purpose built in R for this analysis.
4.1 Methods for Data Processing

A great deal of processing is required to transform raw data from a next-generation sequencing run into something that can be interpreted in a meaningful way (biologically speaking). The importance of these steps is often underestimated when embarking upon a new project, but without fully understanding them, accurate and appropriate analyses cannot be performed. From the naive standpoint of a biologist with little experience in bioinformatics, once the libraries have been prepared, the hard work has been done. Once the data have been generated, analysis should be simply a matter of looking at differences between cases and controls (or whatever groupings one's hypothesis dictates). This section explains the various tools and permutations data must undergo before it can be interrogated with respect to a hypothesis. It will also provide the details of the processing that was performed on the 28 placenta RRBS libraries.

4.1.1 FastQC

Before one can begin analysing data from high throughput sequencers such as the Illumina HiSeq, some quality control checks must be performed to ensure the raw data does not have any inherent problems, and is suitable for analysis. FastQC is a useful tool developed at the Babraham Institute (Cambridge, UK) for assessing the quality of libraries prior to trimming, and can also be used subsequently to check that trimming has had the intended effect. FastQC comprises multiple different quality assessments. Most sequencers will generate an internal QC report as part of their run procedure, but this is usually only focused on identifying problems which were related to technical aspects of the flow cell preparation, cluster generation or data collection. FastQC generates a report on numerous aspects of the library and sequence composition which allows identification of problems in either in the sequencing run or in the starting library material. Initially, it provides one with a summary of the library, including how many reads are in the library, the type of encoding in the fastq file (For all of our li-
libraries - Sanger/Illumina 1.9), the length of the sequences, and the average GC%. As an illustration of the information obtained from FastQC, and how it is used to assess library quality for trimming purposes, FastQC data from the library of Sample 173 is presented to give examples of the outpost both pre- and post- trimming.

The quality of reads across the length of each sequence (0-100 bases) is judged based upon phred scores from individual reads. This measure gives one an idea of how each library performs across the length of the sequence. The first graph in a FastQC report shows the range of quality across all reads at each position in the read using box plots (Figure 4.2). From this graph we can determine the mean, median, and range of data at each base position, and also any trends in the quality of the data. Most libraries will show some reduction in quality as one moves towards the end of the read. This plot aids in decisions about whether to trim the raw data. The plot is colour-coded to offer suggestive cut-off values for ‘good’ quality sequence - i.e. Phred 28+ = Good, 20-28 = Proceed With Caution, <20 = poor. For this library, the decision was made to hard trim to 90 bp because of the extremely low proportion of reads within the acceptable range after 90 bp. Hard trimming is defined as cutting all sequences within a given library to a given length - in this case, regardless of the quality of the original 100 bp read, all reads had the final 10 bp removed.
Figure 4.2: Per base quality as determined by FastQC for the sample 173 RRBS library. Prior to trimming and adaptor removal (a), the first 60 bases of every read are of high quality (Phred 28+), but after 80 bp, the mean quality drops down to 2, and the variation in quality increases markedly. After trimming (b), while the mean remains low after 80 bp, the area above the mean shows enough good quality reads to merit retaining the sequence.
“Per sequence quality” tells us where reads are located within the spectrum of quality. A plot displays the distribution of the average quality score of each sequence. In an ideal run, a single peak above 28 would be observed. Runs that contain a subset of poor quality sequence may have one or more additional peaks at lower Phred scores (quality). After quality trimming and adaptor removal, any secondary peaks are expected to be reduced or removed, as is seen in Figure 4.3.

Figure 4.3: “Per Sequence Quality” for the sample 173 RRBS library. Prior to trimming, the distribution of quality is wider, and centered around a phred score of 30. After trimming however, the distribution has narrowed and shifted left, with the mean now sitting at 33.

The “per base sequence” content is usually of more utility in non-bisulphite converted libraries. In a normal, unconverted library, the base position should not influence the base call. However, RRBS libraries are inherently biased (Figure 4.4) towards high T content because they are bisulphite converted. For RRBS libraries, this measure can give an indication of the quality of the conversion reaction for bisulphite-converted libraries. If there is not an overrepresentation of T and G, and an underrepresentation of C, this may indicate poor conversion. The first 3 bases of this plot are always the same for RRBS libraries, as these bases correspond to the MspI cut site (See Figure 1.7).
Figure 4.4: “Per Base Sequence Content” for the sample 173 RRBS library. A moving average of the relative representation of each base across all reads is shown. G (black), T (red), A (green), and C (blue). After trimming the T and A content becomes more stable after 75 bp.

The plot of “per base GC content” again demonstrates the effect of the invariant first 3 base pairs of each sequence, and then steeply drops down to show the average GC content at each position along all reads (Figure 4.5). This roughly corresponds to the average methylation for the whole library. For assessing RRBS libraries, this plot is not usually uninformative, unless GC content is significantly higher than that of the G content seen in the per base sequence content, which may provide supporting evidence for poor bisulphite conversion.
Figure 4.5: “Per Base GC Content” for the sample 173 RRBS library. A moving average of GC content at each point along all reads is shown. The CGG remaining from the MspI cut site can be clearly seen at bases 1-3.

“Per sequence GC content” shows a plot of the average GC content of reads. A plot of the theoretical normal distribution based upon the same mean and standard deviation as the observed data is overlaid upon the actual distribution of the data. In this sample we actually see a bimodal distribution, which is actually exaggerated after trimming (Figure 4.6). In any other library type, this would be an indicator of contamination, but because this library is bisulphite-converted, and reduced representation, the GC structure is less predictable. The bimodal distribution may in fact reflect the fact that roughly 50% of reads will be higher in G content, and 50% will be higher in C content (Figure 4.10, products 1 and 3 have an overrepresentation of G, and products 2 and 4 have an overrepresentation of C). Another factor that may contribute to this pattern, is that CpG sites tend to be either unmethylated, or show moderate to high levels of methylation.
Figure 4.6: “Per Sequence GC Content” for the sample 173 RRBS library. The distribution of sequences with given GC content within the library (red) is plotted along with the theoretical distribution based upon the same mean and standard deviation (blue). After trimming, the bimodal distribution is exaggerated.

“Per base N content” (figure not shown) is another quality measure provided by FastQC. It gives an indication of the quality of the sequencing, and tells us whether there are any uncalled bases within the library and how common they are at each base along the read.

Prior to trimming, the sequence length distribution is irrelevant - it shows us only what we already know, that all of the reads are 100 bp in length. However, after trimming it can tell us how many of our original fragments were shorter than 100 bp, and read into the adaptors. In the library of sample 173, very little of the first 90 bp was adaptor (this library was hard trimmed to 90 bp), although the small peak at about base 65 demonstrates that some of the original fragments were this length, consistent with the original fragment size range of 40-220 bp.
Finally, the FastQC report searches the library for overrepresented sequences. Overrepresented sequences are defined as individual sequences that represent more than 0.1% of the library. These overrepresented sequences are then screened against possible sources of contamination, with the default list of contaminants being all of the the adaptors and primers used in the various sequencing platforms. The list of overrepresented sequences can also provide evidence of contamination from other libraries, or PCR bias. In the library shown here, most of the overrepresented sequences can be associated with TruSeq Adapters (Figure 4.8).

**Figure 4.7:** “Sequence length distribution” for the sample 173 RRBS library. Prior to trimming, all reads are 100 bp in length, after trimming, the majority of sequences are still above 90 bp in length.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCGCGGAGAACGTATCACATGCATATCATATATAAGTTAT</td>
<td>274021</td>
<td>4.443937923271594</td>
<td>TruSeq Adapter, Index 7 (97% over 36bp)</td>
</tr>
<tr>
<td>AGTCCTGAGGACGTCTTAGCTACATCTATATATATAGTAA</td>
<td>63337</td>
<td>1.033632055353657</td>
<td>TruSeq Adapter, Index 7 (97% over 36bp)</td>
</tr>
<tr>
<td>GATCGAGGAGGACGTCTTAGCTACATCTATATATATAGTAA</td>
<td>51850</td>
<td>0.8465880898494504</td>
<td>TruSeq Adapter, Index 12 (97% over 36bp)</td>
</tr>
<tr>
<td>GATCGAGGAGGACGTCTTAGCTACATCTATATATATAGTAA</td>
<td>37294</td>
<td>0.6049021557124728</td>
<td>TruSeq Adapter, Index 11 (100% over 50bp)</td>
</tr>
<tr>
<td>GATCGAGGAGGACGTCTTAGCTACATCTATATATATAGTAA</td>
<td>25592</td>
<td>0.562406109010302</td>
<td>TruSeq Adapter, Index 8 (100% over 50bp)</td>
</tr>
<tr>
<td>AGTTCGGAGGACGTCTTAGCTACATCTATATATATAGTAA</td>
<td>20772</td>
<td>0.3248767026911306</td>
<td>TruSeq Adapter, Index 8 (100% over 50bp)</td>
</tr>
<tr>
<td>GCAGCGGTCGGTGAATCTCTTCTGTTCTTCGAGTAGGAGCTGCC</td>
<td>17116</td>
<td>0.2757526365136584</td>
<td>No Kit</td>
</tr>
<tr>
<td>AGTTCGGAGGACGTCTTAGCTACATCTATATATATAGTAA</td>
<td>12715</td>
<td>0.2092009114561401</td>
<td>TruSeq Adapter, Index 12 (97% over 36bp)</td>
</tr>
<tr>
<td>GCAGCGGTCGGTGAATCTCTTCTGTTCTTCGAGTAGGAGCTGCC</td>
<td>9104</td>
<td>0.15142377223290202</td>
<td>TruSeq Adapter, Index 9 (97% over 35bp)</td>
</tr>
<tr>
<td>GATCGAGGAGGACGTCTTAGCTACATCTATATATATAGTAA</td>
<td>7133</td>
<td>0.1156780573757364</td>
<td>TruSeq Adapter, Index 9 (97% over 35bp)</td>
</tr>
<tr>
<td>GCAGCGGTCGGTGAATCTCTTCTGTTCTTCGAGTAGGAGCTGCC</td>
<td>7122</td>
<td>0.1154984157529885</td>
<td>TruSeq Adapter, Index 9 (97% over 35bp)</td>
</tr>
<tr>
<td>AGTTCGGAGGACGTCTTAGCTACATCTATATATATAGTAA</td>
<td>5636</td>
<td>0.1018965118208236</td>
<td>TruSeq Adapter, Index 9 (97% over 35bp)</td>
</tr>
<tr>
<td>GCAGCGGTCGGTGAATCTCTTCTGTTCTTCGAGTAGGAGCTGCC</td>
<td>6250</td>
<td>0.1014688052637495</td>
<td>TruSeq Adapter, Index 9 (97% over 35bp)</td>
</tr>
</tbody>
</table>

**Figure 4.8:** “Overrepresented sequences” in the library from sample 173. Of the 13 sequences found to be overrepresented, 12 of them are from TruSeq Adaptors.
4.1.2 Fastx Toolkit

The Fastx Toolkit is a collection of command line tools for processing and editing data from high-throughput sequencing (http://hannonlab.cshl.edu/fastx_toolkit/). Designed by the Hannon Lab at Cold Spring Harbour Laboratories for processing reads from genomic DNA, not all of the tools are applicable in the RRBS context, so only those used will be covered here.

4.1.2.1 Quality Trimming

Quality trimming of reads is a necessary part of the processing when libraries are not consistently of high quality. As determined from an examination of the FastQC analysis, the decision to trim was made for several libraries. The length to which libraries are trimmed depends upon the appearance of the data. Some libraries required only a small amount of trimming - to 90 or 95 bp long - while others required harder trimming. For the libraries used in this project, no hard trimming shorter than 75 bp was performed. The reason for this decision was due to the poorer mapping observed in shorter reads, especially considering that at this stage, adaptors are still present in the reads, and once these have been removed, some reads will be shortened. The trade-off to consider when performing trimming is whether it is worth accepting some poorer quality data so that the chance of it mapping is higher. Short reads cause an increased likelihood of multiple mapping, while poor quality sequence may cause incorrect mapping. Where the yield from the library is low (less than 10 million reads), the desire to retain as much of the information as possible is stronger, so trimming on quality was cautiously undertaken in these cases. Where library yield is high, there is less need to be conservative, so trimming could be less parsimonious.

4.1.2.1.1 Dynamic Trimming

An alternative to hard trimming is dynamic trimming of reads according to the Phred score. However, this technique assumes that quality degrades consistently over the
length of the sequence. While this is certainly the case on average, as with most sequencing, the quality of base calling can actually vary from base to base, and quality may go up as well as down. This idea is illustrated in Figure 4.9. Four reads from the same library have been plotted. Three of these reads are less than 100 bp as adapters have been trimmed prior to generating this figure. The yellow and blue reads stay within acceptable quality bounds for their entire length, but the red and green reads are more variable, dipping down before going up again. The green read, under quality trimming conditions would be trimmed to 20 bp (†), and one would lose the information for the rest of the read despite the fact that the majority of the read sits above Q = 20 (equates to a 0.01 error probability). Perhaps of most significance, the rest of the read includes high quality CpG sites such as the two marked (*), both at Q = 32 (0.000631 error probability in read call).

Figure 4.9: Quality Scoring of Individual Reads. Four individual reads are shown, with their quality (Phred) Score at each position along the read. Phred scores were obtained from the CIGAR string contained within the fastq file for this library.
4.1.2.2 Adaptor Trimming

Reads are 100 bp on the type of Illumina run that was used for this study. The fragments that are being sequenced are 40-220 bp in length. Therefore, not all of each fragment will always be sequenced in one read. For fragments less than 100 bp in length, the read will continue into the 3’ adaptor. To accurately map the read to the genome, those adaptor sequences need to be trimmed off. The DMAP toolkit contains the cleanadaptors program which is essentially a program that allows one to remove specified sequences from the fastq files. The most efficient way to do this is to provide the script with a list of all adaptor sequences, not just the adaptor for that specific library. The same file containing all possible adaptor sequences for all libraries can be used to identify and remove this contamination. The fact that many reads contain adaptor sequence means that some reads may end up shorter than 100 bp once adaptors have been cut off. So when trimming adaptor sequences, one can also specify the command `-x 5`. This tells the script to discard any reads that are shorter than 5 bp after adaptors have been removed, as Bismark will be unable to map these sequences. This ‘x’ number can be altered, e.g. to remove any sequences shorter than 40 bp, but the threshold of 5 was used for this study. Therefore, some ‘rogue’ fragments of length shorter than 40 bp may be present, but the “Sequence length distribution” (Figure 4.7) indicates that this is unlikely to be the case.

4.1.3 Mapping to Reference Genome

For an in depth discussion of alignment and mapping tools, please see the thesis of Aniruddha Chatterjee, and the associated publications in NAR and J Biomed Biotechnol (Chatterjee et al., 2012a,b). As much discussion of this topic has already been published, I will not cover it in depth here. Briefly, the short read length (≤100 bp) coupled with the reduced complexity introduced by bisulphite conversion causes mapping of RRBS reads to be more difficult than mapping genomic reads. Because of these two key problems, a comprehensive reference sequence is needed, so the RRBS
technique is currently only applicable in a limited number of organisms. Hopes of de novo assembly must be ruled out, as the genome has been digested and size selected, so huge swathes of the genome are missing, and there is little or no overlap between fragments. In addition to this, we may expect to see poorer mapping in genomes with lower levels of methylation such as the placental genome. A lack of methylated C means that there is a further reduction in genome complexity. With only three bases out of a possible four represented, there is an increase in the possibility of multiple mapping because the number of possible combinations of bases is decreased (e.g. in a DNA sequence of length 70 bp, there are $4^{70}$ possible combinations, but in a bisulphite-converted sequence, there are closer to $3^{70}$, depending on CpG methylation, a difference of $1 \times 10^{142}$). The placenta is a ‘globally hypomethylated’ tissue compared to other tissues. Pre-eclamptic placentas appear to show a still lower level of methylation (Figure 5.1). Because of this further reduction, it may follow that mapping will also be reduced in placenta libraries as a whole, and Pre-eclamptic libraries in particular. Evidence for this hypothesis is limited at present. One publication by Lim et al. (2012) reported a comparison of mapping efficiencies from different aligners at different levels of methylation, and did not find any variation in the mapping efficiency with differing methylation levels.

It is not enough to map back to a reference genome, but one must map to a completely bisulphite converted reference genome to counter the reduced complexity bisulphite conversion introduces. Instead of working with infinitely variable combinations of three letters (A,T,C,G), we are now working with three letters (A,T,G), with an occasional fourth (meC). When one adds to this fact the fact that there are multiple possible PCR products from bisulphite converted DNA (Figure 4.10), it becomes apparent that we need multiple reference genomes against which to map. Because we cannot predict which strand our sequencing products will come from, we must map to a C->T genome, and a G->A genome for both the top and bottom strands of the reference genome.

Repetitive regions do not map well in a genomic context, let alone a bisulphite
**Figure 4.10**: PCR Products from a Bisulphite Conversion Reaction. After denaturation and bisulphite conversion, it is possible to get 4 different PCR products from a single double-stranded piece of DNA. The original top strand gives rise to a bisulphite converted top strand, and a strand complementary to the bisulphite converted top strand, and the original bottom strand does the same.
converted context. This is unfortunate, as repetitive regions of the genome are some of the most interesting epigenetically and evolutionarily speaking. Human placentas are biologically distinct from the placentas of other mammals, and there is increasing evidence to indicate that repetitive elements play a key role in this distinction.

The Bismark bisulphite mapper (available as a download from Babraham Bioinformatics) was the program used to map the reads back to the reference genome GRCh37/hg19, which was obtained via ftp from ensembl.

### 4.1.3.1 Mapping of Short Reads

Bowtie is the program most used to map short fragments of DNA, and this is the aligner that is used by Bismark (other aligners are available). It attempts to map each fragment against all parts of the genome, and determines where the read is most likely to have come from. Chatterjee et. al (Chatterjee et al., 2012b) attempted to map against an *in silico* RRBS reference genome. This was generated by determining where in the reference genome MspI would cut, and size selecting the theoretical fragments (40-220 bp). However, they found that, rather than increasing the mapping efficiency, it actually caused a slight decrease in efficiency of mapping, with some reads no longer mapping at all. Part of the reason for this may be due to inter-individual genomic variation. Differences in the sequence between people can cause changes not only to methylation sites but to cut sites, so the RRBS libraries from two individuals (regardless of how identically the libraries are prepared) will not consist of exactly the same fragments.

#### 4.1.3.1.1 SAM files

The Bismark aligner was used to produce a SAM (Short Alignment Map) file for each library. These files contain information about genomic location, quality of base-calls, the sequence of the read, and conversion of Cs, along with various other pieces of information relevant to the sequencing platform. The SAM file format is the text equivalent of a BAM file, which stores the same information, although in a binary compressed form. SAM files are larger than BAM files, and often exceed 4 GB in size. This can
mean that manipulation of the files can be difficult on a basic desktop computer. For a bioinformatics beginner, and pipelines that are still in development, the information is more accessible when the file can be opened and the data in it actually looked at. A single line of a bismark SAM file equates to a single uniquely aligned read, and contains the following 16 pieces of information (Krueger & Andrews, 2011):

1. HWI-ST871
2. 16
3. 5
4. 77590844
5. 255
6. 91M
7. *
8. 0
9. 0
10. AAATACCATTTAACACACAAATCACTTTCTCTAAACCGCATCAACAATCTCACTTTCTTCTTACTATTTGCTCTCCTCAATCG
11. BFBBBFBBBFBBBFBBFFIFIFFFIFIFFBIIIIFFIFIIFIFIFIFIIIIIFIIIIIF IIIF المسيح FFBBB
12. NM:i:10
13. XX:Z:4G7G3G3G23G3G14G3G4GG17
14. XM:Z:..................Z..................Z..................Z
15. XR:Z:CT
16. XG:Z:GA

Each piece of information is as follows - 1. The name of the read, 2. a SAM flag, 3. the chromosome, 4. the base position of the left end of the read, 5. mapping quality (>10 is unique), 6. CIGAR string which describes insertions, deletions, and matches 7. and 8. are of relevance only for mate-pair libraries, and 9. the template length. Of major interest are items 10. the sequence read, 11. the quality information which for the bismark aligner is encoded in Sanger Phred 33 format, and 14. the CpG, CHG, and CHH information. Of most relevance is item 14., in which an uppercase Z indicates a read of a methylated C in a CpG context, and a lowercase z indicates an unmethylated C in a CpG context. Items 12., 13., 15., and 16. all contain mapping program-specific flags.
4.1.3.2 Global Methylation and Conversion

The Bismark alignment program also produces a brief mapping report for each library. This report details the mapping efficiency, including how many reads mapped uniquely. Also included in this report is a summary of the total number of methylated cytosines encountered during the mapping process. The number of methylated cytosines found in a CpG context give a rough estimate of the global methylation for the library. Of more importance at this stage of analysis, are the number of cytosines present in a CHG or CHH context. While a small amount of cytosines are expected to be present in these contexts, if the percentage is high, this is an indication of poor bisulphite conversion. For this study, libraries were excluded from further analysis if the amount of ‘methylation’ in either of these two contexts was greater than 5%.

4.1.4 Extraction of Methylation Values

Two main methylation extractors were used for this project. Others are available, but the Bismark methylation extractor and methylKit methylation extractor will be discussed here. Both of these tools scan the reads stored in .sam files looking for Z/z symbols within reads and extracts them along with their location information, making it easy to identify the genomic context of each CpG.

4.1.4.1 Bismark Methylation Extractor

The Bismark methylation extractor, draws information from all CpG sites within a library. Because one has already performed quality trimming, it is assumed that those reads that remain are all of an acceptable quality. This methylation extractor tool scans the SAM file for any CpG sites, and extracts the methylation and location information to a new tab-delimited text file. The lack of filtering based upon quality is the significant difference between this tool, and the MethylKit methylation extractor.
4.1.4.2 MethylKit Methylation Extractor

MethylKit methylation extractor provides an alternative to the idea of dynamic trimming discussed earlier. The MethylKit methylation extractor does not assume that all CpGs are equally valid. It incorporates a quality threshold requirement of Phred \( \geq 20 \) (or otherwise specified). Filtering at this stage means that only the CpG sites are subject to quality requirements. The rest of the read, which would not be analysed regardless of quality, is discarded. The methylKit methylation extractor is, as such, more stringent than the Bismark methylation extractor. As quality trimming has already taken place, one may wish to remove the quality requirement at this stage. For the purposes of this project, a threshold of Phred \( \geq 13 \) was used. This removed all CpG reads with lower than 0.95 confidence in calling, (i.e., an error probability of \( >0.05 \)).

4.2 Statistical Testing of Individual CpGs

After examining the tools available for the significance testing of RRBS data (MethylKit, BiSeq, DMAP), it was found that none of them were appropriate for the experiment, and the decision was made to design a new analysis. While some tools do exist for the analysis of RRBS data, they have for the most part arisen as a result of a need to perform a specific analysis on a specific dataset. In our laboratory, previous work had been done to develop the DMAP software package. DMAP arose as part of a project that was looking for areas of high variation with respect to DNA methylation, both within, and between individuals. In contrast, this project was aimed at identifying regions of low variability within two groups of individuals, but high variation between the groups. As such, at the time of this analysis, DMAP was still under development, and was unable to perform such a test with the appropriate statistical analysis. DMAP uses as its unit of analysis the MspI fragments generated by enzyme digestion during library preparation. This approach has some benefits, as by analysing a fragment instead of an individual site, one reduces the total amount of data and testing required, thus lowering
the false discovery rate. Additionally, it enables the generation of a ‘smoothed’ picture of methylation across the genome by essentially employing a variation of the window-based approach that has also been investigated for use in RRBS analysis (Hansen et al., 2012). However, as the goal of this project was to identify, not regions of high variation, but genomic locations that display differential methylation in Pre-eclampsia, the decision was made to progress using a single-CpG approach. A large contributor to this decision was the downstream use of Sequenom to validate any differences in methylation that were discovered. The Sequenom platform quantifies DNA methylation in amplicons of bisulphite converted DNA at single CpG resolution. Sequenom is a well validated method for quantifying DNA methylation and is already used for clinical tests of cancer risks and subtypes, and pre-natal testing (Gao et al., 2012; Shinojima et al., 2010; Vanaja et al., 2009). The end goal of this project is to develop a diagnostic test that can be used during pregnancy to detect risk associated with abnormal methylation. Therefore, an analysis method was needed for the RRBS data that would produce results that could be directly compared to methylation values obtained via other methods. While it may be possible in some cases to construct amplicons for Sequenom that do exactly match a specific MspI fragment, this is by no means guaranteed. Sequenom amplicons are commonly restricted by where primers can bind, and adding additional restrictions of where MspI cuts would further restrict where amplicons can be validated. The analysis approach that arose out of this reasoning would be performed in parallel with one of the existing tools (MethylKit) as a comparison. We don’t yet fully understand the complexities of RRBS data. The correlations within and between CpG sites, fragments and individuals are not simple. Without the aid of an advanced degree in biostatistics, I did the best I could with the tools of my disposal to design an appropriate analysis, and consulted with a statistician for some aspects regarding the modelling of data. The approach that Dr Matthew Parry and I decided upon combines a conservative approach to data with a high stringency criteria when it came to accepting a result as true. Single CpG sites would be analysed independently, but in such a way that if correlation structures were influencing the data, that non-dependence would be
demonstrated. Part of the approach is illustrated in the lower portions of Figure 4.1. However, further refinement was required to that approach to arrive at a list of 'top hits'. Figure 4.11 illustrates some of the possible filters that can be placed upon a basic analysis. Because, as already stated, there is not a standardised method for analysing RRBS data (Robinson et al., 2014), I chose to select quite stringent criteria for inclusion in a list of candidates for further validation, or “top hits”. Of all of the sites that were covered by any of the libraries, 393611 sites showed 8-fold coverage in two-thirds of the samples in each group. These sites were then analysed for statistical differences in methylation between the case group and the control group.

4.2.1 Regression and Modelling of Data

Regression demonstrates whether there is a relationship between the average values of two variables. The aim of regression is to help predict the average value of one variable (dependent - e.g. disease status, pre-eclampsia) for a given value of another variable (independent - e.g. methylation). It finds the equation around which the data points have least scatter. We can use statistical modelling to determine the best equation to fit the data. At many sites, a simple binomial distribution is observed, and the predictive value of of methylation can be relatively easily tested. Modelling in this way is generally about expressing the mean response in terms of the explanatory variables (plus random effects). Mixed effects models are a type of regression model that take into account variation that is not generalisable to the independent variable - i.e. there is a difference in disease state that is not due to methylation? Some effects may be fixed, such as maternal age or gestation, while others may be more random, such as genetic variation between individuals, and genome-epigenome interactions deriving from differences in the genome.
Figure 4.11: Statistical decisions for calculating significant differences. The final list of most significantly different sites was generated by finding the overlap between the MethylKit-generated results (in green), and the two custom-made analyses (in red). A combination of means calculations and significance values were used.
4.2.2 Logistic Regression

Logistic Regression is used to determine the best equation to describe continuous data linked to a binary variable such as disease status. When continuous data such as methylation data (continuous values between 0 and 1) do not fit to a standard binomial distribution, the equation we are using must be altered to better fit the real distribution of the data. A logistic regression test takes the specified information from each sample (coverage, number of Cs, number of Ts), and tests to compare the fraction of methylated Cs across the test and control groups. At a given base it examines the relationship between methylation level and PET outcome, and allows us to control for confounders (such as gestation, maternal age, smoking etc.). The methylKit program models the methylation proportion $P_i$, for samples $i=1,2,...,30$ (i.e. 30 iterations = 30 samples) using the equation:

$$\log\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 * T_i$$  \hspace{1cm} (4.1)

where $P_i$ is the methylation proportion, $T_i$ is the treatment group (0 for control, 1 for case, this is our binary dependent variable), $\beta_0$ is the log odds of the control group, and $\beta_1$ is the log-odds ratio between the case and control groups. The null hypothesis is that $\beta_1 = 0$. When this hypothesis is rejected, it implies that the methylation proportions (and log-odds) are different between the cases and controls. It also allows for the integration of additional covariates into the analysis, such as the potential confounders previously mentioned. Linear regression has both fixed effects and random effects: we have a response/outcome ($y$) which we think depends on ($X$) but there are other factors we haven’t explicitly accounted for that we incorporate into the error term.

$$\text{i.e. } y = \beta X + \epsilon$$

The major advantage of using logistic regression to model the data before testing is that it allows for the integration of read count as well as methylation percent, i.e., the methylation value is analysed as a proportion rather than a percentage. If these data
were analysed in a t-test, there would be significant biases due to the difficulty of integrating the read count into the analysis.

### 4.2.2.1 Quasibinomial Model

Primarily, all sites that met the criteria for inclusion were run through a simple test for binomial distribution. Some sites did exhibit normally distributed data around a central mean, but this was not the case for all sites. For those samples that did fit to a binomial distribution, a hypothesis test was able to be performed upon the analysis of deviance table. The null hypothesis is that there is no effect of case status. If the null hypothesis is true, then the change in deviance should be chi-squared. Those sites that showed residual deviance higher than expected as detected by an F-test were channeled instead through the Quasibinomial model. The Quasibinomial model is a variation on a binomial model that introduces a dispersion parameter ($\phi$) to account for additional variation within each site (between individuals). In essence this dispersion parameter changes the shape of the distribution curve while keeping the mean and standard deviation the same.

### 4.2.2.2 Random Effects Model

The Random Effects model is similar in principle to the Quasibinomial, but instead of introducing a variable to account for the error, it introduces a variable that we can ascribe some of the over dispersion to. Using a model like this acknowledges that there may be some unknown source of variation that we are not controlling for - this is very possible in clinical samples, which often have incomplete phenotype data. A random effects model has several benefits, in that, as more phenotype data becomes available, we can integrate these into the model as fixed effects. In summary, fixed effects are the explanatory variables we are interested in; random effects are explanatory variables that we want to account for but are not the focus of our investigation. In practice, there will be a parameter ($\beta$) associated with each fixed effect ($X$) that quantifies its contri-
bution - typically $\beta \times X$. A random effect, on the other hand, is a random contribution - in the simplest case, a normal random variate with standard deviation $\alpha$. Inference involves estimating both $\beta$ and $\alpha$.

4.2.3 MethylKit

MethylKit is an R package that has been designed to analyse and annotate data from RRBS. One can use it to generate summary statistics and produce various graphical outputs. It also provides base code for performing differential methylation analysis.

4.2.3.1 Adaptations to MethylKit

Several alterations were made to the MethylKit analysis pipeline. After reading in the sorted Bismark SAM files, a function is available to filter the samples based on read coverage; i.e., remove CpG sites with particularly high coverage that may be suffering due to PCR bias, and remove sites with coverage lower than 8 reads. Following this, one is able to merge reads from both strands of a CpG dinucleotide. However, the flaw in this order of operations is that one would end up discarding a site that had, e.g. 17 reads (9 on one strand and 8 on the other) - together enough to reach the threshold of 10 reads, but separately neither the forward nor the reverse strands meet the threshold criteria. Similarly, if one has 20 reads on one strand, and 7 on the other, only the 20 of those reads that are on the strand that meets the filtering requirements would be analysed. To avoid this unnecessary discarding of potentially informative sites, lower limit filtering based on read count was not performed until after merging of the CpG dinucleotide using the unite function. Another useful tool that can be implemented during the unite function is the ability to specify a minimum number of samples per group that must be covered at each CpG for it to be retained for further analysis. For my samples, as 5 of the libraries had low coverage across the genome, I opted to set a further requirement saying that at least $\frac{2}{3}$ samples per group (cases/controls) had to meet threshold read requirements for the CpG site to be analysed. Because no
filtering based on minimum read count has been applied prior to the unite function, an additional piece of code needed to be written to allow us to perform this minimum samples per group filtering after the merging of data from the two strands. The final read depth requirement for both strands was set to 8. A read depth of 4 has been found to be enough to find agreement between RRBS libraries and the Infinium 450K chip approach, although 10-fold read depth is a more commonly used (Lee et al., 2014; Pan et al., 2012). As already stated in Section 4.2, 393611 CpG sites met the stringency criteria for coverage in sufficient samples, when this stringency was increased to 10-fold read depth, the number of sites that met the threshold was reduced by 30% to 276238 CpG sites. In the interests of retaining a large number of sites to analyse, but also a high threshold for coverage per sample, the threshold for read depth was set to 8 reads per CpG site.

4.3 Identifying Candidate Differentially Methylated Sites

4.3.1 Calculation of Methylation Differences

Calculating the difference in methylation between two groups of complex data is not straightforward. Two methods were utilised in this project. MethylKit, as seen in Figure 4.1, uses a weighted method of calculating the difference between cases and controls. For each sample, there is a count of methylated Cs and a count of unmethylated Ts, which together amount to a measure of total coverage. This coverage can vary remarkably between libraries (and samples), and the weighting method attempts to correct this by giving more importance to samples with higher coverage, and less importance to those with lower coverage. The equation used amounts to the following:
\[
\left( \frac{Cs_1 + Cs_2 + Cs_3 + \ldots + Cs_n}{Cov_1 + Cov_2 + Cov_3 + \ldots + Cov_n} \right)_{\text{Cases}} - \left( \frac{Cs_1 + Cs_2 + Cs_3 + \ldots + Cs_n}{Cov_1 + Cov_2 + Cov_3 + \ldots + Cov_n} \right)_{\text{Controls}}
\] (4.2)

Where the sum of the number of Cs in cases is divided by the sum of the Coverage in cases, and the same equation is performed for all the samples in the Control Group. The weighted methylation value for the Controls is then subtracted from the weighted methylation value for the Cases to find the Methylation Difference (Equation 4.2). Because MethylKit uses this weighted method by default, the same method was used in the two other analyses to enable comparisons between results. However, this was not the only method used to calculate the difference in methylation.

Alternatively, an unweighted, or arithmetic mean was used. This consists of calculating a methylation fraction for each sample (number of Cs divided by the coverage), and taking the average of this for each group (Equation 4.3). The average methylation of Controls is then subtracted from the average methylation of Cases to find the methylation difference. The approach is summarised in the following equation:

\[
\left( \frac{Cs_1}{Cov_1} + \frac{Cs_2}{Cov_2} + \frac{Cs_3}{Cov_3} + \ldots + \frac{Cs_n}{Cov_n} \right)_{\text{n (Cases)}} - \left( \frac{Cs_1}{Cov_1} + \frac{Cs_2}{Cov_2} + \frac{Cs_3}{Cov_3} + \ldots + \frac{Cs_n}{Cov_n} \right)_{\text{n (Controls)}}
\] (4.3)

### 4.3.2 Annotation of Differences

The differences found were annotated using HOMER (Heinz et al., 2010), a set of tools developed at the Salk Institute for identifying motifs and analysing next-gen sequencing. The annotatePeaks program is a multi-purpose tool for annotating genomic locations when given BED, or BED-like files. The output from the analyses performed here were similar enough to the BED files required for HOMER that only very minor cosmetic changes were required. Output files are tab-delimited, and can be opened in
any text-editing software. Basic annotation consists of the following columns in addition to the input columns:

1. Peak Score
2. FDR/Peak Focus Ratio/Region Size
3. Annotation (i.e. Exon, Intron, ...)
4. Detailed Annotation (Exon, Intron etc. + CpG Islands, repeats, etc.)
5. Distance to nearest RefSeq TSS
6. Nearest TSS: Native ID of annotation file
7. Nearest TSS: Entrez Gene ID
8. Nearest TSS: Unigene ID
9. Nearest TSS: RefSeq ID
10. Nearest TSS: Ensembl ID
11. Nearest TSS: Gene Symbol
12. Nearest TSS: Gene Aliases
13. Nearest TSS: Gene description

### 4.4 Summary of Analyses

Because there is not yet a recognised appropriate pipeline for the analysis of RRBS data, this project utilised some of the existing tools, and created some new ones for the purpose. Many judgement calls must be taken when first embarking on a new method, but attempts were made to explore multiple avenues to ensure the best end result. The use of multiple tools, while experimental, also allows for a high threshold for acceptance of results when those results are confirmed via multiple methods. Those results from this analysis that will be investigated further will be those that were deemed significantly different by a variety of methods. The choice of analysis method is an important one, and thus far no one method has been definitively proven to be superior to others. One key aspect that is yet to be resolved by any method relates to the independence of sites. The current tools (DMAP, methylKit, BiSeq) all assume some level
of independence between the CpG sites within a single read. As there are likely to be multiple CpG observations per read, it is probable that CpGs in the same read exhibit some degree of correlation, given that the CpGs in a single read must come from the same piece of DNA in the same genomic region of the same cell. Therefore, the methylation status of these sites is likely to be correlated, and to reflect the biological activity of that region of the genome in that cell. Attempts are currently being made in this lab to address this issue of CpG site independence using Generalised Estimating Equations which are useful in analyzing data where correlated measurements exist, such as multiple measurements of methylation from a single cell (Foster et. al, unpublished). These methods are still under development however, so this project has adopted an approach that does assume some degree of independence between sites, but if adjacent sites are correlated both with each other, and with the disease state, they should exhibit similar p-values, as is seen in analyses for GWAS, where the spatial correlation of sites is due to the frequency of recombination events (Ehret, 2010).
Chapter 5

Results of RRBS Analysis

The results of the analysis outlined in the previous chapter are varied and complex. Because a variety of analysis methods were trialled and combined in different ways, it is necessary for future work in this project to have a clear list of ‘top hits’ - candidates to investigate further and to provide direction to the research. This project has been a pathfinding expedition into RRBS analysis for epigenetics in the placenta. This chapter will show where the paths discussed in Chapter 4 have led us in our search for biomarkers and mechanisms for the development of pre-eclampsia. Results will be shown by following the data through the three different analysis methods (MethylKit, Quasibinomial, and Random Effects), and how the results of these three analyses have been filtered and prioritised to produce a series of lists, stratified by the stringency of the inclusion criteria.

5.1 Library Composition

The data analysed is the result of the construction and sequencing of RRBS libraries. The size and composition of these libraries is a very important indicator of how well the experiment has performed from a technical perspective. Generally, the more reads obtained, the better. All libraries sequenced contained more than 5 million reads, with
the exception of samples 134, 195, and 520, for which 3.3, 4.7, and 4.8 million reads were obtained respectively. The majority of the libraries mapped well - with over 50% of reads mapping uniquely. Again, a few libraries did not perform well in this respect, but due to our small sample size, libraries were not excluded from analysis based solely upon a paucity of reads. The % CpG content gives a rough estimate of the global methylation for each library - it must be noted though that this is merely a measure of the total proportion of methylated cytosines within the sequenced library, and is not adjusted for depth or breadth of library coverage. The percentage of methylation observed in a CHG and CHH context enables us to estimate the efficiency of the bisulphite conversion reaction. Samples with greater than 5% C in either of these measures were excluded from further analysis due to poor conversion. Therefore, from the initial 28 libraries, the libraries of samples 134 and 195 were excluded from further analysis due to high levels of non-CpG methylation, and the libraries of the remaining 26 samples were taken through to data analysis (Table 5.1).
Table 5.1: Summary of library content showing amount of reads obtained, whether trimming took place, how well libraries mapped, and how much methylation was observed for each library. For libraries 47 and 416, the same library was sequenced on two separate Illumina runs. The data from the second run is included in brackets, and the reads combined after mapping.

<table>
<thead>
<tr>
<th>OPuS ID</th>
<th># Reads</th>
<th>Trimming</th>
<th># Reads after trim</th>
<th>% Mapped</th>
<th># Mapped Reads</th>
<th>% CpG</th>
<th>% CHG</th>
<th>% CHH</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>7513824</td>
<td>80 bp</td>
<td>7510962</td>
<td>46.3%</td>
<td>3479654</td>
<td>35.6%</td>
<td>2.1%</td>
<td>2.2%</td>
</tr>
<tr>
<td>46</td>
<td>13705085</td>
<td>none</td>
<td>13705085</td>
<td>66.8%</td>
<td>9159275</td>
<td>45.3%</td>
<td>2.2%</td>
<td>1.9%</td>
</tr>
<tr>
<td>47</td>
<td>5067916 (13493947)</td>
<td>none</td>
<td>5067916 (13493870)</td>
<td>78.7% (76.7%)</td>
<td>14336363</td>
<td>45.3%</td>
<td>2.2%</td>
<td>1.9%</td>
</tr>
<tr>
<td>47</td>
<td>5067916 (13493947)</td>
<td>none</td>
<td>5067916 (13493870)</td>
<td>78.7% (76.7%)</td>
<td>14336363</td>
<td>45.3%</td>
<td>2.2%</td>
<td>1.9%</td>
</tr>
<tr>
<td>68</td>
<td>5359159</td>
<td>95 bp</td>
<td>5359159</td>
<td>60.5%</td>
<td>3244454</td>
<td>31.9%</td>
<td>1.2%</td>
<td>1.1%</td>
</tr>
<tr>
<td>70</td>
<td>10525904</td>
<td>80 bp</td>
<td>10525904</td>
<td>44.2%</td>
<td>4656107</td>
<td>21.3%</td>
<td>2.0%</td>
<td>2.2%</td>
</tr>
<tr>
<td>105</td>
<td>15061373</td>
<td>95 bp</td>
<td>15061373</td>
<td>46.2%</td>
<td>6954104</td>
<td>25.0%</td>
<td>1.6%</td>
<td>1.6%</td>
</tr>
<tr>
<td>110</td>
<td>17750809</td>
<td>90 bp</td>
<td>17750809</td>
<td>69.2%</td>
<td>9159275</td>
<td>34.4%</td>
<td>1.5%</td>
<td>1.5%</td>
</tr>
<tr>
<td>121</td>
<td>12575751</td>
<td>none</td>
<td>12575751</td>
<td>64.1%</td>
<td>8057386</td>
<td>32.5%</td>
<td>1.0%</td>
<td>0.9%</td>
</tr>
<tr>
<td>122</td>
<td>22877910</td>
<td>none</td>
<td>22877910</td>
<td>60.2%</td>
<td>13777234</td>
<td>30.8%</td>
<td>0.7%</td>
<td>0.7%</td>
</tr>
<tr>
<td>134</td>
<td>3313301</td>
<td>80 bp</td>
<td>3313301</td>
<td>45.1%</td>
<td>1493987</td>
<td>38.0%</td>
<td>11.3%</td>
<td>11.5%</td>
</tr>
<tr>
<td>140</td>
<td>22800929</td>
<td>none</td>
<td>22800929</td>
<td>74.3%</td>
<td>16935049</td>
<td>40.0%</td>
<td>0.6%</td>
<td>0.6%</td>
</tr>
<tr>
<td>159</td>
<td>8117270</td>
<td>80 bp</td>
<td>8117270</td>
<td>35.2%</td>
<td>2857378</td>
<td>45.8%</td>
<td>2.5%</td>
<td>2.2%</td>
</tr>
<tr>
<td>170</td>
<td>9180003</td>
<td>95 bp</td>
<td>9180003</td>
<td>62.8%</td>
<td>5768113</td>
<td>40.6%</td>
<td>3.0%</td>
<td>3.0%</td>
</tr>
<tr>
<td>173</td>
<td>6166347</td>
<td>90 bp</td>
<td>6166347</td>
<td>66.1%</td>
<td>4077250</td>
<td>37.8%</td>
<td>2.0%</td>
<td>2.1%</td>
</tr>
<tr>
<td>195</td>
<td>4838915</td>
<td>80 bp</td>
<td>4838915</td>
<td>32.7%</td>
<td>1583020</td>
<td>44.0%</td>
<td>6.9%</td>
<td>6.8%</td>
</tr>
<tr>
<td>199</td>
<td>10173573</td>
<td>none</td>
<td>10173573</td>
<td>64.3%</td>
<td>6543873</td>
<td>30.6%</td>
<td>1.0%</td>
<td>0.9%</td>
</tr>
<tr>
<td>233</td>
<td>7301675</td>
<td>90 bp</td>
<td>7301675</td>
<td>50.3%</td>
<td>3674986</td>
<td>35.0%</td>
<td>1.4%</td>
<td>1.4%</td>
</tr>
<tr>
<td>234</td>
<td>21621581</td>
<td>none</td>
<td>21621581</td>
<td>66.0%</td>
<td>14255778</td>
<td>32.2%</td>
<td>1.1%</td>
<td>1.1%</td>
</tr>
<tr>
<td>250</td>
<td>8236267</td>
<td>none</td>
<td>8236267</td>
<td>77.2%</td>
<td>6355661</td>
<td>35.7%</td>
<td>1.1%</td>
<td>0.8%</td>
</tr>
<tr>
<td>364</td>
<td>30664087</td>
<td>none</td>
<td>30664087</td>
<td>62.4%</td>
<td>19137086</td>
<td>35.0%</td>
<td>0.9%</td>
<td>0.8%</td>
</tr>
<tr>
<td>410</td>
<td>9723001</td>
<td>95 bp</td>
<td>9723001</td>
<td>61.5%</td>
<td>5976174</td>
<td>32.3%</td>
<td>1.3%</td>
<td>1.2%</td>
</tr>
<tr>
<td>416</td>
<td>5441108 (14189543)</td>
<td>80 bp (75 bp)</td>
<td>5436233 (14180260)</td>
<td>29% (28.8%)</td>
<td>5658302</td>
<td>38.0%</td>
<td>2.0%</td>
<td>1.6%</td>
</tr>
<tr>
<td>423</td>
<td>5719098</td>
<td>none</td>
<td>5719098</td>
<td>65.5%</td>
<td>3744942</td>
<td>33.2%</td>
<td>0.9%</td>
<td>0.9%</td>
</tr>
<tr>
<td>424</td>
<td>1183855</td>
<td>none</td>
<td>1183855</td>
<td>70.1%</td>
<td>8296513</td>
<td>40.3%</td>
<td>2.0%</td>
<td>1.8%</td>
</tr>
<tr>
<td>478</td>
<td>10065339</td>
<td>95 bp</td>
<td>10065339</td>
<td>60.9%</td>
<td>6133745</td>
<td>41.2%</td>
<td>1.6%</td>
<td>1.9%</td>
</tr>
<tr>
<td>520</td>
<td>4748158</td>
<td>85 bp</td>
<td>4748158</td>
<td>23.2%</td>
<td>1101801</td>
<td>41.8%</td>
<td>4.1%</td>
<td>3.7%</td>
</tr>
<tr>
<td>526</td>
<td>12684848</td>
<td>none</td>
<td>12684848</td>
<td>73.0%</td>
<td>9255871</td>
<td>37.8%</td>
<td>0.5%</td>
<td>0.6%</td>
</tr>
<tr>
<td>556</td>
<td>15542965</td>
<td>none</td>
<td>15542965</td>
<td>64.0%</td>
<td>9953490</td>
<td>35.8%</td>
<td>1.1%</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

Some libraries yielded more reads than others. It was in part because of this that the filtering criteria for a CpG site to be analysed included the requirement that it be present in two-thirds of samples. For example, while Sample #416 yielded a total of 19 million plus reads, it mapped at a low rate (29%), so the number of analysable reads was just 5,658,302. By not requiring that all libraries have the same complement of CpGs, it enables samples that have low coverage to contribute data where they are able, but does not preclude analysis if those samples are not represented. This also allows for libraries to have different compositions. Further to the observation that some
samples mapped poorly, it could be argued that these samples should be excluded on those grounds. However, this low mapping efficiency may not be a reflection of quality, but instead of composition. All libraries have slightly different compositions due to experimental and genetic variation. Mapping algorithms rely on identifying unique matches in the genome (Krueger & Andrews, 2011). This means that, if a library is particularly enriched for, e.g., repetitive regions, this can drastically reduce mapping efficiency (Doherty & Couldrey, 2014). In the libraries generated for Sample #416, there were a further 1.8 million reads that mapped to the genome, but not uniquely, bringing the actual mapping efficiency up to 38%. No significant differences in either raw read number, number of mapped reads, or mapping efficiency were observed between cases and controls (Figure C.1).

Filtering criteria were applied to the methylation data from the 26 libraries. These criteria are outlined in Chapter 4.2, but to briefly reiterate: for a CpG site to be included in a library, it had to display a Phred score of equal to or greater than 13 (at an error probability of <0.05). The total number of sites covered in any depth by any library was 9,779,215 CpG sites. Sites were then included in the final analysis if they displayed a coverage of at least 8 reads in two-thirds of all samples in each group (cases and controls). This filtering reduced the number of eligible sites to 393,611. The remaining results are based upon this relatively small number of sites.

When the global CpG methylation (as calculated by the bismark mapping tool) for each library is examined, a trend of lower methylation in cases of pre-eclampsia is observed (Figure 5.1). When subsequently a t-test was performed on all of the CpG sites covered to acceptable levels in sufficient libraries, comparing the mean methylation of all cases with all controls, a highly significant result was found. A paired t-test, with pairing based upon CpG site; i.e., mean methylation of cases at site 1 paired with mean methylation of controls at site 1 revealed that the CpGs from pre-eclamptic libraries are 1.98% lower than controls (95% CI 1.97-2.00, p<2.2e-16). Of the 393,611 sites tested, 59% display hypomethylation in pre-eclampsia compared to controls, compared to 37% that display hypermethylation in pre-eclampsia. The remaining 4% of sites show
Figure 5.1: Methylation of placentas in RRBS libraries. Cases are shown in red, and Controls in blue, black lines connect matched pairs with each other. Students matched t-test shows a significance level of \( p=0.028 \).

no differences between cases and controls.

5.2 Testing for Statistically Significant CpG Sites

The first step towards identifying candidate sites of interest is to identify those sites that reach a pre-ordained threshold of significance. While the commonly used threshold for any individual test is a \( p \)-value of 0.05, when multiple tests are performed, adjustments must be made to account for the number of tests. Two of the commonly used methods of correcting for multiple testing are the Bonferroni method, and the Benjamini-Hochberg method. The Bonferroni method simply divides the pre-ordained \( p \)-value threshold (0.05) by the number of tests performed (393,611). The Benjamini-Hochberg method is slightly less stringent, and instead of simply taking those sites that fall under a fixed threshold of significance, takes sites that fall under a threshold with a slope of 0.05 the number of tests. In the plots that follow, both thresholds are
illustrated (Figures 5.2-5.4). Due to the use of three methods for analysing this data, three different sets of significant sites have been identified. There is, however, some overlap between these sets, as described in Section 5.5.

5.2.1 MethylKit

Because MethylKit has its own inbuilt method of performing multiple correction (SLIM (Wang et al., 2011)), it is the q-value derived from this method that has been used as the measure of significance, and sites falling under the significance threshold of q<0.01 are deemed significant. Of the 393,611 CpG sites that met the filtering criteria for analysis, the majority show no statistical evidence of difference. 6,537 sites do show statistically significant differences between cases and controls (Figure 5.2).
Figure 5.2: Significance values for CpG sites from MethylKit analysis. Sites were sorted by significance level, and then plotted. The upper plot shows all sites, while the lower plot shows the contents of the small region highlighted by the blue box and arrow in the upper plot. Significant CpG sites fall under the red line.

5.2.2 Quasibinomial

The quasibinomial model feeds data from each CpG site through the binomial model. If the data from a CpG site doesn’t fulfil the distribution criteria of the binomial model, then that site is channelled instead through the quasibinomial model. If the model
variance could be estimated, as with the binomial distribution, a chi-squared test was used to test the null hypothesis that disease status does not affect methylation status. If the residual deviance is higher than the binomial model expects, the quasibinomial model can be used to introduce a dispersion parameter to explain the deviation of the data from a binomial distribution. Following this modelling, the F-test was used to test the null hypothesis. This method is relatively inflexible, and thus provides us with the lowest number of sites that are significantly different. With only 69 of the 393,611 sites interrogated returning p-values that were significant, and only 8 of these at Bonferroni significance, this method seems the most stringent (Figure 5.3). Of the 393,611 sites analysed, 198,512 sites were close enough to a binomial distribution (did not demonstrate over-dispersion), and were thus tested via the Chi-squared. The remaining 195,099 sites all demonstrated over-dispersion, and thus referred to the F-test. Of note is the observation that all of the sites that demonstrate significance under this method are sites that fit to a binomial model, rather than those that were tested using an F-test.
Figure 5.3: Significance values for CpG Sites From Quasibinomial modelling analysis. Sites were sorted by significance level, and then plotted. The upper plot shows all 393,611 sites analysed. The middle plot is a magnification of the area bounded by the blue box (indicated by a blue arrow) in the upper plot, and shows that 69 sites fall under the Benjamini-Hochberg Significance level. The lower plot is again a magnification, this time of the area within the green box in the centre plot (indicated by a green arrow). 8 CpG sites fall under both the Benjamini-Hochberg significance level, and the Bonferroni significance level.
5.2.3 Random Effects Model

In terms of the number of significant sites, the Random Effects model brings us closer again to the number identified using MethylKit. Because the random effects model can’t be fitted reliably if the data contains too few methylation reads, an additional filtering step when using the Random Effects model reduces the number of sites analysed to 328,710. That is, for the random effects model to be fitted reliably, each group must have at least one non-zero sample for methylation. Therefore, a site which has a high amount of coverage but no methylated Cs in one of the groups is excluded. Of the 328,710 sites analysed, we see in Figure 5.4 that 1876 sites meet Benjamini-Hochberg significance, and 35 meet Bonferroni significance.
Figure 5.4: Significance values for CpG sites from Random Effects modelling analysis. Sites were sorted by significance level, and then plotted. The upper plot shows all 393611 sites analysed in this analysis. The middle plot is a magnification of the area bounded by the blue box in the upper plot (indicated by a blue arrow), and shows that 1876 sites fall under the Benjamini-Hochberg Significance level. The lower plot is again a magnification of the area inside the green box in the centre plot (indicated by a green arrow). 35 CpG sites fall under both the Benjamini-Hochberg significance level, and the Bonferroni significance level.
5.3 Differentially Methylated Sites

As described in Chapter 4.3.1, there are multiple ways of calculating the difference in methylation between two groups. The weighted and arithmetic mean differences are the two methods that were used in this analysis. Depending on how little variation there is in coverage between samples in the same group, these two methods can show anything from less than 5% variation in over half of sites, to nearly 30% variation between the two methods in a small proportion of sites. While superficially Figures 5.5 and 5.6 may look similar, we do see movement in an appreciable number of sites. The two 40% differences seen in chromosome 4 when a weighted average is used (Figure 5.5, one at around +40, one at around -40) both disappear when an arithmetic mean difference is calculated (Figure 5.6). For the methylation difference as calculated by the weighted method, the range of values is -42.48 to 42.36, with a mean of -0.95, and a median difference of -0.24 difference (cases minus controls). For the arithmetic method, the range of values is -39.9 to 40.83, with a mean of -1.98, and a median of -0.53. All methylation differences are represented as the percentage difference between the mean of cases and the mean of controls.
Figure 5.5: Difference in Methylation Between Cases and Controls, Weighted by Coverage. Sites are plotted by their location on the chromosomes. Large gaps observed in chromosomes 1, 9, and X correspond to centromeric regions. All differences are calculated as Cases-Controls. Red lines at +15 and -15% difference indicate the magnitude of the difference required to meet threshold difference.

Figure 5.6: Difference in Methylation Between Cases and Controls, Arithmetic Means. Sites are plotted by their location on the chromosomes. Large gaps observed in chromosomes 1, 9, and X correspond to centromeric regions. All differences are calculated as Cases-Controls. Red lines at +15 and -15% difference indicate the magnitude of the difference required to meet threshold difference.
5.4 Significantly Different Sites

Statistical significance and large differences are the two key outcomes desired from this analysis. To be a feasible biomarker or to reflect a plausible biological mechanism for the development of pre-eclampsia, we require both a large difference, and to be assured that this difference is unlikely to have occurred by chance. Therefore, we wish to identify the sites which display both characteristics. The following volcano plots allow us to illustrate the intersect of methylation difference with statistical significance.

5.4.1 MethylKit

As previously mentioned, the MethylKit approach yields the largest number of sites for further investigation. The use of only the weighted methylation difference calculation to find differences between groups, and the relatively lenient q-value result in 3061 CpG sites that demonstrate both a large difference and statistical significance.
**Figure 5.7:** MethylKit Significantly Different Sites. For each site, the $-\log_{10}$ of the q-value generated using MethylKit is plotted against the methylation difference between cases and controls. For MethylKit, methylation difference is calculated using a weighted mean methylation value for each group. The red line denotes the $-\log_{10}$ of q-value = 0.01. Purple lines show the threshold requirement for methylation difference. Significantly different sites are highlighted in blue.

### 5.4.2 Quasibinomial

When the methylation differences between cases and controls as calculated by weighted and unweighted methods are plotted against the significance values assigned using the quasibinomial method, a small number of sites are identified. The effects that different weighting scenarios can have on the outcome of an analysis are clearer in Figure 5.8 than when we look at differences alone. In Figure 5.8a, we see a relatively even fan-shaped spread, with each half of the plot being a rough mirror-image of the other. There is increased density of data around those sites that have little difference, but relatively low p-values, and also corresponding dense regions that have large differences, but high p-values. When compared to Figure 5.8b, a shift in distribution is observed. While the dense region of sites with small differences and small p-values remains, there is a flattening out of the base of the volcano, and sites that were clustered into prongs...
of large and small differences in Figure 5.8a, are now distributed more evenly across
the range. Fewer sites meet the significance and difference thresholds when the methy-
lation is calculated by an arithmetic mean than by a weighted mean (29 vs. 32 CpG
sites at reach significance when corrected for multiple testing using the Benjamini-
Hochberg method, and >15% difference).
Figure 5.8: Quasibinomial Significantly Different Sites. For each site, the $-\log_{10}$ of the p-value generated using the Quasibinomial model is plotted against the methylation difference between cases and controls. 5.8a. Methylation difference is calculated using a weighted mean methylation value for each group. 5.8b. Methylation difference is calculated using and arithmetic mean methylation value for each group. The red line denotes the $-\log_{10}$ of Benjamini-Hochberg significance for 393611 tests, while the green line shows the Bonferroni threshold of significance. Purple lines show the threshold requirement for methylation difference. Significantly different sites are located in the upper left and upper right sixths, and are highlighted in blue.
5.4.3 Random Effects Model

The same plots were made for the reduced number (328710) of CpG sites that were analysed using Random Effects modelling. In Figure 5.9, a similar effect as that observed in Figure 5.8 is seen. In Figure 5.9a, which shows a v-shaped pattern, with a shorter arm on the right hand, or positive portion of the plot. Again, a protrusion is observed around those sites that have little difference, but relatively low p-value (-log scores greater than 3, but only on the negative side of the plot. When compared to a non-weighted analysis (Figure 5.9b), a shift in distribution is observed. There is, as with Figure 5.8b, a flattening out of the volcano, and the plot takes on a squatter, more rounded appearance, and sites in Figure 5.9b, are now distributed more evenly across the range of differences. As with the quasibinomial analysis, fewer sites meet the significance and difference thresholds when the methylation is calculated by an arithmetic mean than by a weighted mean (1075 vs. 983 CpG sites at Benjamini-Hochberg significance, and >15% difference).
Figure 5.9: Random Effects Model Significantly Different Sites. For each site, the $-\log_{10}$ of the p-value generated using the Random Effects model and an F-test is plotted against the mean methylation difference between cases and controls. 5.9a. Methylation difference is calculated using a weighted mean methylation value for each group. 5.9b. Methylation difference is calculated using and arithmetic mean methylation value for each group. The red line denotes the $-\log_{10}$ of Benjamini-Hochberg significance for 328710 tests, while the green line shows the Bonferroni threshold of significance. Purple lines show the threshold requirement for methylation difference. Significantly different sites are located in the upper left and upper right sixths, and are highlighted in blue.
5.5 Selecting Candidates

The progression from 9 lists of significantly different sites to one single list of 'top hits' was a relatively simplistic one, and is illustrated in Figure 4.11. Firstly, the weighted means were compared to the arithmetic means, and those sites that showed greater than 15% difference in both were selected. Subsequently, those sites deemed significant (according to the Benjamini-Hochberg method of correction for multiple testing) for the Quasibinomial and Random Effects approaches were selected, and the single MethylKit list were compared, and those sites that occurred in all three lists were identified. The Quasibinomial approach was the limiting factor in this, having only 26 sites identified as significantly different. All 26 sites were identified in one or both of the other lists. The majority of sites identified using the Random Effects approach were also identified by MethylKit. In total, 25 sites were statistically different according to all three analysis methods, and these 25 were identified as candidates for further analysis and validation.
Figure 5.10: Overlapping sites between all three analysis approaches. 25 Sites are present in all three approaches to analysing RRBS data. Each of the three lists was generated by taking the Benjamini-Hochberg significantly different sites (or \( q < 0.01 \) for MethylKit). The Random Effects approach and MethylKit have 734 sites in common, while the Quasibinomial and MethylKit have 26 sites. All sites that were identified as differentially methylated by the Quasibinomial analysis were also identified by one or both of the other two analyses.

5.5.1 Examining Top Hits In Detail

While these 25 sites show statistic and mathematical differences, for them to be useful as biomarkers, or convincing as molecular mechanisms, the difference in methylation must be observable. To this end, the methylation percentage of each sample was plotted according to its classification as a case or control. The following box plots demonstrate the mean and range of data points within each group for the 25 sites that showed statistical difference in all three analyses (Figures 5.11 - 5.13). Figure 5.11 shows that while the means of each group may show a large difference, the ranges are often broad and show overlap. All of the plots show an overlap between the ranges of cases and controls. The same is shown of the sites in Figures 5.12 and 5.13.
Figure 5.11: Box plots of methylation percentages for all samples as calculated by the arithmetic mean for the top 25 candidate sites (Part 1). The first 9 significantly different sites all show reduced methylation in cases (red) compared to controls (blue).
Of the 25 sites, 17 sites do not show overlap between the interquartile ranges, suggesting that, if they are representative of the population, these sites may be of further interest. In all bar three of these 25 sites, pre-eclamptic placentas show lower levels of methylation than their control counterparts. These sites show no spatial clustering, being spread across the genome, although chromosome 5 appears relatively overrepresented (4 of the 25 sites).

Figure 5.12: Box plots of methylation percentages for all samples as calculated by the arithmetic mean for the top 25 candidate sites (Part 2). The second 9 significantly different sites all show reduced methylation in cases (red) compared to controls (blue), with the exception of the chromosome 6 site at 167,070,067, which shows higher methylation in cases.
Figure 5.13: Box plots of methylation percentages for all samples as calculated by the arithmetic mean for the top 25 candidate sites (Part 2). The final 8 significantly different sites all show reduced methylation in cases (red) compared to controls (blue), with the exception of the chromosome 19 site at 3,136,430, and the chromosome 21 site at 45,845,683, which show higher methylation in cases. The chromosome 16 site at 64,378,035 shows the largest difference in means of all the 'top hits'.

5.5.1.1 Gene Annotation of Top Hits

Gene annotation of the locations gives a clearer idea of the types of sites that demonstrate significant differences in pre-eclamptic placentas. The HOMER programme for motif-discovery and gene-annotation in next-gen sequencing data was used to identify the gene regions that these 25 CpGs are located in. Of the 25, 6 are within 2 kb of a transcription start site (TSS), which may prove to be biologically interesting. None of the 25 are located in or near imprinted genes. The largest difference in methylation is observed at chromosome 16:64378035, which is located in a relative gene desert
- 777 kb from the nearest gene (CDH11) downstream, and 2.2 Mb from the nearest gene upstream (CDH8). The most different site is also the most significant site in all three analyses. CDH11 and CDH8 both encode cadherin precursors that were originally identified in osteoblasts and human brain respectively. CDH11 has also been associated with synovial cell adhesion in rheumatoid arthritis, and both CDH8 and CDH11 have both been implicated in some forms of autism and learning disability (Hussman et al., 2011). CDH11 has also been shown to have possible tumour suppressor activity in a mouse model of retinoblastoma by promoting cell death (Marchong et al., 2010). As apoptosis is increased in pre-eclampsia (Crocker et al., 2003; Ishihara et al., 2002; Longtine et al., 2012), and methylation has decreased by 28% to levels unlikely to suppress transcription (Macaulay et al., 2011), there is the possibility for this gene to play a role in pre-eclampsia. However, the CpG site that is being associated with pre-eclampsia in this study is located a long way from either of these genes, so if it does play a role it is likely that it is acting as either a distant repressor or enhancer. One of the 25 is located in a transcription start site, for the antisense RNA transcript of TRPM2. This long non-coding RNA has been identified as a prognostic marker in prostate cancer, and the knockdown of TRPM2-AS leads to apoptosis (Orfanelli et al., 2015). Interestingly, this is one of only three CpG sites in the top 25 differentially methylated sites that demonstrates higher methylation in pre-eclamptic samples. Three of the remaining sites are located in exons. The remainder of sites are located either between genes (intergenic), or in intronic regions. Weighted and arithmetic methylation differences range from 15-35% differences. As seen in Figures 5.11 - 5.13, only three of the sites show higher methylation in cases than in controls - those in intron two of the ribosomal protein kinase RPS6KA2, and in the 5’UTR of the G protein-binding signalling molecule GNA15 in addition to the TRPM2-AS site.

A brief gene ontology/term enrichment search using the AmiGO2 Term Enrichment Service (http://amigo.geneontology.org/rte) to look at associated biological processes does not reveal any obvious link between these genes and pre-eclampsia or placental development, with seven of the top 11 terms returned being related to neurogenesis.
The remaining four terms are related to locomotion, metal ion transport, inorganic cation transport, and movement of cellular and sub cellular components - all very generalised terms which could, nonetheless, be linked to placental development.

**Table 5.2:** HOMER annotation of top 25 differentially methylated CpG sites. Genomic Locations, Context, Closest Gene, Methylation Differences and Significance Measures, are shown, arranged by chromosome and location. Percentage differences in methylation are calculated as Cases - Controls.

<table>
<thead>
<tr>
<th>Chr</th>
<th>BP</th>
<th>Annotation</th>
<th>Nearest Gene Name</th>
<th>Distance to TSS</th>
<th>Weighted Difference</th>
<th>Arithmetic Difference</th>
<th>Quasibinomial p-value</th>
<th>Random Effects p-value</th>
<th>MethylKit q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11591766</td>
<td>Exon 17 of 21</td>
<td>PTCHD2</td>
<td>52471</td>
<td>-20.77</td>
<td>-21.11</td>
<td>1.04E-06</td>
<td>6.47E-05</td>
<td>2.50E-04</td>
</tr>
<tr>
<td>2</td>
<td>201039692</td>
<td>Intron 1 of 15</td>
<td>RBBM44</td>
<td>270</td>
<td>-21.50</td>
<td>-21.70</td>
<td>7.61E-06</td>
<td>2.31E-04</td>
<td>1.05E-03</td>
</tr>
<tr>
<td>3</td>
<td>49620583</td>
<td>Intergenic</td>
<td>CWH43</td>
<td>631924</td>
<td>-22.61</td>
<td>-20.66</td>
<td>1.70E-07</td>
<td>2.12E-05</td>
<td>2.12E-05</td>
</tr>
<tr>
<td>4</td>
<td>188097861</td>
<td>Intergenic</td>
<td>LOC339975</td>
<td>328906</td>
<td>-18.60</td>
<td>-16.66</td>
<td>3.26E-06</td>
<td>1.52E-04</td>
<td>5.75E-04</td>
</tr>
<tr>
<td>5</td>
<td>85521993</td>
<td>Intergenic</td>
<td>NBPF22P</td>
<td>-56269</td>
<td>-18.51</td>
<td>-17.91</td>
<td>6.92E-06</td>
<td>1.69E-04</td>
<td>9.81E-04</td>
</tr>
<tr>
<td>6</td>
<td>134344570</td>
<td>Intron 5 of 8</td>
<td>PEPER3</td>
<td>-40975</td>
<td>-20.13</td>
<td>-16.73</td>
<td>1.07E-07</td>
<td>2.31E-05</td>
<td>4.37E-05</td>
</tr>
<tr>
<td>7</td>
<td>167070067</td>
<td>Intron 2 of 21</td>
<td>RPS6KA2</td>
<td>-29341</td>
<td>22.06</td>
<td>17.84</td>
<td>7.26E-09</td>
<td>4.93E-06</td>
<td>5.40E-06</td>
</tr>
<tr>
<td>8</td>
<td>16238322</td>
<td>Intron 8 of 8</td>
<td>ISPD</td>
<td>-11789</td>
<td>-24.65</td>
<td>-27.17</td>
<td>3.27E-08</td>
<td>9.16E-06</td>
<td>1.72E-05</td>
</tr>
<tr>
<td>9</td>
<td>73339715</td>
<td>Intron 9 of 30</td>
<td>CDH23</td>
<td>183025</td>
<td>-23.19</td>
<td>-18.33</td>
<td>2.66E-06</td>
<td>1.87E-04</td>
<td>4.94E-04</td>
</tr>
<tr>
<td>10</td>
<td>134016133</td>
<td>Intron 11 of 13</td>
<td>DPYSL4</td>
<td>15719</td>
<td>-20.27</td>
<td>-19.10</td>
<td>6.70E-06</td>
<td>1.37E-04</td>
<td>9.60E-04</td>
</tr>
<tr>
<td>11</td>
<td>134049826</td>
<td>Intron 2 of 11</td>
<td>STK32C</td>
<td>71651</td>
<td>-23.64</td>
<td>-21.17</td>
<td>1.20E-06</td>
<td>6.27E-05</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>12</td>
<td>72565335</td>
<td>Intergenic</td>
<td>PAP6</td>
<td>-1707</td>
<td>15.77</td>
<td>-20.64</td>
<td>5.71E-06</td>
<td>1.30E-04</td>
<td>8.61E-04</td>
</tr>
<tr>
<td>13</td>
<td>27660251</td>
<td>Intron 6 of 27</td>
<td>KIAA0556</td>
<td>98783</td>
<td>-22.79</td>
<td>-22.11</td>
<td>3.35E-06</td>
<td>8.91E-05</td>
<td>5.87E-04</td>
</tr>
<tr>
<td>14</td>
<td>64378035</td>
<td>Intergenic</td>
<td>CDH11</td>
<td>777884</td>
<td>-30.04</td>
<td>-34.35</td>
<td>2.35E-10</td>
<td>3.56E-05</td>
<td>7.35E-07</td>
</tr>
<tr>
<td>15</td>
<td>3136430</td>
<td>5' UTR (Exon 1 of 7)</td>
<td>DNA15</td>
<td>400</td>
<td>19.61</td>
<td>21.41</td>
<td>7.78E-07</td>
<td>4.27E-05</td>
<td>2.01E-04</td>
</tr>
<tr>
<td>16</td>
<td>10454625</td>
<td>Exon 21 of 25</td>
<td>TYK2</td>
<td>26992</td>
<td>-17.31</td>
<td>-19.57</td>
<td>3.70E-06</td>
<td>9.45E-05</td>
<td>6.31E-04</td>
</tr>
<tr>
<td>17</td>
<td>60434721</td>
<td>Intron 6 of 15</td>
<td>CDH4</td>
<td>259805</td>
<td>-24.99</td>
<td>-19.76</td>
<td>1.00E-06</td>
<td>5.77E-05</td>
<td>2.43E-04</td>
</tr>
<tr>
<td>18</td>
<td>22369363</td>
<td>Intergenic</td>
<td>NCA3M2</td>
<td>-1270</td>
<td>-25.54</td>
<td>-25.11</td>
<td>2.77E-07</td>
<td>3.92E-05</td>
<td>9.05E-05</td>
</tr>
<tr>
<td>19</td>
<td>45845683</td>
<td>Promoter-TSS</td>
<td>TRPM2-AS</td>
<td>-237</td>
<td>19.41</td>
<td>15.86</td>
<td>6.03E-06</td>
<td>2.07E-04</td>
<td>8.92E-04</td>
</tr>
<tr>
<td>20</td>
<td>153880762</td>
<td>Exon 2 of 2</td>
<td>CTAG2</td>
<td>1091</td>
<td>-19.58</td>
<td>-23.24</td>
<td>5.84E-07</td>
<td>3.60E-05</td>
<td>1.62E-04</td>
</tr>
</tbody>
</table>

Chrs - Chromosome, BP - Base Position, TSS - Transcription Start Site

### 5.6 Validation of RRBS

In this section, results are presented for the validation of methylation at selected CpG sites for Sequenom. The candidate validation that is presented in this section was based upon an early version of this analysis pipeline which produced different results to those
presented in the previous sections of this chapter. While this analysis pipeline was in development, it was hoped that the MethylKit package alone would be sufficient to produce a reliable result. In addition, this early analysis was based upon incomplete phenotypic data and the MethylKit approach which relies upon only weighted methylation differences. As such, the sites that were validated using Sequenom Sequenom MassArray® Epityper® do not overlap with the list of ‘Top 25’. The sites validated here were analysed using the finalised methods, and used the same samples, so we do know the methylation values as detected by RRBS, and can therefore use them to assess the validity of using Sequenom for targeted resequencing of candidates identified by RRBS.

5.6.1 Sequenom MassArray® Epityper®

The Sequenom platform is a commonly used method for analysing the methylation of specific DNA sequences. Amplicons of bisulphite-converted DNA are fragmented, and the proportion of methylated cytosines within a fragment is measured based on molecular weight. Sequenom is a good method for analysing the same DNA sequence in a large number of samples, and the intent is to use the platform for future targeted validation of methylation differences in other placental cohorts. Therefore, it is important to know that the methylation as detected by RRBS can be validated using Sequenom; i.e., to determine if it is a valid method of validation for this type of data. For the methods used to design the Sequenom assays, please refer to Chapter 2.9. The following figures show the arithmetic mean methylation as calculated from RRBS against the Sequenom mean methylation per group (cases/controls). The plots shown are of the individual sites for which both RRBS data and Sequenom data were available. The number of sites for which data from both platforms is available ranges from 1-4 CpGs per amplicon. Sequenom amplicons were designed, where possible, to cover sites found to be different by RRBS, but also to overlap either with adjacent fragments, or with adjacent CpGs not covered by RRBS fragments. The Sequenom amplicons that were designed
contained more CpG sites than those shown for each gene, but either RRBS data was unavailable for those sites (due to library composition, or low coverage), or Sequenom data was unavailable where it was unable to distinguish between the molecular weights of Sequenom fragments. To determine how well the methylation as measured by Sequenom corresponded to the methylation as measured by RRBS, matched t-tests between Sequenom and RRBS were carried out at each site for the cases and controls separately.

**Figure 5.14:** Methylation in Cases and Controls as measured by Sequenom® and RRBS, for the amplicons of abParts, ADARB2, CAMSAP3, CDH4, CPLX1, and FLG-AS1. The methylation percentage for cases as calculated by each method are shown in shades of red, and the corresponding controls in shades of blue. Paired t-tests between cases and between controls indicate the presence or absence of significant differences between the methods.

In **Figure 5.14**, the majority of the sites validate well, with notable exceptions in the abParts and ADARB2 sites. If the site shows no difference in methylation between
the two methods, a significance level of ‘na’ is expected, demonstrating that there is little or no difference between the methylation in the same samples as measured by two different platforms. abParts and ADARB2 both show large differences between the two measures of methylation. The CAMSAP3 sites perform well, with the exception of the controls at site 19:7,680,266, which do show a significant difference. CDH4 and CPLX1 likewise show good concordance between the two methods.

**Figure 5.15:** Methylation in Cases and Controls as measured by Sequenom® and RRBS, for the amplicons of FLJ26245, GALNT9, GLOD5, GMPPA, GNAS, and GREM1. The methylation percentage for cases as calculated by each method are shown in shades of red, and the corresponding controls in shades of blue. Paired t-tests between cases and between controls indicate the presence or absence of significant differences between the methods.

In Figure 5.15, GMPPA, GNAS and GREM1 all show good concordance. At many sites the difference in methylation between the two methods is actually larger than the difference in methylation between the cases and controls. Much of this is
due to the way in which the candidate sites were selected, which relied solely upon a weighted methylation difference. The fact that this difference disappears when the methylation difference is calculated arithmetically suggests that one or more samples were skewing the mean of the group, and reinforces the choice in the final pipeline to select candidates that show greater than 15% difference in both weighted and arithmetic means. A weighting method like that employed by MethylKit lumps all reads from all cases together, and this not only tips the analysis in favour of CpG sites with high coverage, but it also ignores the variation in phenotype that can occur between cases.

Figure 5.16: Methylation in Cases and Controls as measured by Sequenom® and RRBS, for the amplicons of HAR1B, KCNJ12, KCNQ2, OXCT1, ROHBTB1, SLC16A7, and SNRPN. The methylation percentage for cases as calculated by each method are shown in shades of red, and the corresponding controls in shades of blue. Paired T-tests between cases and between controls indicate the presence or absence of significant differences between the methods.

Figure 5.16 shows good concordance in OXCT1, ROHBTB1, SLC16A7, and
SNRPN, HAR1B, while demonstrating no significant difference between the two measures of methylation, displays a very large range of methylation values in the Sequenom columns of the plot - this reflects the low number of samples that were successfully analysed by Sequenom (n=8 cases, and n=5 controls).

Overall, while Sequenom does appear to validate the methylation values observed in RRBS in most instances, greater sample sizes, and more reliably different candidates must be investigated. What has not been validated here is the methylation differences that caused these sites to be chosen originally. However, as aforementioned, the selection of these sites for their large differences was based upon an incomplete analysis pipeline using the MethylKit approach only. It is of utmost importance however, that the actual methylation values observed by RRBS can be replicated by other methods, and for this type of validation, it is not necessary to select sites that display marked differences between groups. A thoughtful approach, which considers the nature of both RRBS data and Sequenom data is required. These results, while still preliminary, provide a solid platform for future research projects, and indicate that RRBS is able to detect methylation values in a manner that can be replicated using the Sequenom method. Research is currently underway that will utilise the 25 candidate sites identified in Section 5.5, and utilise the pipeline that has been developed to generate a clear picture of how useful Sequenom is as a validation tool for RRBS. In addition, a small but significant level of hypomethylation in pre-eclamptic placentas has been identified.
Chapter 6

Discussion

This project undertook to investigate the genome-wide methylation of the placenta in both its normal state, and in the disease state known as pre-eclampsia. In addition, it aimed to examine the role of imprinting or allele-specific methylation in the placenta. In the way of many PhD projects, the original aims changed over the course of the project, as more was discovered and understood about the technicalities of such research. The method used in the first instance to address these aims - Reduced Representation Bisulphite Sequencing - is not yet fully understood from a bioinformatic and statistical perspective. Therefore, what has resulted from this project has been, in the large part, a methodological advancement towards the analysis of Reduced Representation Bisulphite Sequencing data. Through these advances, a new set of CpG sites have been identified as being differentially methylated in pre-eclamptic placentas. These findings provide a crucial platform for future work on the epigenetic mechanisms of pre-eclampsia, and identifying biomarkers for the development of pre-eclampsia.

6.1 Overview of Main Findings

The aim in the first instance was to generate a genome-wide epigenetic profile of the placenta using Reduced Representation Bisulfite Sequencing - a goal that has
been achieved for 28 placentas of varying gestations and phenotypes. Alongside this achievement, the first RRBS libraries have been produced for pre-eclamptic placentas. By comparing the DNA methylation of pre-eclamptic placentas to placentas not affected by the disorder, some interesting new discoveries have been made. Chief among these is that pre-eclamptic placentas routinely show lower levels of methylation than their unaffected counterparts. Given that placentas are already hypomethylated when compared to somatic tissues, this finding could have implications not only in the clinical diagnosis of pre-eclampsia, but also in our understanding of how, when, and why pre-eclampsia develops. As part of this discovery, a large number of differentially methylated CpG sites have been identified in pre-eclamptic placentas, and work is currently underway to validate these differences in an additional cohort, and to investigate downstream effects of these differences with reference to RNA expression.

6.1.1 Pre-eclamptic Placentas Are Consistently Globally Hypomethylated

While this research was largely targeted towards discovering large, localised differences in methylation between pre-eclamptic placentas and control placentas, these differences proved to be modest (15-30% difference at individual sites). What was a somewhat unexpected outcome was the discovery that, across all 393611 CpG sites tested, there was a highly significant global reduction in methylation in pre-eclampsia. While this difference was relatively small at 1.98% (95% CI 1.97-2.00), this result was highly significant. Of all sites tested, 59% were hypomethylated in pre-eclamptic placentas, and 37% were hypermethylated compared to control placentas. This is consistent with findings by other groups using chip-based methods to examine pre-eclamptic placentas with or without associated IUGR (Blair et al., 2013; Ching et al., 2015; Chu et al., 2014; Hogg et al., 2013). Placental hypomethylation is also observed in other pathologies, notably gestational diabetes mellitus (Nomura et al., 2014; Rong
et al., 2015), as well as being associated with air particulate exposure during pregnancy (Janssen et al., 2013). Obesity appears to have the inverse effect, with global hypermethylation in the placentas from obese mothers compared to controls (Nomura et al., 2014).

6.2 Considerations When Analysing RRBS Data

6.2.1 What Processing and Filtering Must Be Applied To Raw Data?

Raw sequence reads can either be hard trimmed or dynamically trimmed based on the Phred quality score to remove poor quality bases. Hard trimming may remove data that is still of sufficient quality. Dynamic trimming may cut prematurely if quality varies (Figure 4.9). It is important to combine trimming with considerations about how to extract methylation values. If one is using the Bismark methylation extractor, this extracts information for all CpGs on all reads, regardless of quality. While most sites on most reads are expected to be of a high quality, there will inevitably be some poor quality reads or poor quality base calls. Therefore, if using the Bismark methylation extractor, prior trimming may be required. However, if one opts to use the MethylKit methylation extractor to obtain single CpG site information from SAM files, trimming of reads may not actually be required. The piece of MethylKit code that searches for CpG sites within the reads also assesses the quality score of each CpG before extracting the information. By default, a threshold quality requirement is for the Phred score to be greater than or equal to 20. Therefore there is a decision to be made when trimming and extracting CpG data that involves a trade-off between quantity and quality? More data increases coverage and read depth, which in turn may increase confidence in the results. If strict quality controls are applied, even though coverage and depth are decreased, one may end up with more confidence in the result because more of the included data are accurate. A more stringent option, and the approach taken in this study, is to use both quality trimming, and quality-based extraction of CpG values. In future RRBS
experiments that look at individual CpG sites, extracting CpG site information direct from .sam files without any hard trimming should be the method used if the reads are of a high enough quality to successfully map back to the genome.

6.2.2 Technical Considerations in Library Construction and Sequencing

There are numerous possible sources of bias and error during library construction and sequencing. Cluster generation during sequencing can introduce variation into libraries, and is a possible source of variation if some sequences generate clusters more efficiently. In any genetic or epigenetic study, the most ubiquitously used method - PCR - can provide potential variation and error. Quality of starting material, starting copy number, sequence variation due to SNPs, degradation due to bisulphite conversion, and incomplete conversion can all contribute to variation in amplification efficiency. When relatively few cycles of PCR are employed, or an input is of particularly high (and known) quality, this source of variation can be negligible. However, PCRs on bisulphite converted DNA can be easily influenced by degradation, and thus prone to amplification bias in that respect. Differential methylation results in different sequence at the same location after bisulphite conversion, which can also cause differences in amplification efficiency (Warnecke et al., 1997). In studies of differential methylation, this variation can have a large impact. This concept is illustrated well in Chapter 3.5.1.3 - 3.5.1.4, where even when cloned, the A allele of IGFBP1 the exon 1 amplicon yielded very few results. Similarly, the ZNF264 exon 1 amplicon produced clones of only two of the possible four products from a PCR of bisulphite converted DNA. Because of the degradation of DNA during bisulphite conversion, alleles may by chance become differentially degraded. High levels of DNA degradation by bisulphite treatment decrease the number of DNA molecules which are effectively available for PCR amplification (Ehrich et al., 2007). If copies of a small subset of fragments were sufficiently abundant in the RRBS library, then even the few number of cycles used in final
amplification would be enough to get copies to ‘saturation’, and thus be overrepresented in the library. Conversely, scarcely represented fragments may be swamped by their more abundant counterparts, as with each round of amplification, the difference in relative abundance widens, and competition for reagents increases.

6.2.2.1 How Much Does Inter-Individual Genetic Variation Affect Library Constitution?

While humans share 99.9% of their DNA with each other, the remaining 0.1% is responsible for the wide variation that exists among us. Recent estimates place the number of SNPs alone at 10 million in human populations (International HapMap 3 Consortium et al., 2010), without including the variation that is due to microsatellites, copy number variation, and de novo mutations. All of this variation can affect methylation of CpG sites. Allele-specific methylation has been briefly discussed previously in Chapter 1, but genetic variation can affect epigenetic variation in other ways. With specific reference to how RRBS libraries are constructed, genetic variation can create or destroy MspI cut sites, thus creating a slightly different pattern of both CpG sites, and of RRBS fragments. Every RRBS library constructed from a unique individual will contain a different set of MspI fragments. Attempts have been made previously to increase the mapping efficiency of RRBS genomes, and reduce the computational power needed to map by mapping to a ‘Reduced Representation’ genome consisting of 40-220 base pair fragments generated by an in silico digestion of the GRCh37 human genome (Chatterjee et al., 2012b). This approach actually resulted in a slight reduction in mapping efficiency, and this may well be in part because of the aforementioned inter-individual differences.

SNPs and other variants are important to consider not only during library construction and mapping, but also in the final selection of candidate loci. Currently, formal exclusion of CpG sites that overlap with common SNPs is not an exclusion criterion in this pipeline. When the list of top 25 candidate sites showing differential
methylation was investigated for the presence of SNPs using dbSNP 142 in the UCSC Genome Browser, an astonishing 9 of the 25 overlapped with one or more SNPs. Upon closer examination of the allele frequencies, these SNPs were however revealed to have extremely low frequencies, the most common having a minor allele frequency of less than 0.01, thus classifying them all as rare variants that are unlikely to exist in our small sample of individuals. Previous permutations of the analysis (including that used to design the Sequenom assays) also demonstrated high levels of putative SNPs. More than 10% of sites in a single analysis overlapped with SNPs, many at high frequency in human populations. Future analyses should formalise this process of examining CpG sites for the presence of SNPs; setting a threshold for acceptable minor allele frequency that is dependent on the sample size. With larger sample sizes, the chance of encountering rarer alleles that may complicate analyses increases.

6.2.3 A Weighty Issue

A somewhat unexpected side-effect of understanding and developing systems for analysing RRBS data was the need to regress back to questions of basic mathematics. In essence, one of the biggest questions in developing this analysis pipeline can be boiled down to: “How do you calculate the average?”. Until preliminary attempts to use Sequenom to validate the sites shown in Chapter 5.6.1, this question had not arisen. When it became apparent that the methylation difference shown by the initial MethylKit analysis couldn’t be replicated from the raw data (using the previously noted ‘Arithmetic Method’ - Equation 4.3), it forced a re-evaluation of the methods used to obtain those methylation differences in the first place. MethylKit uses a weighted average of each group of samples - a fact that is not explicitly stated in their documentation. This may provide a stumbling block for anyone performing similar case-control cohort analyses.

The reasons against weighting are discussed further in Section 6.6, but the reasons why one might choose to weight next generation sequencing data will be discussed in more detail here. A weighted mean provides a way to control for the differences that
we see in coverage between samples. These differences can on occasion be as much as 20-30 fold due to differences in library constitution, cluster generation, and inter-individual variation. Most immediately obvious however, is that there can be large differences in the number of reads obtained from each library (See Figure 5.1). However, by using a weighted mean calculation, we can place more importance upon those samples with more data. There is a need for a more sophisticated approach than simply adding together all the methylated reads and dividing them by the total number of reads. Technical library variation that occurs during PCR and cluster generation during sequencing can mean that libraries of similar input can still vary widely in coverage and fragment composition. One approach to fixing this would be to ‘spike’ each library with a sequence of known quantity, amplification efficiency, and methylation status. This ‘spike’ could act as a normalising factor by which libraries could be weighted. Similarly, a reference library, again of known constitution, could be used to normalise the data by essentially adjusting the number of reads in a sequenced library; i.e., increasing the relative depth of coverage in a library with low depth, and decreasing the relative depth for libraries with extremely high depth. A third way to control for differences in cluster generation and PCR would be to include two additional samples in each Illumina HiSeq run. These two samples would consist of a pooled library of cases, and a pooled library of controls. Using a weighted approach, while of limited clinical value, may still provide valuable help towards uncovering the causation of a disorder like pre-eclampsia. As with GWAS, if a sufficiently large sample size (thousands of cases and controls) can be obtained, looking at the weighted means of groups of individuals may provide insight into disease mechanisms.

6.2.4 Statistics of Complex Data

Part of the problem in performing appropriate statistical tests on RRBS data is that many aspects of epigenetics are still poorly understood. The prime example of this relates to the independence of sites. When one analyses data, it is necessary to know
whether the data one is examining are correlated or independent. When analyses are performed on multiple CpG sites, as is the case with RRBS, a problem arises. It is not known (except in a very few cases) how the methylation at one site affects the methylation at adjacent sites. Furthermore, due to 3D interactions in the genome, methylation affecting transcription factor binding and methyltransferases may affect methylation at distant sites. Conversely, single CpG sites may be methylated independently of methylation at other locations, or in very small clusters. For these stated reasons, I believe that at this point in our understanding, it would be presumptuous to analyse RRBS data as anything other than individual CpG sites. If sites are analysed independently, patterns of correlation should still be resolvable. A lesson can be taken from the genetics theory behind Genome-Wide Association Studies (GWAS). These studies use independent testing of sites to reveal patterns of linkage, recombination, and correlation.

The first approach to analysing RRBS data that I explored was to use a modified version of the approach used commonly by GWAS. GWAS uses logistic regression to search for regions of the genome (as defined by single nucleotide polymorphisms or SNPs) that associate strongly with a disease state. The idea is that this method could be adapted for use in analysing CpG methylation, if one makes the assumption that a SNP ≈ a CpG site. Both are essentially a single nucleotide polymorphism with two possible states - one allele or the other / methylated or unmethylated. There is added complexity in methylation due to the fact that, while methylation is a binary state in a single cell, we must analyse it as a population of cells, which may not all carry the same methylation pattern (unlike DNA sequence - which is static between cells of the same individual). Techniques for analysing methylation at a single cell level are not yet widely available, so we must look at a pool of DNA from multiple cells and analyse the methylation as a continuous variable between 0 and 1; i.e., as an average methylation score for all of the cells that contributed DNA to the pool.

With those assumptions made, a logistic regression aims to find the equation to best describe the data. It also allows for the integration of covariates such as birth weight, BMI, and gestation if one desires. Early attempts in this project to perform a
simple logistic regression seemed promising. Unfortunately, these early analyses were performed on incomplete data, and with a few key steps of the pipeline missing. Once the gaps in the pipeline had been resolved (such as analysing the CpG as a dinucleotide, rather than two separate observations), applying the most basic logistic regression to this RRBS data did not show any significant values. This result was surprising given the size and variation of the data set. Upon consultation with a statistician (M. Parry, pers. comm.), it was determined that the simple logistic regression was not the ideal approach for this type of data because it is not always normally distributed, and is bounded by limits at 0 and 1. What was gratifying to learn however, was that it was not the strategy itself that was flawed, but the modelling required to fit the data. It was determined that a logistic regression was the appropriate approach to take, but that some alterations needed to be made to better fit the data. Based upon this knowledge, several other models were trialled. The betabinomial performed poorly, and failed to accurately describe the distribution of the data, but trials with the Quasibinomial and Random Effects Models both showed promise, and the decision was made to proceed with both of these models.

6.3 What is the role of Imprinting in Pre-eclampsia?

Although no direct evidence was found to support the hypothesis that imprinted genes play a conspicuous role in pre-eclampsia, absence of evidence is not evidence of absence. This study has just scratched the surface of the complexity that is the epigenetics of pre-eclampsia. Neither IGFBP1 or ZNF264 appear to show pronounced evidence of imprinting or allele-specific methylation, although small but significant differences have been observed in the promoters and first exons of both genes. Regrettably, the RRBS libraries prepared from placental tissue did not contain fragments that overlapped with the promoter of either of these genes. Expression studies of these genes will reveal whether these differences are large enough to produce a phenotype in the
blood of normal individuals, and once the normal state has been established, it should be possible to determine with better clarity, whether either of these genes may in fact play a role in fetal development. What was noted was the presence of relatively low methylation in the promoters of both genes (less than 50 percent at all bar a few sites), and the very low level of methylation in the gene bodies of both genes. While DNA methylation in promoter regions is usually associated with transcriptional silencing, methylation within the gene body is often associated with increased transcriptional efficiency, and use of alternative promoters (Karimi et al., 2011).

There are already examples of the important roles that imprinted genes play in placental development. The paternally expressed gene IGF2 is expressed almost exclusively during embryonic development, and has a number of placental-specific transcripts (Monk et al., 2006b; Shen et al., 1988). IGF2 is involved in the regulation of cell proliferation, growth, differentiation, migration and survival. Cells from pre-eclamptic placentas show increased cell proliferation, poor differentiation and migration, and increased syncytial shedding. Mice that have had this placental transcript knocked down exhibit growth restricted placentas and pups, indicating that IGF2 in the placenta impacts upon resource allocation in the context of the feto-maternal unit, and plays a crucial role in embryonic growth (Jacob et al., 2013). IGFBP1 is a key binder of IGF2 in fetal circulation, and has been isolated in amniotic fluid (Brewer et al., 1988). IGFBP1 is able to bind and sequester IGF2, and is elevated in placental insufficiency and placental hypoxia associated with intrauterine growth restriction - which is often concurrent with pre-eclampsia. When IGFBP1 is elevated in placental insufficiency, there may be increased sequestration of IGF2, causing inappropriate regulation of fetal growth, as observed in the mouse knockdown of IGF2. Dysregulation of IGFBP1 by aberrant methylation during placental development could result in a pre-eclampsia-like phenotype, and warrants further investigation.
6.4 The Complexity of Pre-eclampsia

Ultimately, pre-eclampsia is an incredibly complex disease consisting of many layers that all need to be carefully unpicked in order be fully understood. Many questions remain to be answered, and more still must be asked. This disorder is an amalgamation of clinical disease, cell stress responses, oxidative stress, impaired trophoblast invasion and implantation, and fetal and paternal antigens interacting with the maternal immune system. Underlying everything is the maternal genotype and phenotype, and how this interacts with the fetal genotype and phenotype to form a functional placenta and sustain a pregnancy. Epigenetics gives us an opportunity to look directly at that maternal-fetal interaction, and see how it impacts upon growth and development. We have the opportunity to discover more about the causes and effects of pre-eclampsia.

6.5 A Proposed Role for Methylation in the Genesis of Pre-eclampsia

DNA methylation plays a critical role during embryonic development, particularly in the suppression of plasticity. As development continues, cells become less and less plastic, and more and more specialised. The trophectoderm lineage stays relatively hypomethylated compared to the somatic cells, and it is possible that this pattern of epigenetic control contributes to the robustness of the placenta. Cancers which share phenotypic similarities with the placenta, such as melanoma employ DNA methylation-mediated silencing in cancer progression and survival (Macaulay et al., 2011; Novakovic & Saffery, 2013). The placenta is able to sustain growth and development in adverse circumstances such as confined placental mosaicism, which has parallels with the chromosomal abnormalities that are seen in cancers where a group of cells of one genotype exists in alongside cells with a different genotype (Calvanese et al., 2008; Novakovic et al., 2008). If this global low methylation is reduced even further, as it is
in pre-eclampsia, is it preventing the specialisation of trophoblastic lineages? Are pre-eclamptic placentas ‘immature’ compared to normal placentas because of this methylation aberration, or is the methylation a product of an underlying mechanism? Certainly from a physiological standpoint, pre-eclamptic placentas can be characterised by an immature hyper-proliferative trophoblast phenotype, with poor invasion of the maternal decidua and incomplete remodelling of the maternal spiral arteries (Rolfo et al., 2013).

When the wave of re-methylation occurs after fertilisation, extra-embryonic tissues become re-methylated to a lower extent than do somatic tissues (Reik & Walter, 2001). This global hypomethylation is crucial for the correct function of the placenta and amnion as life support for the developing embryo (Dean et al., 2003; Messerschmidt et al., 2014). However, if the re-methylation process is altered, the hypomethylation may become more pronounced and may in turn contribute to abnormal placentalation, which ultimately results in the placental dysfunction known as pre-eclampsia. We already know that de novo re-methylation of the developing embryo after implantation is highly susceptible to the maternal environment (Dolinoy et al., 2007; Kobor & Weinberg, 2011). It is not a large leap to conclude that the maternal environment around the time of conception can also affect the re-methylation of the extra-embryonic lineages and thus alter the developmental trajectory.

### 6.6 Validation of RRBS Data

Validation techniques that are quantitative (like Sequenom), calculate absolute methylation in individual samples, meaning that it is not necessarily valid to do a direct comparison between the methylation differences as calculated using a weighted equation and an absolute quantifiable methylation value from Sequenom. If one aims to (as this project aims to) use genome-wide approaches to identify new candidates for biomarkers or control mechanisms, it is important that differences in methylation are identifiable at the level of the individual, not just a population level. The crude coverage-based
weighting system of MethylKit may work in a large population to, for example, identify risk factors associated with different ethnicities, or BMIs, but it would provide little utility when attempting to diagnose a patient. Because this study is small, weighting based solely on coverage obscured many results that were just as valid; i.e., any results from those libraries with low coverage. The hypothesis and aim of this project relied on identifying an element (in this case methylation state) that is reliably different between normal and pre-eclamptic placentas. If one wishes to identify a feature that could be exploited as part of a diagnostic test, the weighted mean alone contains little potential. For a diagnostic test to be of use in a clinical setting, there must be a distinction between cases and non-cases, regardless of technical variation. Looking at a weighted mean like Equation 4.2 equates to treating all cases as a single unit, and all controls as a single unit. The approach taken ultimately in this project is, in some senses, a ‘best of both worlds’ approach. The sites that have been identified as differentially methylated were required to show greater than 15% difference between cases and controls regardless of the way that that difference was calculated. A site was deemed differentially methylated if it showed a greater than 15% difference between methylation in cases and controls in both weighted and unweighted calculations.

The issues of PCR bias discussed in Section 6.2.2 may also be contributing to the poor validation of some sites by Sequenom (see Chapter 5.6.1). Both RRBS and Sequenom use PCR as a method to increase the concentration of DNA to analysable levels. Where the difference between them may lie is in the number of rounds of replication each method requires. RRBS uses 15-20 cycles of amplification, and the purpose of the semi-quantitative PCR performed prior to final amplification is to determine the lowest possible number of PCR cycles, and thus reduce the chance of bias. Sequenom however uses 40 rounds of amplification. In part this is due to the fact that in the Sequenom reaction, there is a relatively small amount of starting material, and only one specific sequence is being amplified, while in RRBS there is a large amount of starting material, and many thousands of different sequences. In the higher specificity of Sequenom may lie an inherent weakness. Every piece of DNA exists as two alleles,
and those two alleles may amplify with different efficiencies. In targeted studies, these differences can be teased apart using methods like those used in this thesis to analyse the differential methylation of IGFBP1 and ZNF264. However, such methods are time consuming and laborious, and for the validation of large numbers of candidates a system like Sequenom is much more effective. Unfortunately, Sequenom is unable to differentiate between alleles, and both are analysed as a single unit. Again, for most experiments this would not necessarily be a problem, but if there is a bias, and one allele amplifies more efficiently than the other, the methylation level will be biased towards that allele. In Sequenom, 40 cycles are used - this is enough cycles of amplification to get even a single original copy to saturation.

6.7 Future Directions for Research

It must be recognised that this research has been a pilot study for a wide-ranging ongoing project and as such, there is much more work to be undertaken before a true working knowledge of the epigenetics of pre-eclampsia can be gained. In addition, as has already been stated, the technique of RRBS is still under development. Therefore, work going forward will continue along these two avenues.

6.7.1 Methods for Analysing RRBS Data

Several groups worldwide are currently undertaking the task that is understanding the complexity of understanding the complex nature of epigenetic data. RRBS, while having many positive aspects, also has increased complexity because of the fragmentation and size-selection method it employs. The MethylKit (that has been used and discussed in this project) and BiSeq R packages are the most advanced tools we currently have for analysing this data and, while they incorporate many useful tools for the manipulation and visualisation of RRBS data, there is still a paucity of statistical support for this kind of work. Therefore, future work in this area should be devoted to creating
a solid statistical framework for the analysis of RRBS data, with particular attention paid to validation of the data, so that direct comparisons between methods of detecting methylation can be easily performed.

6.7.2 Epigenetic Markers for Disease

One of the ultimate goals of this work is to develop a reliable test for the early detection and clinical diagnosis of pre-eclampsia. To this end, we aim to identify not a single marker for pre-eclampsia, but multiple markers. These markers can then be combined to create a panel for screening. If multiple sites are tested, then the overall predictive value of the test can be increased. Because of the complex nature of pre-eclampsia, one marker alone is unlikely to have a high degree of power in distinguishing cases and controls, but a panel of moderately powered tests can combine to give high specificity and sensitivity.

It has been demonstrated by multiple research groups that methylation and gene expression can vary spatially across the placenta (Avila et al., 2010; Tzschoppe et al., 2010; Wyatt et al., 2005). In addition, while some of the variation exhibited in methylation and gene expression may be inherent, localised damage, or variation in perfusion can cause stochastic variation between sites. This spatial variability in placental DNA methylation may arise due to the branching villous tree structure of the placenta, such that methylation changes in one villus progenitor may be propagated through that tree independently of events in another villus (Hogg et al., 2014). The recommended way to overcome this spatial variation, and find changes that are systematic is to sample multiple sites across the placenta (Burton et al., 2014) and pool the material to reduce variability due to sampling. Performing multiple sampling can be difficult when there is only a limited amount of tissue. In the case of limited tissue, it may be preferable to focus on specific cell types or structures. Primary trophoblasts, mesenchymal cells, and fibroblasts can be isolated and cultured (Grigoriu et al., 2011; Kim et al., 2012), or micro-dissection can be used to isolate trophoblast cells (Kim et al., 2013), and methy-
lation analyses can be performed on these isolated cell types. The placenta is a complex tissue composed of multiple cell types, and this composition changes throughout pregnancy. The different cell types exhibit different methylation signatures (Novakovic et al., 2009), and therefore the global changes in methylation observed in this study may in actuality reflect differences in tissue composition rather than differences in gene regulation. Use of immunofluorescence based cell sorting can allow for the isolation of different cellular components of the placenta (Grigoriu et al., 2011). By isolating and analysing the different cell lineages of the placenta, we may come closer to identifying the key developmental pathways that lead to pre-eclampsia.

One of the main limitations of this study is the small sample size. As this is a pilot study, and much optimisation of analysis technique was required, a large sample size would have become unwieldy. This small sample size has allowed the generation and streamlining of the processes that will be used in a larger study. A larger study cohort, and a cohort unrelated to the OPuS cohort is desirable to allow the performance of independent validation of findings. A larger cohort would also allow for the differentiation between early and late onset Pre-eclampsia, and the disentanglement of the complexities of placental development with respect to pre-eclampsia in all its variation. However, as a small but diverse range of gestations have been analysed here, it has allowed us to identify features common to both forms of pre-eclampsia.

A main element of the future of this research, is, as already stated, to develop a diagnostic test for pre-eclampsia. As two of the three currently used methods for testing fetal and placental abnormalities during pregnancy are both invasive, and can carry high risk of miscarriage, a better method was needed. Lo et al. have developed a third, less invasive method. Fetal DNA can actually be detected in maternal blood from the 7th week of gestation (Lo et al., 1998), and continues to increase in concentration throughout pregnancy. This ‘cell-free fetal DNA’ (cffDNA) actually consists of fragmented DNA shed from the placenta, and can constitute up to 20% of the DNA in maternal blood by the third trimester (Lun et al., 2008). The methylation signature of cffDNA can be used to differentiate between maternal and fetal DNA fragments.
The DNA methylation seen in cffDNA is comparable to that seen in placental tissue (Jensen et al., 2015). Once differences in placental methylation due to pre-eclampsia have been validated, the next aim is to detect these differences in cffDNA isolated from maternal blood (van den Oever et al., 2013). This would allow for non-invasive testing of pre-eclampsia, such as is already being implemented worldwide for the detection of chromosomal disorders.

If a diagnostic test using biomarkers like the epigenetic ones this project has putatively identified can be created, this will open the door for the development of better treatments. If pre-eclampsia can be detected earlier, and with higher accuracy, then clinicians can make more informed judgements in their management of the disease. Not only this, but having reliable markers for disease can aid in identifying the factors that cause pathogenesis in the first place, and interventions that do not require the premature delivery of an infant can be developed, reducing the risk of complications for both mother and baby. To achieve these goals of both understanding and treating pre-eclampsia, we need good communication and cooperation between clinicians and researchers, so that the complexities of each crucial aspect of development, diagnosis, and treatment can be understood and overcome.

### 6.7.3 Understanding the Pathogenesis of Pre-eclampsia

Our finding of genome-wide hypomethylation prompts the question of cause and effect. Is the hypomethylation observed a result of a systematic placental pathology, or a cause? Until further validation experiments have been performed into the functional relevance of these findings, it is difficult to hypothesise. Functional experiments like RNA sequencing, immunofluorescence, and primary trophoblast invasion assays can be used to determine the biological effects of hypo/hypermethylation at specific candidate regions. If the differences found are having a biological impact upon gene expression and cell function, then their involvement in the pathogenesis of pre-eclampsia is likely
to be of interest.

There is a clear gap in the research of pre-eclampsia, and this lies in how it affects the child born of a pre-eclamptic pregnancy. In addition to the fact that these babies are often born prematurely, and underweight, during a crucial window in their development, many of these babies have been exposed to increased levels of maternal antigens, hypoxia, and inflammation amongst other things. There is much research currently being done (mostly in the first world!) about the effect that metabolic disorders, gestational diabetes, smoking, and maternal obesity have upon how babies develop, and particularly how their brains develop differently, but very little of this research has been aimed at pre-eclampsia. As the most common pregnancy disorder worldwide, maternal hypertension and pre-eclampsia expose millions of developing babies every year to similar physiological conditions as maternal obesity, smoking and similar environmental influences. But we know very little of the long-term consequences. Perhaps part of the reason why so little research has been done into this area is because we currently are able to do very little about pre-eclampsia once it has begun to exhibit symptoms. Knowing more about the long-term consequences of this ancient and frighteningly common disorder will encourage renewed vigour into developing interventions that do not require the delivery of the baby.
6.8 Conclusions

The journey from raw sequencing data to easily understood results can be a tricky one, and much of this PhD has been a matter of navigating way through the murky waters of statistics and bioinformatics - not an easy task for a biologist more used to laboratory work.

Once the data have been generated, it quickly becomes clear that while our molecular biology techniques advance in leaps and bounds, the amount of data being generated is also increasing exponentially, and the bioinformatics theory and tools for analysing complex datasets are playing catch-up. As alluded to in Chapter 4, as molecular biology advances, bioinformatics has a tough job to keep up. Not only that, but biologists have to keep up with the bioinformatics if we hope to have a chance at performing appropriate analyses. In the current environment individuals who not only have the capability to translate research between biology, informatics, and statistics have tremendous value. Attempts are currently being made at an undergraduate level to ensure that molecular biologists will not be thrown in the deep end when it comes to modern analysis techniques, but we still have a long way to go before there is true integration between disciplines.
Appendix A

Reagents, Commercial Kits, and Recipes

A.1 Reagents and Kits

• Agilent Technologies - Pfu Turbo Cx Hotstart DNA Polymerase

• Applied Biosystems - dGTP BigDye™ Terminator v3.0 Ready Reaction Sequencing Kit, BigDye™ Terminator v3.1 Cycle Sequencing Kit

• Beckman Coulter - Agencourt™ AMPure XP

• Illumina - TruSeq® DNA LT Kit

• Invitrogen - TOPO TA Cloning® Kit for Sequencing, LB Broth, LB Agar, Qubit® dsDNA HS Assay Kit

• New England BioLabs - BsrDI, Bsu36I and EcoRI restriction enzymes, Buffers 2 and 3

• Qiagen - QIAmp® DNA Mini & Blood Mini Kit, QIAquick® PCR Purification Kit, MinElute® PCR Purification Kit, QIAquick® Gel Extraction Kit, QIAEX® II Gel Extraction Kit

• Roche - FastStart Taq DNA Polymerase, 10x PCR reaction buffer, 25mM MgCl2, 5x GC-RICH solution, 2.5µM dNTPs, Agarose

• Sigma-Aldrich - Dimethyl Sulfoxide (DMSO), Kanamycin Sulfate, Ampicillin

• Zymo - EZ DNA MethylationTM Kit, DNA Clean & Concentrator-5 Kit, Zyppy® Plasmid Miniprep Kit
A.2 Recipes

Xylene Cyanol Loading Dye
0.25% bromophenol blue
0.25% xylene cyanol
15% Ficoll
TE Buffer

TAE Buffer (10x)
96.8 g Tris base
8.2 g sodium acetate anhydrous
7.6 g Na₂ EDTA (pH 8.0)
MilliQ H₂O to 2000 mL

TE Buffer
10 mM TrisCl (pH 7.6)
1 mM EDTA (pH 8.0)
100 mg/mL Ampicillin
2g Ampicillin Sodium Salt
20 mL 50% methanol
Aliquoted into 400 µL amounts and stored at -20°C. No aliquot was used more than twice.

50 mg/mL Kanamycin
1g Kanamycin Sulfate
20 mL MQ H₂O
Aliquoted into 400 µL amounts and stored at -20°C. No aliquot was used more than twice.

Luria-Bertani Agar and Broth
Both LB Agar and LB Broth were made up from Invitrogen powder stocks (Catalogue #12780029 and #22700041). Both were made up in 400 mL amounts; LB Broth at a concentration of 20 g/Litre, and LB Agar at a concentration of 32 g/Litre.
Ampicillin or Kanamycin (at the above stated concentrations) was added to the LB Broth and LB Agar at a volume of 1 µL/mL before use.
Appendix B

Placenta Documentation

**OPuS PATIENT INFORMATION CONFIDENTIALITY AGREEMENT**

Name: __________________________________________
Position: ________________________________________

I recognize that, in the course of my duties as an investigator or agent of an investigator in the OPuS study that I may gain access to patient information, which is required by law to be kept confidential and which may be disclosed only under limited conditions. I agree that:

1. I will keep confidential all patient information to which I gain access.
2. I will access and use patient information only in connection with the OPuS research protocol.
3. I will not re-disclose patient information except to the extent required by applicable laws, including but not limited to HIV information.
4. I will not discuss patient information in public places or outside of work.
5. I will access information only concerning patients for whom OPuS approval has been given, and will not access information for other patients.
6. I will take all necessary precautions to ensure that the access and handling of patient information are conducted in ways that protect patient confidentiality to the greatest degree possible. This includes maintaining such information in a locked file cabinet.
7. Unless there is written consent from the patient/guardian, if, in the course of my review of patient information, I recognize the patient outside the scope of my practice (for example, if the patient is my acquaintance or neighbour), I will immediately stop reviewing the information and return the chart or, in the case of electronic records, close the applicable file. I will not record any information, even if such information does not identify the patient.

I understand that it is my obligation and responsibility to maintain the confidentiality of all patient information. Improper disclosure or misuse of patient information, whether intentional or due to neglect on my part, is a breach of patient confidentiality, which can result in the loss of access to clinical information and may result in disciplinary action by the university or hospital.

Signature: __________________________ Date: ______________________________

Name _____________________________
Position ____________________________
Witness_____________________________
Name:_____________________________

200
B.1 Gross Pathology Datasheet - Example

OPuS- Otago Placenta Study

Sample Collection Instructions

1) Consent Form signed
   (photocopy and place one copy in medical notes and one for patient)
   Return to patient (either fresh or frozen) YES/ NO
   (if still on ward return to patient at QM)
   If histology required, send to SCL with request for fresh and fresh tissue blocks
   for OpuS. (Third floor hospital "cut-up" room)

2) Take the photo first (maternal and then fetal side)
   (Paper ruler with sticker attached for identification)
   YES

3) Sample cord blood- separate if fresh enough → Freezer
   Clot or separated?

4) Placenta Weight (minus UC and membranes) = ______ g UC weight ______ g

Shape nuclear Haemorrhage

Cord insertion

Intervillous margin closest length to edge ______ mm Length of Cord ______ cm

Placental Size  width ______ cm  x breadth ______ cm

   x thickness ______ mm max ______ mm

Velamentous NO / YES Membranes Normal nonconcealed opaque green

5) Biopsies for formalin - 3 pieces of placental bed and several pieces of UC cord - label cassette

6) And biopsies fresh - 3 pieces of placental bed and several pieces of UC cord – label cassette
   
   ➢ Placental Bed
     o 2x 1cm full-thickness punch/slice biopsies of the placenta from the thickest most healthy
       looking part of the centre of the placental bed
     o Place samples in blocks with OPuS number top and side
   
   ➢ Umbilical Cord
     o 2x 2cm biopsies of the umbilical cord from the most healthy
       looking part divided
   
   ➢ Maternal blood sample and buccal swab, attach NHl stickers

Fresh blocks and blood tube go into minus 80°C freezer
Formalin blocks 48 hours fixation and then to processing
Bag placenta and seal with label, freeze or dispose.
Appendix C

RRBS Library Gross Characteristics

Figure C.1: Number of Reads and Coarse Methylation in RRBS Libraries. A. Comparison between raw number of reads and number of mapped reads in RRBS libraries from cases and controls. No statistical significance was observed between cases and controls, although controls did yield slightly more reads in total, and this translated into more mapped reads. B. Percentage of methylated CpG sites detected in each library. Cases and controls are plotted by gestation along the axis, with rough groupings of pre-term (<37 weeks), and term (37+ weeks) placentas.
References


205


Constancia, M., Hemberger, M., Hughes, J., Dean, W., Ferguson-Smith, A., Fundele, R., Stewart, F., Kelsey, G., Fowden, A., Sibley, C. & Reik,


208


GRIEKSPOOR, T., A. GROOTHUIS (2005). 4Peaks. 46


KERKEL, K., SPADOLA, A., YUAN, E., KOSEK, J., JIANG, L., HOD, E., LI, K., MURTY, V.V., SCHUPF, N., VILAIN, E., MORRIS, M., HAGHIGHI, F. & TY-


**LONGTINE, M.S., CHEN, B., ODIBO, A.O., ZHONG, Y. & NELSON, D.M.** (2012). Villous trophoblast apoptosis is elevated and restricted to cytotrophoblasts in pregnancies complicated by preeclampsia, IUGR, or preeclampsia with IUGR. *Placenta*, **33**, 352–9. 170


219


221


creased preeclampsia at high altitude (3100 meters) in Colorado. _Am J Obstet Gynecol_, **180**, 1161–8. 4


PETERS, J. (2009). Overview of Mammalian Genome special issue on epigenetics. _Mammalian Genome_, **20**, 529–531. 6, 7


224


231


ming in human somatic tissues between fetus and adult. *Epigenetics and Chromatin*, 4, 7. 33


234