The University of Adelaide

DOCTORAL THESIS

Omic characterisation of placental development and phenotype

Author:
Benjamin MAYNE

Primary Supervisor:
Professor Claire ROBERTS

A thesis submitted in fulfilment of the requirements for

the degree of Doctor of Philosophy

in the

Adelaide Medical School

Discipline of Obstetrics and Gynaecology

January 2018

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree. I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: _____

Benjamin MAYNE

Date: 2nd of January_____

The University of Adelaide

Abstract

Discipline of Obstetrics and Gynaecology

Doctor of Philosophy

Omic characterisation of placental development and phenotype

By Benjamin MAYNE

Gene expression is influenced by precise epigenetic mechanisms. In the context of pregnancy proper placental development and pregnancy outcome are dependent upon these mechanisms. These are poorly understood in the placenta and historically have not been investigated. In many biomedical research fields epigenetic modifications such as DNA methylation have been proven to be an effective biomarker. However, this has yet to be shown in the reproduction research field.

The overall aim of this thesis was to investigate new epigenetic mechanisms in placental development and to identify novel biomarkers for phenotype prediction. This thesis firstly focuses on sex-biased gene expression in multiple human tissues to identify targets of sexual dimorphism. Secondly, it investigates novel transcripts in the placenta and finally focuses on using DNA methylation as a biomarker.

Firstly, the research has identified potential new gene targets and mechanisms which may explain sexual dimorphism in many phenotypic traits and diseases. These results suggest that sex-biased gene expression is dynamic and tissue specific. It also highlights the need to consider sex as a biological variable in biomedical research and to address the lack of female representation in many studies.

Secondly, by performing a *de novo* transcript analysis on the placenta this thesis has identified new non-coding RNAs. These placental transcripts were also found to be specific to the placenta and were differentially expressed across gestation and in preeclampsia compared to uncomplicated pregnancies. This suggests these transcripts

may be involved in placental development and may have roles in the pathogenesis of preeclampsia. Identifying novel placenta specific transcripts has uncovered new research opportunities involving the placenta. There are potentially hundreds of other unannotated transcripts in the placenta which may have roles in placental development and may be crucial to a successful pregnancy outcome.

Thirdly, using DNA methylation as a biomarker has led to the development of two key prediction models. The first one used the level of methylation at 62 cytosine-phosphate-guanosine (CpG) sites to determine the gestational age of a placenta. This computational tool was also used to identify placental aging in placentas from women with early onset preeclampsia. This tool points to potential mechanisms underpinning placental aging which may have an impact on pregnancy complications. The second prediction tool has identified 84 methylated sites in the methylome of maternal circulating leukocytes which can distinguish five pregnancy outcomes. This tool has potential clinical application to identify women at risk of a pregnancy complication. This would enable clinicians to intervene and potentially prevent or reduce morbidity and mortality for mother and child.

In summary, this thesis has focused on sex differences in gene expression and DNA methylation in placental development. It has also shown that DNA methylation has potential as an effective biomarker in the field of reproduction research.

Acknowledgements

The work presented here in this thesis is the result of a collaborative effort. I would like to thank and express my gratitude to my supervisors Claire Roberts, Tina Bianco-Miotto and James Breen. I appreciate their time and patience throughout the course of my candidature. I am also thankful for their knowledge and skills in many areas such as reproductive biology, molecular biology, epigenetics and bioinformatics. I would also like to thank support outside of my supervisor panel. I thank Carlos Rodriguez Lopez who was involved in the data creation of work that has been presented here in this thesis. In addition, this thesis has used genomic data available in public repositories. I would therefore like to thank the scientific community for making their data available which has been used to great lengths in this thesis.

I would also like to acknowledge individuals who took time to assist me during my project. I would also like to thank Shalem Leemaqz who has assisted me with statistical support. I would also like to thank Dylan McCullough and Dale McAninch who have assisted me with laboratory work. I would also like to thank other current and past postdoctoral fellows in the CTR lab or have provided helpful discussions; Prabha Andraweera, Jessica Phillips, Tanja Jankovic-Karasoulos and Jessica Grieger. I would also like to acknowledge other post-graduate students in the CTR lab; Rebecca Wilson, Petra Verburg, Michelle Plummer and Amy Garrett.

I would also like to acknowledge the generosity of funding sources that have made this project possible. I would like to thank the Department of Education for awarding me an Australian Postgraduate Award, enabling personal financial stability during my candidature. I would also like to thank the National Health and Medical Research Council (NHMRC) for awarding project grants to the Placental Development Laboratory.

Finally, I would like to thank my mum, sister and brother who have supported me through both of my undergraduate and postgraduate studies.

Publications Arising from this Thesis

1. **Mayne BT**, Bianco-Miotto T, Buckberry S, Breen J, Clifton V, Shoubridge C, Roberts CT: Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans. ***Frontiers in Genetics*** 2016, 7.

2. Mayne BT, Leemaqz SY, Smith AK, Breen J, Roberts CT, Bianco-Miotto T: **Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation.** *Epigenomics* 2017, **9:**279-289.

Contents

List of Figures

List of Tables

1 Introduction

1.1 The epigenetic landscape of the developing placenta is essential to a successful pregnancy outcome

The placenta comprises an important part of the conceptus but is only present for a short period of time, where term is defined as 37-41 weeks' gestation. Despite its short life it has the substantial task of providing for the needs of the growing fetus. It protects the fetus from the maternal immune system, acts as interface for the exchange of nutrients, gases and wastes between the mother and fetus, and orchestrates maternal adaptation to pregnancy. The placenta itself undergoes drastic changes during development which are essential for a successful pregnancy outcome.

In order for the placenta to undergo dynamic change across gestation to facilitate the needs of the developing fetus, there must be tightly controlled mechanisms at the genomic level. Alterations in the mechanisms controlling placental transcription during development have been implicated in pregnancy complications such as preeclampsia and intrauterine growth restriction [1]. Many defects in placental function may stem from early on in gestation during trophoblast invasion of the maternal vasculature. Impaired implantation and placentation during the early stages of development can increase risk for pregnancy complications such as miscarriage, preeclampsia, preterm birth and intrauterine growth restriction [2]. Despite the knowledge of the link between placental development and pregnancy complications, the placenta remains the most poorly understood organ in the human body [3].

The placenta originates from the conceptus and is therefore genetically identical to the fetus. Placental development is directed by the fetal genome but there are a number of factors that tightly regulate transcription such as the epigenetic landscape. For example, DNA methylation has been extensively studied in the placenta in terms of genomic imprinting, which is essential for placental and fetal development. Furthermore, alterations in placental DNA methylation levels have been shown to associate with pregnancy complications [4]. Despite increasing literature, the precise mechanisms by which

epigenetics and transcription interact with one another and how these interactions relate to pregnancy success remains to be elucidated.

1.2 The developing placental transcriptome

The massive amount of high resolution gene expression data that is generated by next generation sequencing (NGS) enables the identification of previously unannotated transcripts in the human genome [5, 6]. These analyses are essential to identifying new biological pathways which could potentially play a role in various diseases. There have been a number of RNA sequencing studies that have focused on the placenta [7-9], however these have focused on the term placenta and therefore the first trimester placenta remains an enigma. In addition, most studies have been confined to group comparisons and have not dived deeper into transcriptomic analysis of the placenta.

Thus far, gene expression analyses have been carried out on the first trimester placenta using gene expression microarrays [10, 11]. These microarray experiments rely on a priori transcript information and hence, cannot identify novel transcriptional events. In other areas of genomic research, studies have been identifying novel transcripts and determining their association in a range of diseases including cancers [12, 13] and neurological disorders [14]. Detecting novel transcripts in a tissue-specific manner enables the identification of new biological pathways and deeper understanding of the transcriptome environment. Such studies are essential to improve understanding of the role of gene expression in the development of the placenta in health and disease. Although large scale analyses have focused on novel transcripts in a range of human tissues [15], no such study has conducted these type of analyses in the placenta. In this thesis, RNA-seq will be used to perform a de novo transcript analysis on placental tissue to fill this gap in our knowledge.

1.3 The importance of considering sex as a biological variable

Many biomedical researchers avoid the use of female mammals in experimental manipulations. There is a common assumption that results from males can be directly applied to females. However, since sex differences in body metabolism are well known

[16, 17] research should conduct experiments on both sexes. One of the misconceptions of females not being used in biomedical research is that their hormonal cycles may be a confounding factor [18, 19]. However, some studies have shown that females can still give consistent results comparable to males [20, 21]. Despite this, there is a bias of using males in 8 out of 10 biological disciplines [22]. Conducting research only on male subjects can have implications for women's health. Since there are sex differences in the absorption and metabolism of drugs [23], it is potentially unknown how a drug may behave specifically in females since the research was conducted only on males.

Sex differences in physiology and pathology are well known, however it is unknown how these differences arise. Sexually dimorphic traits can be contributed by sex chromosomes, sex hormones and reproductive factors, but may extend beyond these factors. Despite sharing similar genomes, differences in gene regulation between sexes may be an underlying mechanism in many sexually dimorphic traits. Previous studies have reported sex biased gene expression in the human brain [24-28], pancreas [29], heart [30], placenta [31] and the liver [32]. One major difference between males and females is the presence of the sex chromosomes (XX in females and XY in males) that contribute a large proportion of the total number of sex biased genes within a tissue. Genes located on the Y chromosome, which females don't have, are not considered to be differentially expressed between sexes. However, Y chromosome genes may have an important role in the male phenotype in a range of tissues. To compensate for gene dosage in mammalian female somatic cells, one X chromosome is randomly inactivated by the long non-coding RNA, XIST, in a process referred to as X chromosome inactivation (XCI). In humans, up to 15% of genes escape XCI, unlike in mice where there is relatively no escape of XCI [33, 34]. Sex biased gene expression can extend beyond the X chromosome and into genes located on autosomes. These differences may contribute to sex differences in the prevalence of certain diseases and disorders.

Sex differences in autosomal gene expression and non-reproductive factors have been shown to be associated with brain disorders such as multiple sclerosis (MS) [35, 36] and epilepsy [37]. Furthermore, genes that encode for pro-inflammatory factors such as heat shock proteins (HSPs) have been shown to be upregulated in female brains compared to

males [38]. In addition to other organs, isolated systolic hypertension can be up to 14% more prevalent in females compared to males [39]. Sex biased gene expression may result in sex differences in how many prescribed medications are absorbed and metabolized [40]. Not taking these factors into consideration may have devastating consequences for females and it may also be the reason why females are 1.5 times more likely to have an adverse reaction to a prescription drug compared to males [41]. Therefore, studying sex differences in gene expression, if the true extent of sex differences can be elucidated, may help identify possible mechanisms for sexually dimorphic traits. In this thesis I will make use of publicly available data to investigate sex differences in gene expression and DNA methylation.

1.4 Machine learning applications in epigenetic studies

Advances in technology in the past decade has substantially reduced the time and cost of sequencing a human genome. The cost has reduced from $3 billion to currently $1000 [42]. This has created a wealth of data in which to conduct large scale genomic studies. However, our ability to identify meaningful patterns in data has proven to be difficult. Genomic data such as DNA methylation data suffers from the problem of high dimensionality [43]. This can result in high computational costs when analysing data with potentially millions of variables.

Machine learning comprises techniques in which computer algorithms can be used to identify meaningful results from large sets of data [43]. In the field of genomics machine learning can be used to identify potential biomarkers and possible biological mechanisms. The methods used in machine learning can be divided into two main categories, unsupervised and supervised learning [44]. Unsupervised techniques summarise the data such that an overview of the data can be interpreted. By contrast, supervised learning constructs models using the data to predict an outcome. Supervised learning can be used with epigenetic data to construct models which can be used as either prognostic or diagnostic tools. These methods have been widely used in cancer prognosis and diagnostics [45, 46]. However, these methods are gaining traction in other applications in

clinical research and can potentially be used with epigenetic data for the prediction of a wide variety of diseases.

1.5 DNA methylation can be used as biomarkers for pregnancy complications

A biomarker is a factor that can be measured to give a prediction of either a normal biological or pathogenic process [47]. Many complex, pleiotropic human traits and disease states are the result of a mixture of both genetic and environmental factors and as such, biomarkers should reflect each so that they provide accurate prediction. Epigenetic modifications, influence gene expression but do not alter the DNA sequence [48], they are relatively stable [49] and reflective of both genetic and environmental factors, making them ideal biomarkers for identifying diseased states.

Epigenetic modifications include DNA methylation which is the addition of a methyl group to a cytosine residue. In the human genome, DNA methylation occurs predominantly at cytosine-phosphate-guanine (CpG) dinucleotides [50]. Most epigenetic studies in humans have focused on CpG methylation through the use of microarrays or sequencing based methods. Currently, pregnancy research has focused on epigenome-wide studies which have identified DNA methylation differences in pregnancy complications [4]. While multiple studies have focused on individual sites [51-53], differentially methylated regions [54] or trait association [55] few studies have investigated the use of stable methylated sites that could be biomarkers and act as diagnostic tools. Currently, no epigenetic biomarkers exist for the prediction of pregnancy complications. Furthermore, previous studies have focused on microarray technology which only captures a limited number of CpG sites in the genome. However, with the increased resolution of epigenomic markers generated by NGS technology, there is the potential for using such data for biomarker discovery in pregnancy research. Furthermore, NGS technology is capable of capturing more markers as opposed to microarray technology and therefore can be used to identify novel biomarkers.

1.6 Summary

The aim of this thesis was to investigate the epigenetic dysregulation in pregnancy complications. In addition, it also focused on the epigenetic landscape during normal placental development. The analysis presented here in this thesis was primarily conducted using DNA methylation data. However, gene expression data was also generated to gain a further understanding of the placental transcriptome. Gene expression data was also used in other human tissues to build a comprehensive resource of sex biased gene expression. Furthermore, many of the bioinformatic methods did not exist prior to this thesis and to my knowledge this is the first time they have been used for these applications.

References

1. Guo L, Tsai SQ, Hardison NE, James AH, Motsinger-Reif AA, Thames B, Stone EA, Deng C, Piedrahita JA: **Differentially expressed microRNAs and affected biological pathways revealed by modulated modularity clustering (MMC) analysis of human preeclamptic and IUGR placentas.** *Placenta* 2013, **34:**599-605.
2. Roberts CT: **IFPA Award in Placentology Lecture: Complicated interactions between genes and the environment in placentation, pregnancy outcome and long term health.** *Placenta* 2010, **31 Suppl:**S47-53.
3. Burton GJ, Fowden AL: **The placenta: a multifaceted, transient organ.** *Philos Trans R Soc Lond B Biol Sci* 2015, **370:**20140066.
4. Bianco-Miotto T, Mayne B, Buckberry S, Breen J, Rodriguez Lopez C, Roberts C: **Recent progress towards understanding the role of DNA methylation in human placental development.** *Reproduction* 2016.
5. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al: **The landscape of long noncoding RNAs in the human transcriptome.** *Nat Genet* 2015, **47:**199-208.
6. Bonnal RJ, Ranzani V, Arrigoni A, Curti S, Panzeri I, Gruarin P, Abrignani S, Rossetti G, Pagani M: **De novo transcriptome profiling of highly purified human lymphocytes primary cells.** *Sci Data* 2015, **2:**150051.
7. Sober S, Reiman M, Kikas T, Rull K, Inno R, Vaas P, Teesalu P, Marti JM, Mattila P, Laan M: **Extensive shift in placental transcriptome profile in preeclampsia and placental origin of adverse pregnancy outcomes.** *Sci Rep* 2015, **5:**13336.
8. Kaartokallio T, Cervera A, Kyllonen A, Laivuori K: **Gene expression profiling of pre-eclamptic placentae by RNA sequencing.** *Sci Rep* 2015, **5:**14107.
9. Kim J, Zhao K, Jiang P, Lu Z, Wang J, Murray JC, Xing Y: **Transcriptome landscape of the human placenta.** *BMC Genomics* 2012, **13:**115.
10. Founds SA, Conley YP, Lyons-Weiler JF, Jeyabalan A, Hogge WA, Conrad KP: **Altered global gene expression in first trimester placentas of women destined to develop preeclampsia.** *Placenta* 2009, **30:**15-24.
11. Uuskula L, Mannik J, Rull K, Minajeva A, Koks S, Vaas P, Teesalu P, Reimand J, Laan M: **Mid-gestational gene expression profile in placenta and link to pregnancy complications.** *PLoS One* 2012, **7:**e49248.
12. Kazemian M, Ren M, Lin J-X, Liao W, Spolski R, Leonard WJ: **Comprehensive assembly of novel transcripts from unmapped human RNA-Seq data and their association with cancer.** *Mol Syst Biol* 2015, **11:**826.
13. Verma A, Jiang Y, Du W, Fairchild L, Melnick A, Elemento O: **Transcriptome sequencing reveals thousands of novel long non-coding RNAs in B cell lymphoma.** *Genome Med* 2015, **7:**110.

14. Johnson MB, Wang PP, Atabay KD, Murphy EA, Doan RN, Hecht JL, Walsh CA: **Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex.** *Nat Neurosci* 2015, **18:**637-646.
15. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y: **The landscape of long noncoding RNAs in the human transcriptome.** *Nat Genet* 2015, **47**.
16. Blaak E: **Gender differences in fat metabolism.** *Curr Opin Clin Nutr Metab Care* 2001, **4:**499-502.
17. Tipton KD: **Gender differences in protein metabolism.** *Curr Opin Clin Nutr Metab Care* 2001, **4:**493-498.
18. Mogil JS, Chanda ML: **The case for the inclusion of female subjects in basic science studies of pain.** *Pain* 2005, **117:**1-5.
19. Beery AK, Zucker I: **Sex bias in neuroscience and biomedical research.** *Neurosci Biobehav Rev* 2011, **35:**565-572.
20. Gold SM, Voskuhl RR: **Estrogen treatment in multiple sclerosis.** *J Neurol Sci* 2009, **286:**99-103.
21. Veliskova J: **Estrogens and epilepsy: why are we so excited?** *Neuroscientist* 2007, **13:**77-88.
22. Zucker I, Beery AK: **Males still dominate animal studies.** *Nature* 2010, **465:**690.
23. Hughes RN: **Sex does matter: comments on the prevalence of male-only investigations of drug effects on rodent behaviour.** *Behav Pharmacol* 2007, **18:**583-589.
24. Trabzuni D, Ramasamy A, Imran S, Walker R, Smith C, Weale ME, Hardy J, Ryten M: **Widespread sex differences in gene expression and splicing in the adult human brain.** *Nat Commun* 2013, **4:**2771.
25. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, et al: **Spatio-temporal transcriptome of the human brain.** *Nature* 2011, **478:**483-489.
26. Vawter MP, Evans S, Choudary P, Tomita H, Meador-Woodruff J, Molnar M, Li J, Lopez JF, Myers R, Cox D, et al: **Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes.** *Neuropsychopharmacology* 2004, **29:**373-384.
27. Weickert CS, Elashoff M, Richards AB, Sinclair D, Bahn S, Paabo S, Khaitovich P, Webster MJ: **Transcriptome analysis of male-female differences in prefrontal cortical development.** *Mol Psychiatry* 2009, **14:**558-561.
28. Reinius B, Jazin E: **Prenatal sex differences in the human brain.** *Mol Psychiatry* 2009, **14:**987, 988-989.
29. Hall E, Volkov P, Dayeh T, Esguerra JL, Salo S, Eliasson L, Ronn T, Bacos K, Ling C: **Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets.** *Genome Biol* 2014, **15:**522.
30. Fermin DR, Barac A, Lee S, Polster SP, Hannenhalli S, Bergemann TL, Grindle S, Dyke DB, Pagani F, Miller LW, et al: **SEX AND AGE DIMORPHISM OF**

MYOCARDIAL GENE EXPRESSION IN NONISCHEMIC HUMAN HEART FAILURE. *Circulation Cardiovascular genetics* 2008, **1:**117-125.
31. Buckberry S, Bianco-Miotto T, Bent SJ, Dekker GA, Roberts CT: **Integrative transcriptome meta-analysis reveals widespread sex-biased gene expression at the human fetal-maternal interface.** *Mol Hum Reprod* 2014, **20:**810-819.
32. Zhang Y, Klein K, Sugathan A, Nassery N, Dombkowski A, Zanger UM, Waxman DJ: **Transcriptional Profiling of Human Liver Identifies Sex-Biased Genes Associated with Polygenic Dyslipidemia and Coronary Artery Disease.** *PLoS One* 2011, **6:**e23506.
33. Carrel L, Willard HF: **X-inactivation profile reveals extensive variability in X-linked gene expression in females.** *Nature* 2005, **434:**400-404.
34. Yang F, Babak T, Shendure J, Disteche CM: **Global survey of escape from X inactivation by RNA-sequencing in mouse.** *Genome Res* 2010, **20:**614-622.
35. Voskuhl RR, Palaszynski K: **Sex hormones in experimental autoimmune encephalomyelitis: implications for multiple sclerosis.** *Neuroscientist* 2001, **7:**258-270.
36. Ebers GC, Sadovnick AD, Dyment DA, Yee IM, Willer CJ, Risch N: **Parent-of-origin effect in multiple sclerosis: observations in half-siblings.** *Lancet* 2004, **363:**1773-1774.
37. Christensen J, Kjeldsen MJ, Andersen H, Friis ML, Sidenius P: **Gender differences in epilepsy.** *Epilepsia* 2005, **46:**956-960.
38. Lin LC, Lewis DA, Sibille E: **A human-mouse conserved sex bias in amygdala gene expression related to circadian clock and energy metabolism.** *Mol Brain* 2011, **4:**18.
39. Maas A, Appelman YEA: **Gender differences in coronary heart disease.** *Neth Heart J* 2010, **18:**598-602.
40. Anderson GD: **Sex and racial differences in pharmacological response: where is the evidence? Pharmacogenetics, pharmacokinetics, and pharmacodynamics.** *J Womens Health (Larchmt)* 2005, **14:**19-29.
41. Zopf Y, Rabe C, Neubert A, Gassmann KG, Rascher W, Hahn EG, Brune K, Dormann H: **Women encounter ADRs more often than do men.** *Eur J Clin Pharmacol* 2008, **64:**999-1004.
42. Hayden EC: **Technology: The $1,000 genome.** *Nature* 2014, **507:**294-295.
43. Libbrecht MW, Noble WS: **Machine learning applications in genetics and genomics.** *Nat Rev Genet* 2015, **16:**321-332.
44. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer New York; 2009.
45. Levenson VV: **DNA methylation as a universal biomarker.** *Expert Rev Mol Diagn* 2010, **10:**481-488.
46. Shi T, Gao G, Cao Y: **Long Noncoding RNAs as Novel Biomarkers Have a Promising Future in Cancer Diagnostics.** *Dis Markers* 2016, **2016:**9085195.
47. Naylor S: **Biomarkers: current perspectives and future prospects.** *Expert Rev Mol Diagn* 2003, **3:**525-529.

48. Bird A: **Perceptions of epigenetics.** *Nature* 2007, **447:**396-398.
49. Teh AL, Pan H, Chen L, Ong ML, Dogra S, Wong J, MacIsaac JL, Mah SM, McEwen LM, Saw SM, et al: **The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes.** *Genome Res* 2014, **24:**1064-1074.
50. Mikeska T, Craig JM: **DNA Methylation Biomarkers: Cancer and Beyond.** *Genes* 2014, **5:**821-864.
51. Liu L, Zhang X, Rong C, Rui C, Ji H, Qian YJ, Jia R, Sun L: **Distinct DNA methylomes of human placentas between pre-eclampsia and gestational diabetes mellitus.** *Cell Physiol Biochem* 2014, **34:**1877-1889.
52. Anton L, Brown AG, Bartolomei MS, Elovitz MA: **Differential methylation of genes associated with cell adhesion in preeclamptic placentas.** *PLoS One* 2014, **9:**e100148.
53. Chu T, Bunce K, Shaw P, Shridhar V, Althouse A, Hubel C, Peters D: **Comprehensive Analysis of Preeclampsia-Associated DNA Methylation in the Placenta.** *PLoS One* 2014, **9**.
54. Hoyo C, Murtha AP, Schildkraut JM, Jirtle R, Demark-Wahnefried W, Forman MR, Iversen ES, Kurtzberg J, Overcash F, Huang Z, Murphy SK: **Methylation variation at IGF2 differentially methylated regions and maternal folic acid use before and during pregnancy.** *Epigenetics* 2011, **6:**928-936.
55. Joubert Bonnie R, Felix Janine F, Yousefi P, Bakulski Kelly M, Just Allan C, Breton C, Reese SE, Markunas Christina A, Richmond Rebecca C, Xu C-J, et al: **DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis.** *Am J Hum Genet* 2016, **98:**680-696.

2 Literature Review

Abstract

Pregnancy complications can cause perinatal and lifelong health problems for the child. Prediction of complications and interventions that prevent them could significantly improve the future quality of life for the child. Prediction of pregnancy complications prior to the onset of disease is a major challenge for clinicians and researchers as there are currently no accurate predictors. Non-invasive biomarkers are promising candidates for the prediction of pregnancy complications due to their easy accessibility. Previous studies have focused on differences in profiles of placental DNA methylation in uncomplicated and complicated pregnancies, showing that identified methylated sites may be potential biomarkers. In addition, these studies have also focused on expression profiles of non-coding RNAs (ncRNA) in maternal blood and the placenta. These studies have identified biological mechanisms where ncRNAs have a role in the development of certain pregnancy complications. Furthermore, specific ncRNAs have been shown to be present in maternal plasma depending on pregnancy outcome. This review focuses on DNA methylation and ncRNAs as biomarkers of pregnancy complications. It also focuses on the roles of epigenetic modifications in the development of the placenta and how they relate to the onset of pregnancy complications. Finally, it will also describe recent work relating to the use of epigenetic modifications as biomarkers in pregnancy.

2.1 Introduction

Pregnancy is a physiological state where the mother's body faces significant challenges. In order for a successful pregnancy outcome, maternal physiological adaptations include insulin resistance, increased cardiac output and increased glomerular filtration rate [1]. Despite the adaptability of a woman's body in pregnancy, complications can arise. Pregnancy complications including preeclampsia (PE), which is characterised by high maternal blood pressure and proteinuria, preterm birth (PTB), when the baby is born prior to 37 weeks' gestation, intrauterine growth restriction (IUGR), when birthweight is < 5th

centile and gestational diabetes mellitus (GDM), affect up to 1 in 4 pregnancies [1]. These pregnancy complications can be life threatening to the growing fetus and mother but also impact the lifelong health of the child. Pregnancy complications can result from defective early placental morphogenesis which impacts placental function and, thereby, the health of both the mother and fetus later in gestation [1]. Genetic and environmental factors contribute to pregnancy success. These impact epigenetic modifications in the placenta, fetus and maternal tissues which may have a profound impact on fetal development and pregnancy outcome [2].

An increasing number of studies have demonstrated that aberrant epigenetic modifications may be the underlying mechanism in many diseases such as some cancers, cardiovascular and autoimmune diseases and pregnancy complications [3, 4]. One of the most commonly studied epigenetic modifications, DNA methylation, has increasing prominence as a biomarker in many biomedical research fields. DNA methylation has been used as a biomarker in diagnosis and prognosis in many types of cancers [5]. In addition, non-coding RNAs (ncRNAs) also have great promise as biomarkers and have been used in cancer diagnostics [6]. A biomarker is any biological measure that can be used to predict the risk of a pathological condition occurring [7]. Generally, peripheral fluids such as blood and saliva are ideal, as they have been widely used in studies as a non-invasive method for disease prediction [8, 9]. For example, DNA methylation levels at specific sites in the peripheral blood have been shown to be associated with colorectal cancer [10]. Furthermore, detection of a biomarker must not only be reproducible but provide high sensitivity and specificity for accurate prediction.

Currently, cancer related applications of epigenetic modifications as biomarkers, dominate research. However, there is the potential for using epigenetic modifications, including DNA methylation and ncRNAs as biomarkers of monitoring pregnancy development and outcome. In this review, we highlight studies that have identified key epigenetic biomarkers of pregnancy development and outcome. We also discuss possible biological functions of the epigenetic modifications in the context of fetal development.

2.2 Epigenetics

Epigenetics can be defined as the study of heritable modifications that alter gene expression without changing the underlying DNA sequence [11, 12]. Epigenetics covers a broad range of modifications that occur to the underlying DNA sequence without changing the genetic code. We will confine our review to DNA methylation and ncRNAs. The form of DNA methylation discussed will specifically be the addition of a methyl group (CH3) to a cytosine residue. However, it is acknowledged that there are other forms of DNA methylation that exist [13]. DNA methylation is maintained by DNA methyltransferases (DNMT1, DNMT3A and DNMT3B) which generally act to repress gene expression through the addition of methyl groups within promoter regions [14]. On the other hand ncRNAs, which were originally thought to be non-functional junk, also feature in transcriptional regulation [15]. The length of ncRNAs is used to define their class. The first class is short ncRNAs (20-200 nt) and the other class, which makes up 80% of ncRNAs, is long non-coding RNAs (lncRNAs) (200nt – 1kb) [6, 16, 17]. Placental development is dependent upon ncRNA expression and genomic imprinting. Imprinted genes are expressed in a parent of origin fashion and are epigenetically silenced by DNA methylation, histone modifications or lncRNAs [18, 19].

Despite epigenetic modifications having great promise as biomarkers, there is little knowledge on the molecular mechanisms by which chosen markers act. . DNA methylation is well known for its role in regulating gene expression levels [20]. DNA methylation at cytosine-phosphate-guanosine (CpG) sites nearby the transcription start site of genes commonly represses gene expression [21] by preventing binding of transcription regulatory factors which would otherwise increase expression. However, DNA methylation has also been shown to act over long range distances through chromatin remodelling [21], increasing the complexity of regulation of gene expression levels. Moreover, DNA methylation is not the only epigenetic regulator of gene expression. The expression of ncRNAs can also influence gene expression by recruiting transcription factors or mediating heterochromatin assembly [22]. The regulation of gene expression is mediated by different

epigenetic factors which act together [23]. However, many of the precise mechanisms remain to be elucidated but they may uncover novel targets for a range of future research questions.

2.3 Epigenetic modifications in placental development

The placenta still remains the most poorly understood organ in the human body [24]. Although the placenta only exists for a short period of time, it has the substantial task of providing protection from the maternal immune system and nutrient, gaseous and waste exchange for the growing fetus among other functions [1, 25]. Implantation of the blastocyst into the decidua initiates the differentiation of various trophoblast cell types which later form the epithelial populations in the placenta. Implantation and early placental morphogenesis are mediated by extravillous cytotrophoblast invasion of the decidua and its vasculature, which uses mechanisms similar to tumour metastasis [26]. Improper implantation and placental morphogenesis during the early stages of development can lead to an increased risk of pregnancy complications such as miscarriage, PE, PTB and fetal growth restriction (FGR) [1, 25].

Studies have shown that epigenetic modifications such as DNA methylation [27, 28] and ncRNAs [29] undergo rapid changes during placental development. Global levels of DNA methylation in the placenta increase during gestation which are associated with gene expression levels [28]. This has been observed using the Illumina Infinium HumanMethylation27 BeadChip which assesses DNA methylation at 27,000 CpG sites. The level of methylation was averaged across these sites to show that a significant increase in the level of DNA methylation from first, second and third trimester [28]. This suggests that DNA methylation is required for proper placental development.

MicroRNAs (miRNAs) (~22nt) also regulate placental development. For example, specific miRNAs have been found to be associated with enhancing or downregulating trophoblast proliferation. Trophoblast proliferation is tightly controlled throughout gestation.

Proliferation is dependent upon the expression of specific miRNAs which activate downstream targets, thereby increasing cell growth. Several miRNAs, including miR-376c [30] and miR-195 [31] enhance proliferation and trophoblast invasion during early gestation by downregulating nodal signalling and inhibiting transforming growth factor-β. Other miRNAs such as miR-675 [32], miR-155 [33] and miR-29b [34] inhibit trophoblast invasion by downregulating angiogenic factors which would otherwise increase cell proliferation. There are more than 500 known miRNAs expressed in the placenta [35], many of which are specific to pregnancy. Despite the known changes of DNA methylation and ncRNA expression during placental development only a limited number of CpG sites and ncRNAs have been analysed. Hence, further research is required to understand the full impact of DNA methylation and ncRNA expression during placental development.

2.4 Placental DNA methylation

DNA methylation occurs when a methyl group is added to a cytosine residue at the C5 position of a pyrimidine ring or to an adenine at the N6 position of a purine ring [36]. Cytosine and adenine methylation occur in bacteria as a defence mechanism [37]. However, in mammals most research investigating DNA methylation is focused on cytosine methylation [36]. DNA methylation is mediated by DNA methyltransferases (DNMTs) along with S-adenosylmethionine which is the universal methyl donor [12, 38]. Traditionally, DNA methylation is well known for its role at promoters and enhancers in regulating gene expression [39].

The placenta is unique compared to other tissues, with the exception of cancer tissues [40, 41] and a human fetal fibroblast cell line (IMR90) [42] in that it has low levels of genome wide cytosine methylation [27, 43]. Despite low levels of DNA methylation in this tissue, it has been observed that levels of DNA methylation increase during gestation [28, 44]. Furthermore, this hypomethylation is not consistent across the genome but rather occurs in large domains (> 100kb) [27]. These large domains, which are referred to as partially methylated domains (PMDs), cover approximately 40% of the genome and are specific to the placenta [27]. Genes within PMDs are typically repressed and this is maintained throughout gestation [27]. In addition, PMDs have only been investigated in placentas

from uncomplicated pregnancies. It still remains to be determined if PMDs are disrupted in pregnancy complications.

In addition, the placenta has a variety of cell types with their own distinct patterns of methylation. This has been demonstrated using DNA methylation profiles of cytotrophoblasts and fibroblasts analysed by Illumina Infinium Human Methylation 27K BeadChip Array [45]. The cytotrophoblasts and fibroblasts were found to cluster separately by hierarchical clustering which highlights distinct cell type methylation patterns in the placenta [28]. Moreover, novel ncRNAs have been found to be in trophoblast subpopulations and differentially expressed in placentas from preeclamptic pregnancies [46]. This research highlights distinct DNA methylation and gene expression profiles in different placental cell types. Future work should focus on capturing the complete DNA methylation and gene expression profiles of trophoblast subpopulations in the placenta which may uncover targets critical for placental development.

2.5 Non-coding RNA expression in the placenta

It is currently estimated that only 2% of all genomic transcripts have protein coding capability [6, 47]. The remaining genomic transcripts are classified as ncRNAs and fall into either one of two main classes depending on their length as previously described above. lncRNAs play an important role in placental development. One of the first lncRNAs to be discovered, H19, is an imprinted gene that is essential to placental development. H19 is situated on chromosome 11 and is expressed by the maternal allele. It also shares cis-regulatory elements with IGF2 which encodes an insulin-like growth factor. Mutations within the H19-IGF2 locus have implications in placental development and are associated with pregnancy complications [48]. For example, altered epigenetic regulation of H19 and IGF2 are associated with PE [49]. Expression of ncRNAs are important for proper placental development. The complete landscape of ncRNA expression in the placenta remains unknown. There are potentially thousands of undiscovered ncRNAs expressed in the placenta. Moreover, the precise mechanisms by which these ncRNAs act remains to be

determined. Further research is needed to identify potentially novel transcripts and their targets within the placenta, advancing our knowledge.

2.6 Lifelong effects of perinatal exposures in utero

In utero exposures play a critical role in perinatal and childhood development. These effects can potentially have long lasting effects on the health of the child into adulthood. Studies have identified epigenetic changes in relation to in utero exposures such as maternal smoking [50-53]. Although it is well known that smoking impacts on DNA methylation profiles [54], it is unclear what the long lasting effects are on the developing fetal epigenome. It has also been shown that if the mother ceases smoking prior to 18 weeks gestation that there is evidence in DNA methylation profiles of cord blood [50]. However, smoking is not the only environmental exposure to impact on DNA methylation. Levels of micronutrients in maternal plasma have also been shown to impact the fetal epigenome [55]. Folate, which is involved in the production of S-adenosylmethionine, is vital for fetal development. One study has shown that levels of maternal plasma folate are associated with DNA methylation in cord blood at 443 CpG sites which are related to 320 genes [55]. These genes are involved in embryonic development and birth defects such as neural tube defects. In utero exposures can impact DNA methylation profiles, although it is not clear what is the precise mechanism.

Unforeseen events, including natural disasters and famines can affect women and the developing fetus at an epigenetic level. One of the first studies to show this was “Project Ice Storm” [56], which assessed the DNA methylation profiles of blood samples from offspring at 13 years of age from women who experienced the 1998 Quebec ice storm. The researchers identified 1675 CpG sites that were associated with maternal stress [56]. Other natural disasters have been shown to impact on pregnancy complications [57]. These studies also strengthen the notion that perinatal exposures during in utero development can have lifelong epigenetic consequences.

2.7 DNA methylation as a biomarker for pregnancy outcome

Epigenetic dysregulation, including differential DNA methylation is associated with pregnancy complications [58]. Analysis of DNA methylation in relation to pregnancy outcome has been conducted in the placenta and cord blood extensively. Using microarray technology it has been shown that differential methylation occurs in placentas from PE [59-61], IUGR [62] and GDM [61, 63] in comparison to uncomplicated controls. In addition, studies have also identified the same regions to be differentially methylated in cord blood [64, 65]. These studies identify epigenetic mechanisms that are associated with pregnancy complications.

Studies have also focused on assessing differential methylation in the maternal blood of women early in pregnancy to identify women at risk of pregnancy complications [66, 67]. In maternal peripheral blood, it has been found that hypomethylation occurs in genes related to cell morphology and cell cycle in women who develop GDM [66]. This study assessed DNA methylation using the Illumina HumanMethylation 27 BeadChip at 16 weeks' gestation and identified 27 target sites that may be novel biomarkers for the identification of women at risk of developing GDM. Furthermore, differences in DNA methylation profiles of women who developed PE, GDM, PTB or SGA compared to uncomplicated controls have been identified in a preliminary study using peripheral blood collected at 15 weeks' gestation [67]. These differences in DNA methylation profiles are detectable well before the onset of any of the pregnancy complications. There have been several suggestions of using DNA methylation as a biomarker for pregnancy complications [68-70]. Furthermore, studies have also used miRNAs as biomarkers for the prediction of pregnancy outcomes in maternal blood (Table 2-1). These studies demonstrate the potential of DNA methylation and miRNAs as biomarkers in predicting pregnancy complications.

Table 2-1: Overview of studies that have identified epigenetic biomarkers of pregnancy outcomes.

Reference	Biomarker(s)	Prognosis, diagnosis, assessment
Tsochandaridis *et al.* 2015 [82]	miR-144	Prognosis of PE
Ge *et al.* 2015 [80]	miR-141-3p and miR-200c-3p	Fetal macrosomia
Anton *et al.* 2013 [81]	miR-210	Prognosis of PE
Li *et al.* 2013 [34]	miR-29a	Prognosis of PE
Zhao *et al.* 2011 [83]	miR-29a, miR-222 and miR-132	Prognosis of GDM
Anderson *et al.* 2014 [70]	207 CpG sites	Prognosis of PE

CpG: cytosine phosphate guanosine, GDM: gestational diabetes mellitus, PE: preeclampsia, miR: micro RNA.

2.8 DNA methylation as a biomarker of gestational age

Rapid development changes occur every week during gestation and these significantly impact neonatal mortality and morbidity [71]. The gold standard for measuring gestational age has typically been through ultrasound based methods [72]. Differential methylation has been observed at specific CpG sites at different gestational age time points in cord blood [73, 74]. Recently, using Illumina Infinium Human Methylation 27K and 450K BeadChip Arrays, DNA methylation has been used to predict gestational age from cord blood [75]. Gestational age can be calculated from the level of methylation at 148 CpG sites from either cord blood or blood spot samples [75]. The importance of this study highlights aging mechanisms are associated with DNA methylation.

2.9 Association of non-coding RNA expression with pregnancy complications

Recently, it has been found that differential placental expression of ncRNAs occurs in PE [76] and PTB [77] compared to that in uncomplicated pregnancies. Pathway analyses of these differentially expressed ncRNAs in PE have been shown to be related to pregnancy, lipid metabolism and the immune response [76]. In addition, pathway analyses have shown upregulated ncRNA expression in PTB to be associated with inflammation [77]. The results from these studies indicate that PE and PTB are immune and metabolic conditions that are possibly driven by ncRNAs. However, these studies have used microarray technology to assess the expression profiles, which only assess the expression of a limited number of genes.

Sequencing-based methods have the ability to assess the expression of all expressed ncRNAs within a tissue and discover novel ncRNAs. One study has used high throughput sequencing of maternal blood at 16-19 weeks' gestation, to identify 32 differentially expressed miRNAs from pregnant women who went on to have either GDM or uncomplicated pregnancies [78]. There are only a few studies that have used high throughput sequencing to identify biomarkers of pregnancy complications. However, this may change in the future due to the diminishing cost of high throughput sequencing.

Large sequencing studies and databases have been used to discover thousands of previously unannotated ncRNAs. In a recent study, using 25 independent studies of sequencing data from a variety of human tissues, it found 48,952 previously unannotated lncRNAs [79]. These novel lncRNAs are important as they can be potentially used in a broad range of applications. For example they can be used in biomarker development and studying the biological mechanisms in placental gene expression during development. Despite the size of studies such as these, little work has been conducted in the placenta. Future work should focus on ncRNA discovery in the placenta and the mechanisms by which ncRNAs regulate gene expression in placental development.

2.10 Maternal circulating non-coding RNA

Non-invasive biomarkers such as ncRNAs in maternal blood are becoming promising predictors of pregnancy complications. Studies have identified plasma miRNAs as potential biomarkers for predicting pregnancy outcomes. Expression of specific miRNAs during second trimester, such as miR-141-3p and miR-200c-3p in maternal blood have been shown to be candidate biomarkers for predicting fetal macrosomia [80]. Other miRNAs such as miR-210 have also been shown to be a biomarker for PE, which acts by inhibiting trophoblast invasion [81]. Inadequate trophoblast invasion results in reduced maternal spiral artery remodelling which can result in an increase of placental oxidative stress. miR-144 is an important regulator of hypoxia and has also been shown as a potential biomarker for PE [82]. These findings involving miR-210 and miR-144 show an important role in the pathogenesis in PE. They can potentially be used as biomarkers in the prediction of PE.

2.11 Conclusions and recommendations

Epigenetic modifications including DNA methylation are relatively stable within tissues and plasma. This makes epigenetic modifications ideal candidates for the prognosis and diagnosis for a range of diseases. Circulating ncRNAs within maternal plasma offer a non-invasive method to predict pregnancy complications. DNA methylation is also a strong candidate as a biomarker due to its stability. An epigenetic biomarker for

pregnancy complications would significantly improve the assessment and interventions for women at high risk. Targeted early interventions could reduce health burdens of pregnancy. Expression of protein coding genes varies across pregnancy and is influenced by multiple environmental factors. Specific epigenetic modifications are ideal as biomarkers as they can be used to distinguish health outcomes from varying environmental exposures.

This review, has brought to attention the current research to identify epigenetic modifications as biomarkers of development and pregnancy outcome. It has focused primarily on DNA methylation profiles and expression of ncRNAs in maternal blood and the developing placenta. Most studies have used microarray technology in their biomarker discovery. However, future research using sequencing technology has the ability to identify potential novel and pregnancy specific biomarkers. These novel biomarkers may increase the sensitivity and specificity in the prediction of pregnancy complications. Successful development and progress of use of epigenetic modifications as biomarkers have the potential to improve maternal and perinatal health.

References

1. Roberts CT: **IFPA Award in Placentology Lecture: Complicated interactions between genes and the environment in placentation, pregnancy outcome and long term health.** *Placenta* 2010, **31 Suppl:**S47-53.
2. Barua S, Junaid MA: **Lifestyle, pregnancy and epigenetic effects.** *Epigenomics* 2015, **7:**85-102.
3. Bishop KS, Ferguson LR: **The Interaction between Epigenetics, Nutrition and the Development of Cancer.** *Nutrients* 2015, **7:**922-947.
4. Weinhold B: **Epigenetics: The Science of Change.** *Environ Health Perspect* 2006, **114:**A160-167.
5. Levenson VV: **DNA methylation as a universal biomarker.** *Expert Rev Mol Diagn* 2010, **10:**481-488.
6. Shi T, Gao G, Cao Y: **Long Noncoding RNAs as Novel Biomarkers Have a Promising Future in Cancer Diagnostics.** *Dis Markers* 2016, **2016:**9085195.
7. Strimbu K, Tavel JA: **What are Biomarkers?** *Curr Opin HIV AIDS* 2010, **5:**463-466.
8. Nisenblat V, Bossuyt PM, Shaikh R, Farquhar C, Jordan V, Scheffers CS, Mol BW, Johnson N, Hull ML: **Blood biomarkers for the non-invasive diagnosis of endometriosis.** *Cochrane Database Syst Rev* 2016**:**CD012179.
9. Chahine LM, Stern MB, Chen-Plotkin A: **Blood-Based Biomarkers for Parkinson's Disease.** *Parkinsonism Relat Disord* 2014, **20:**S99-103.
10. Luo X, Huang R, Sun H, Liu Y, Bi H, Li J, Yu H, Sun J, Lin S, Cui B, Zhao Y: **Methylation of a panel of genes in peripheral blood leukocytes is associated with colorectal cancer.** *Sci Rep* 2016, **6:**29922.
11. Tarrade A, Panchenko P, Junien C, Gabory A: **Placental contribution to nutritional programming of health and diseases: epigenetics and sexual dimorphism.** *J Exp Biol* 2015, **218:**50-58.
12. Dupont C, Armant DR, Brenner CA: **Epigenetics: Definition, Mechanisms and Clinical Perspective.** *Semin Reprod Med* 2009, **27:**351-357.
13. Wu TP, Wang T, Seetin MG, Lai Y, Zhu S, Lin K, Liu Y, Byrum SD, Mackintosh SG, Zhong M, et al: **DNA methylation on N(6)-adenine in mammalian embryonic stem cells.** *Nature* 2016, **532:**329-333.
14. Denis H, Ndlovu MN, Fuks F: **Regulation of mammalian DNA methyltransferases: a route to new mechanisms.** *EMBO Reports* 2011, **12:**647-656.
15. Cao J: **The functional role of long non-coding RNAs and epigenetics.** *Biol Proced Online* 2014, **16:**11-11.
16. Wapinski O, Chang HY: **Long noncoding RNAs and human disease.** *Trends Cell Biol* 2011, **21:**354-361.

17. Brosnan CA, Voinnet O: **The long and the short of noncoding RNAs.** *Curr Opin Cell Biol* 2009, **21:**416-425.
18. Ferguson-Smith AC, Surani MA: **Imprinting and the epigenetic asymmetry between parental genomes.** *Science* 2001, **293:**1086-1089.
19. Reik W, Walter J: **Genomic imprinting: parental influence on the genome.** *Nat Rev Genet* 2001, **2:**21-32.
20. Newell-Price J, Clark AJ, King P: **DNA methylation and silencing of gene expression.** *Trends Endocrinol Metab* 2000, **11:**142-148.
21. Siegfried Z, Simon I: **DNA methylation and gene expression.** *Wiley Interdiscip Rev Syst Biol Med* 2010, **2:**362-371.
22. Holoch D, Moazed D: **RNA-mediated epigenetic regulation of gene expression.** *Nat Rev Genet* 2015, **16:**71-84.
23. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.** *Nat Genet* 2003, **33 Suppl:**245-254.
24. Burton GJ, Fowden AL: **The placenta: a multifaceted, transient organ.** *Philos Trans R Soc Lond B Biol Sci* 2015, **370:**20140066.
25. Regnault TR, Galan HL, Parker TA, Anthony RV: **Placental development in normal and compromised pregnancies-- a review.** *Placenta* 2002, **23 Suppl A:**S119-129.
26. Murray MJ, Lessey BA: **Embryo implantation and tumor metastasis: common pathways of invasion and angiogenesis.** *Semin Reprod Endocrinol* 1999, **17:**275-290.
27. Schroeder DI, Blair JD, Lott P, Yu HO, Hong D, Crary F, Ashwood P, Walker C, Korf I, Robinson WP, LaSalle JM: **The human placenta methylome.** *Proc Natl Acad Sci U S A* 2013, **110:**6037-6042.
28. Novakovic B, Yuen RK, Gordon L, Penaherrera MS, Sharkey A, Moffett A, Craig JM, Robinson WP, Saffery R: **Evidence for widespread changes in promoter methylation profile in human placenta in response to increasing gestational age and environmental/stochastic factors.** *BMC Genomics* 2011, **12:**529.
29. Li J, Zhang Y, Li D, Liu Y, Chu D, Jiang X, Hou D, Zen K, Zhang C-Y: **Small non-coding RNAs transfer through mammalian placenta and directly regulate fetal gene expression.** *Protein & Cell* 2015, **6:**391-396.
30. Fu G, Ye G, Nadeem L, Ji L, Manchanda T, Wang Y, Zhao Y, Qiao J, Wang Y-L, Lye S: **MicroRNA-376c impairs transforming growth factor-β and nodal signaling to promote trophoblast cell proliferation and invasion.** *Hypertension* 2013, **61:**864-872.
31. Bai Y, Yang W, Yang HX, Liao Q, Ye G, Fu G, Ji L, Xu P, Wang H, Li YX, et al: **Downregulated miR-195 detected in preeclamptic placenta affects trophoblast cell invasion via modulating ActRIIA expression.** *PLoS One* 2012, **7:**e38875.
32. Gao WL, Liu M, Yang Y, Yang H, Liao Q, Bai Y, Li YX, Li D, Peng C, Wang YL: **The imprinted H19 gene regulates human placental trophoblast cell**

proliferation via encoding miR-675 that targets Nodal Modulator 1 (NOMO1). *RNA Biol* 2012, **9:**1002-1010.

33. Zhang Y, Diao Z, Su L, Sun H, Li R, Cui H, Hu Y: **MicroRNA-155 contributes to preeclampsia by down-regulating CYR61.** *Am J Obstet Gynecol* 2010, **202:**466 e461-467.
34. Li P, Guo W, Du L, Zhao J, Wang Y, Liu L, Hu Y, Hou Y: **microRNA-29b contributes to pre-eclampsia through its effects on apoptosis, invasion and angiogenesis of trophoblast cells.** *Clin Sci (Lond)* 2013, **124:**27-40.
35. Morales-Prieto D, Ospina-Prieto S, Schmidt A, Chaiwangyen W, Markert U: **Elsevier Trophoblast Research Award Lecture: origin, evolution and future of placenta miRNAs.** *Placenta* 2014, **35:**S39-S45.
36. von Kanel T, Huber AR: **DNA methylation analysis.** *Swiss Med Wkly* 2013, **143:**w13799.
37. Casadesús J, Low D: **Epigenetic Gene Regulation in the Bacterial World.** *Microbiol Mol Biol Rev* 2006, **70:**830-856.
38. Furness DL, Fenech MF, Khong YT, Romero R, Dekker GA: **One-carbon metabolism enzyme polymorphisms and uteroplacental insufficiency.** *Am J Obstet Gynecol* 2008, **199:**276.e271-278.
39. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al: **Conserved role of intragenic DNA methylation in regulating alternative promoters.** *Nature* 2010, **466:**253-257.
40. Hon GC, Hawkins RD, Caballero OL, Lo C, Lister R, Pelizzola M, Valsesia A, Ye Z, Kuan S, Edsall LE, et al: **Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer.** *Genome Res* 2012, **22:**246-258.
41. Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CP, van Dijk CM, Tollenaar RA, et al: **Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains.** *Nat Genet* 2012, **44:**40-46.
42. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, et al: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462:**315-322.
43. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, et al: **Human body epigenome maps reveal noncanonical DNA methylation variation.** *Nature* 2015, **523:**212-216.
44. Price EM, Cotton AM, Penaherrera MS, McFadden DE, Kobor MS, Robinson W: **Different measures of "genome-wide" DNA methylation exhibit unique properties in placental and somatic tissues.** *Epigenetics* 2012, **7:**652-663.
45. Grigoriu A, Ferreira JC, Choufani S, Baczyk D, Kingdom J, Weksberg R: **Cell specific patterns of methylation in the human placenta.** *Epigenetics* 2011, **6:**368-379.

46. Gormley M, Ona K, Kapidzic M, Garrido-Gomez T, Zdravkovic T, Fisher SJ: **Preeclampsia: novel insights from global RNA profiling of trophoblast subpopulations.** *Am J Obstet Gynecol* 2017.
47. Ponting CP, Belgard TG: **Transcribed dark matter: meaning or myth?** *Hum Mol Genet* 2010, **19:**R162-168.
48. Tabano S, Colapietro P, Cetin I, Grati FR, Zanutto S, Mando C, Antonazzo P, Pileri P, Rossella F, Larizza L, et al: **Epigenetic modulation of the IGF2/H19 imprinted domain in human embryonic and extra-embryonic compartments and its possible role in fetal growth restriction.** *Epigenetics* 2010, **5:**313-324.
49. Yu L, Chen M, Zhao D, Yi P, Lu L, Han J, Zheng X, Zhou Y, Li L: **The H19 gene imprinting in normal pregnancy and pre-eclampsia.** *Placenta* 2009, **30:**443-447.
50. Joubert BR, Haberg SE, Bell DA, Nilsen RM, Vollset SE, Midttun O, Ueland PM, Wu MC, Nystad W, Peddada SD, London SJ: **Maternal smoking and DNA methylation in newborns: in utero effect or epigenetic inheritance?** *Cancer Epidemiol Biomarkers Prev* 2014, **23:**1007-1017.
51. Ungerer M, Knezovich J, Ramsay M: **In Utero Alcohol Exposure, Epigenetic Changes, and Their Consequences.** *Alcohol Research : Current Reviews* 2013, **35:**37-46.
52. Perera F, Herbstman J: **Prenatal environmental exposures, epigenetics, and disease.** *Reproductive toxicology (Elmsford, NY)* 2011, **31:**363-373.
53. Lee H-S: **Impact of Maternal Diet on the Epigenome during In Utero Life and the Developmental Programming of Diseases in Childhood and Adulthood.** *Nutrients* 2015, **7:**9492-9507.
54. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, Barrdahl M, Boeing H, Aleksandrova K, Trichopoulou A, et al: **Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study.** *Epigenomics* 2016, **8:**599-618.
55. Joubert BR, den Dekker HT, Felix JF, Bohlin J, Ligthart S, Beckett E, Tiemeier H, van Meurs JB, Uitterlinden AG, Hofman A, et al: **Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns.** *Nature Communications* 2016, **7:**10577.
56. Cao-Lei L, Massart R, Suderman MJ, Machnes Z, Elgbeili G, Laplante DP, Szyf M, King S: **DNA methylation signatures triggered by prenatal maternal stress exposure to a natural disaster: Project Ice Storm.** *PLoS One* 2014, **9:**e107653.
57. Antipova A, Curtis A: **The post-disaster negative health legacy: pregnancy outcomes in Louisiana after Hurricane Andrew.** *Disasters* 2015, **39:**665-686.
58. Parets SE, Bedient CE, Menon R, Smith AK: **Preterm birth and its long-term effects: methylation to mechanisms.** *Biology (Basel)* 2014, **3:**498-513.
59. Anton L, Brown AG, Bartolomei MS, Elovitz MA: **Differential methylation of genes associated with cell adhesion in preeclamptic placentas.** *PLoS One* 2014, **9:**e100148.

60. Chu T, Bunce K, Shaw P, Shridhar V, Althouse A, Hubel C, Peters D: **Comprehensive analysis of preeclampsia-associated DNA methylation in the placenta.** *PLoS One* 2014, **9:**e107318.
61. Liu L, Zhang X, Rong C, Rui C, Ji H, Qian YJ, Jia R, Sun L: **Distinct DNA methylomes of human placentas between pre-eclampsia and gestational diabetes mellitus.** *Cell Physiol Biochem* 2014, **34:**1877-1889.
62. Hillman SL, Finer S, Smart MC, Mathews C, Lowe R, Rakyan VK, Hitman GA, Williams DJ: **Novel DNA methylation profiles associated with key gene regulation and transcription pathways in blood and placenta of growth-restricted neonates.** *Epigenetics* 2015, **10:**50-61.
63. Petropoulos S, Guillemin C, Ergaz Z, Dimov S, Suderman M, Weinstein-Fudim L, Ornoy A, Szyf M: **Gestational Diabetes Alters Offspring DNA Methylation Profiles in Human and Rat: Identification of Key Pathways involved in Endocrine System Disorders, Insulin Signaling, Diabetes Signaling and IL-K Signaling.** *Endocrinology* 2014**:**en20141643.
64. Finer S, Mathews C, Lowe R, Smart M, Hillman S, Foo L, Sinha A, Williams D, Rakyan VK, Hitman GA: **Maternal gestational diabetes is associated with genome-wide DNA methylation variation in placenta and cord blood of exposed offspring.** *Hum Mol Genet* 2015.
65. Ruchat SM, Houde AA, Voisin G, St-Pierre J, Perron P, Baillargeon JP, Gaudet D, Hivert MF, Brisson D, Bouchard L: **Gestational diabetes mellitus epigenetically affects genes predominantly involved in metabolic diseases.** *Epigenetics* 2013, **8:**935-943.
66. Enquobahrie DA, Moore A, Muhie S, Tadesse MG, Lin S, Williams MA: **Early Pregnancy Maternal Blood DNA Methylation in Repeat Pregnancies and Change in Gestational Diabetes Mellitus Status-A Pilot Study.** *Reprod Sci* 2015, **22:**904-910.
67. Bianco-Miotto T, Lopez CR, Leemaqz S, Buckberry S, McCullough D, Zhuang Z, Dekker G, Wilkinson M, Roberts C: **DNA methylation biomarkers for predicting pregnancy complications.** *Placenta* 2015, **36:**A38-A39.
68. Menon R, Conneely KN, Smith AK: **DNA methylation: an epigenetic risk factor in preterm birth.** *Reprod Sci* 2012, **19:**6-13.
69. Tendl KA, Schulz SM, Mechtler TP, Bohn A, Metz T, Greber-Platzer S, Kasper DC, Herkner KR, Item CB: **DNA methylation pattern of CALCA in preterm neonates with bacterial sepsis as a putative epigenetic biomarker.** *Epigenetics* 2013, **8:**1261-1267.
70. Anderson CM, Ralph JL, Wright ML, Linggi B, Ohm JE: **DNA methylation as a biomarker for preeclampsia.** *Biol Res Nurs* 2014, **16:**409-420.
71. Engle WA: **Morbidity and mortality in late preterm and early term newborns: a continuum.** *Clin Perinatol* 2011, **38**.
72. Lynch CD, Zhang J: **The research implications of the selection of a gestational age estimation method.** *Paediatr Perinat Epidemiol* 2007, **21**.

73. Parets SE, Conneely KN, Kilaru V, Fortunato SJ, Syed TA, Saade G, Smith AK, Menon R: **Fetal DNA methylation associates with early spontaneous preterm birth and gestational age.** *PLoS One* 2013, **8**.
74. Schroeder JW, Conneely KN, Cubells JC, Kilaru V, Newport DJ, Knight BT, Stowe ZN, Brennan PA, Krushkal J, Tylavsky FA: **Neonatal DNA methylation patterns associate with gestational age.** *Epigenetics* 2011, **6**.
75. Knight AK, Craig JM, Theda C, Bækvad-Hansen M, Bybjerg-Grauholm J, Hansen CS, Hollegaard MV, Hougaard DM, Mortensen PB, Weinsheimer SM, et al: **An epigenetic clock for gestational age at birth based on blood methylation data.** *Genome Biol* 2016, **17:**206.
76. He X, He Y, Xi B, Zheng J, Zeng X, Cai Q, Ouyang Y, Wang C, Zhou X, Huang H, et al: **LncRNAs expression in preeclampsia placenta reveals the potential role of LncRNAs contributing to preeclampsia pathogenesis.** *PLoS One* 2013, **8:**e81437.
77. Luo X, Pan J, Wang L, Wang P, Zhang M, Liu M, Dong Z, Meng Q, Tao X, Zhao X, et al: **Epigenetic regulation of lncRNA connects ubiquitin-proteasome system with infection-inflammation in preterm births and preterm premature rupture of membranes.** *BMC Pregnancy Childbirth* 2015, **15:**35.
78. Zhu Y, Tian F, Li H, Zhou Y, Lu J, Ge Q: **Profiling maternal plasma microRNA expression in early pregnancy to predict gestational diabetes mellitus.** *Int J Gynaecol Obstet* 2015, **130:**49-53.
79. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al: **The landscape of long noncoding RNAs in the human transcriptome.** *Nat Genet* 2015, **47:**199-208.
80. Ge Q, Zhu Y, Li H, Tian FEI, Xie X, Bai Y: **Differential expression of circulating miRNAs in maternal plasma in pregnancies with fetal macrosomia.** *Int J Mol Med* 2015, **35:**81-91.
81. Anton L, Olarerin-George AO, Schwartz N, Srinivas S, Bastek J, Hogenesch JB, Elovitz MA: **miR-210 inhibits trophoblast invasion and is a serum biomarker for preeclampsia.** *The American journal of pathology* 2013, **183:**1437-1445.
82. Tsochandaridis M, Nasca L, Toga C, Levy-Mozziconacci A: **Circulating microRNAs as clinical biomarkers in the predictions of pregnancy complications.** *Biomed Res Int* 2015, **2015:**294954.
83. Zhao C, Dong J, Jiang T, Shi Z, Yu B, Zhu Y, Chen D, Xu J, Huo R, Dai J, et al: **Early second-trimester serum miRNA profiling predicts gestational diabetes mellitus.** *PLoS One* 2011, **6:**e23925.

Statement of Authorship

Title of Paper	Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Mayne BT, Bianco-Miotto T, Buckberry S, Breen J, Clifton V, Shoubridge C, Roberts CT: Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans. *Frontiers in Genetics* 2016, 7.

Principal Author

Name of Principal Author (Candidate)	Benjamin Mayne
Contribution to the Paper	Designed, conducted the study, analysed and interpreted the data and wrote the manuscript.
Overall percentage (%)	85
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 14-7-17

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Tina Bianco-Miotto
Contribution to the Paper	Involved in the study design, provided critical discussion and intellectual input into the manuscript.
Signature	Date 26/7/17

Name of Co-Author	Sam Buckberry
Contribution to the Paper	Conceived the initial part of the study and provided intellectual input into the manuscript.
Signature	Date 16 July 2017

Name of Co-Author	James Breen
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.
Signature	Date 20/7/2017

Name of Co-Author	Vicki Clifton
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.
Signature	Date 25/7/17

Name of Co-Author	Cheryl Shoubridge
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.
Signature	Date 25-7-17.

Name of Co-Author	Claire T Roberts
Contribution to the Paper	Involved in the study design, provided critical discussion and intellectual input into the manuscript.
Signature	Date 14.7.17

3 Large scale gene expression meta-analysis reveals tissue-specific, sex-biased gene expression in humans

BENJAMIN T MAYNE, TINA BIANCO-MIOTTO, SAM BUCKBERRY, JAMES BREEN, VICKI CLIFTON, CHERYL SHOUBRIDGE and CLAIRE T ROBERTS

Abstract

The severity and prevalence of many diseases are known to differ between the sexes. Organ specific sex-biased gene expression may underpin these and other sexually dimorphic traits. To further our understanding of sex differences in transcriptional regulation, we performed meta-analyses of sex biased gene expression in multiple human tissues. We analysed 22 publicly available human gene expression microarray data sets including over 2500 samples from 15 different tissues and 9 different organs. Briefly, by using an inverse-variance method we determined the effect size difference of gene expression between males and females. We found the greatest sex differences in gene expression in the brain, specifically in the anterior cingulate cortex, (1818 genes), followed by the heart (375 genes), kidney (224 genes), colon (218 genes) and thyroid (163 genes). More interestingly, we found different parts of the brain with varying numbers and identity of sex-biased genes, indicating that specific cortical regions may influence sexually dimorphic traits. The majority of sex-biased genes in other tissues such as the bladder, liver, lungs and pancreas were on the sex chromosomes or involved in sex hormone production. On average in each tissue, 32% of autosomal genes that were expressed in a sex-biased fashion contained androgen or estrogen hormone response elements. Interestingly, across all tissues, we found approximately two-thirds of autosomal genes that were sex-biased were not under direct influence of sex hormones. To our knowledge this is the largest analysis of sex-biased gene expression in human tissues to date. We identified many sex-biased genes that were not under the direct influence of sex chromosome genes or sex hormones. These may provide targets for future development of sex-specific treatments for diseases.

3.1 Introduction

Differences in both disease severity, prevalence, symptoms and age of onset vary greatly between males and females [1]. For example, cardiovascular disease is one of the leading causes of death, affecting up to 55% of females but only 44% of males in Europe [2]. Sex differences are also evident in the risk factors for cardiovascular disease, such as diabetes which increases the risk for cardiovascular disease 2-3 fold in males but 3-7 fold in females [3]. Sex differences have also been identified in the age of onset of brain diseases such as schizophrenia, where males develop symptoms between 18-25 years of age whereas females develop symptoms between 25-35 years [4]. Moreover, reported atonic seizures in epilepsy are more frequent in males compared to females (6.5% vs 1.7%) [5]. These sex differences in diseases may be the result of tissue-specific differential gene expression between males and females. In schizophrenia, genes relating to energy metabolism have been found to have altered expression in the prefrontal cortex of only males [6]. Therefore, gene expression may have a role in orchestrating sex differences in the prevalence of diseases.

Many studies neglect to account for sample sex in the design and analysis of their experiments [7, 8]. Historically, females have been excluded from biomedical studies, due to the assumption that their hormonal cycles are a confounding factor in experimental manipulations [7, 9]. Despite females and males sharing highly similar genomes, there are numerous sex-specific traits in phenotype, physiology and pathology. Sexually dimorphic traits can be influenced by sex chromosome genes or sex hormones, but may extend beyond these influences. Sex differences may arise through alterations in autosomal gene regulation but the true extent of sex specific differential gene regulation is not fully known. Understanding these differences may dictate that future research should consider sex as a biological confounder [9]. Sex differences in many traits are often small and require large sample sizes for studies to be sufficiently powered. The substantial increase in the number of large publicly available genomic data sets could

assist in determining the true extent of sex-biased gene expression but to date there are no large-scale meta-analyses investigating this in adult human tissues.

Previous studies have reported sex-biased gene expression in the human brain [10-14], pancreas [15], heart [16] and liver [17]. Most studies identify sex-biased genes as those located on the sex chromosomes and it is well known that these are a source of differentially expressed genes between the sexes [18]. In mammalian, female, somatic cells, one X chromosome is randomly inactivated by a process referred to as X chromosome inactivation (XCI) [18, 19]. In normal human XX females, up to 15% of genes on the X chromosome escape XCI, unlike the case in mice where very few escape inactivation [18, 19]. Escape from XCI results in a number of genes that are expressed more highly in females compared to males. In addition, autosomal genes have also been shown to be sex-biased in human tissues including the brain [10], heart [16] and placenta [20]. Furthermore, sex differences in the brain in diseases such as multiple sclerosis (MS) are related to autosomal genes and are not regulated by sex chromosome genes [21, 22]. These studies highlight the importance of investigating sex differences outside the context of reproductive and sex chromosome factors. In order to characterize the true extent of sex-biased gene expression in humans, we performed a large meta-analysis of publicly available microarray data. We limited our analysis to tissue samples from healthy individuals, reducing the possible effect that diseases may have on gene expression. Our analysis revealed consistencies in sex differences that are widespread in a range of human tissues. Furthermore, we have identified sex-biased genes that are disease-related, suggesting possible mechanisms for the associations of sex with an increased risk of certain diseases.

3.2 Materials and methods

3.2.1 Data collection

Data sets were from different microarray platforms and therefore pre-processing was tailored to each platform. Briefly, data from Illumina platforms were pre-processed using Beadarray prior to quantile normalisation [23]. Data from Affymetrix platforms were pre-processed and quantile normalised using the robust multiarray average (RMA) or

GeneChip-RMA (GC-RMA) where appropriate that is implemented in Simpleaffy [24]. Batch effects in data sets were corrected for using the 'combat' function in the SVA package [25]. Outliers were identified and removed using ArrayQualityMetrics by analysing MA plots [26].

3.3 Sample sex identification

To identify sample sex in each data set we used the massiR Bioconductor package [27]. This R package uses unsupervised clustering of probes that target Y chromosome genes to identify sample sex. In data sets where sample sex was supplied, we found an agreement in all predicted and supplied sample sex identification.

3.3.1 Differential gene expression analysis

Probes were re-annotated to Ensembl gene identifiers using biomaRt [28]. In tissues where only one data set was found to be useable, sex-biased gene expression was determined using the Empirical Bayes methods within limma [29]. For tissues that were present in >1 data set, differential gene expression analysis was performed using the metaGEM package (https://spiral.imperial.ac.uk/handle/10044/1/4217) and using the inverse–variance method as previously described [30]. For each probe, study specific effect sizes were calculated, by determining the mean and standard deviation for each probe which was corrected using Hedges' g (accounts for the number of samples in each dataset). Z statistics were calculated for each gene identifier which was used to calculate a nominal p-value to give a corrected p-value (false discovery rate, FDR).

3.3.2 Androgen and estrogen response elements

To determine which genes contained androgen response elements (AREs), we firstly downloaded the coordinates of AREs from JASPAR [31, 32] and determined the positions within the genome in relation to genes and genomic locations. This was performed using the matchGenes function in the bumphunter Bioconductor package [33]

and UCSC hg19 annotation package [34]. For estrogen response elements (EREs) we used a previous study that lists genes that are targets of ERα [35].

3.3.3 Identifying enriched transcription factors

Transcription factor binding sites within 10kb upstream/downstream of sex-biased genes were analysed using oPOSSUM-3 and the JASPAR vertebrate core profiles [31, 36]. We chose 10kb upstream/downstream of genes as this was the largest range the oPOSSUM-3 would allow. Thus we sought to identify all possible transcription factor (TF) binding sites enriched within sex-biased genes. For each sex-biased gene in each tissue, the TF binding site motifs were searched with a conservation cut-off of 0.4, an 85% threshold for the matrix score and minimum specificity of 8 bits. The resulting TF analysis was limited to the most enriched TFs which were defined as those with the highest Fisher's exact test and z score rankings.

3.3.4 Gene ontology

Gene ontology (GO) analysis was performed using all human genes in the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 [37] and g:Profiler [38]. GO terms were considered significant if the corrected p-value (FDR) < 0.05.

A more detailed account of the methodology is provided in Supplementary File 3-1.

3.4 Results and Discussion

3.4.1 Overview of publicly available microarray data

Using the Gene Expression Omnibus (GEO) [39] and ArrayExpress [40] we identified 22 microarray data sets containing a total of 2502 samples, in 15 different human tissues (Table 3-1). We excluded pooled samples and limited our analyses to data sets with >10 samples to allow better determination of sample sex. To increase the number of useable data sets we used massiR [27] to identify and to verify the sample sex in all data sets. From the 22 chosen studies, 10 had sample sex metadata and within these we found

concordance with all the predicted and supplied sample sex information. Female samples (N = 803) made up 32% of all samples across all data sets (Table 3-1).

Table 3-1: Gene expression data involving 15 healthy tissues.

Each row corresponds to a data set where only healthy tissue was used within this analysis. The columns report the Microarray manufacturer, total number of samples and w

Organ/Tissue	GEO Accession	Microarray manufacturer	Samples in data set	Control Samples	Sample after pre-processing	Males	Females
Bladder	GSE13507	Affymetrix	256	68	68	48	20
Brain	GSE45642	Affymetrix	670	670	659	493	166
Brain	GSE11512	Affymetrix	80	44	44	29	15
Brain	GSE54572	Affymetrix	24	12	12	5	7
Brain	GSE36192	Illumina	911	911	911	622	289
Brain	GSE44456	Affymetrix	39	39	39	28	11
Colon	GSE8671	Affymetrix	62	25	23	15	8
Colon	GSE41328	Affymetrix	20	10	10	8	2
Heart	GSE55231	Illumina	129	129	118	69	49
Heart	GSE26887	Affymetrix	24	24	23	19	4
Heart	GSE57338	Affymetrix	313	136	136	97	39
Kidney	GSE43974	Illumina	554	118	118	73	45
Kidney	GSE50892	Affymetrix	17	17	15	9	6
Liver	GSE61276	Illumina	106	50	48	22	26
Liver	GSE23649	Illumina	69	69	68	42	26
Liver	GSE38941	Affymetrix	27	10	10	4	6
Lung	GSE10072	Affymetrix	107	49	46	32	14
Lung	GSE18995	Affymetrix	35	35	34	15	19
Lung	GSE51024	Affymetrix	96	41	39	34	5
Pancreas	GSE15471	Affymetrix	78	36	35	19	16
Thyroid	GSE33630	Affymetrix	105	45	35	10	25
Thyroid	GSE65144	Affymetrix	25	13	12	7	5
		Total	3747	2551	2502	1699	803

ed sample sex.

Sex differences in autosomal gene expression are typically small so in order to increase statistical robustness, we performed multiple testing corrections in three different analyses for each tissue. We determined the adjusted p-value implemented by Benjamini & Hochberg [41] for each autosomal gene where 1) all the chromosomes were included, 2) the Y chromosome was excluded, and 3) both the X and Y chromosomes were excluded in the analysis (Table 3-2). In general, we observed a reduction in the number of autosomal genes that were significantly sex-biased when we removed sex chromosomes from the analysis. Since most genes located on the sex chromosomes had the smallest adjusted p-value, their removal from the analysis slightly increased the adjusted p-value for all other genes. Here we supply the adjusted p-values for all three analyses (Supplementary Table 3-1, 3-2, and 3-3) but discuss only autosomal genes that were significantly different in all three cases. Furthermore, the sample size in each tissue was not reflective of the total number of genes differentially expressed between males and females (Figure 3-1). For example, despite the frontal lobe of the cerebral cortex or frontal cortex (FC) and cerebellum (CB) data sets containing the greatest number of samples, with 455 and 553 samples, respectively, we detected only a small number of sex-biased genes compared to other tissues such as the anterior cingulate cortex (AnCg) and the heart which contained the greatest number of sex-biased genes with average sample sizes (Figure 3-1 and Table 3-2).

Table 3-2: Total number of sex-biased genes in each tissue.

Each column corresponds to the total number of genes that were differentially expressed between males and females in each analysis.

Organ/Tissue	No. of Sex-biased Genes (All Chromosomes)	No. of autosomal sex-biased genes (Sex chromosomes included in analysis)	No. of autosomal sex-biased genes (Sex chromosomes removed)	No. of autosomal sex-biased genes (Y chromosome removed)
Bladder	16	0	0	0
Brain (Nucleus Accumbens)	264	239	216	244
Brain (Amygdala)	17	6	0	0
Brain (Cerebellum)	98	59	45	52
Brain (Anterior Cingulate Cortex)	1818	1726	1690	1728
Brain (Dorsolateral Prefrontal Cortex)	198	180	165	169
Brain (Frontal Cortex)	45	10	27	7
Brain (Hippocampus)	205	183	174	180
Colon	218	199	162	190
Heart	375	348	334	346
Kidney	224	196	194	194
Liver	32	21	16	28
Lung	36	14	2	12
Pancreas	22	0	0	0
Thyroid	163	151	133	135

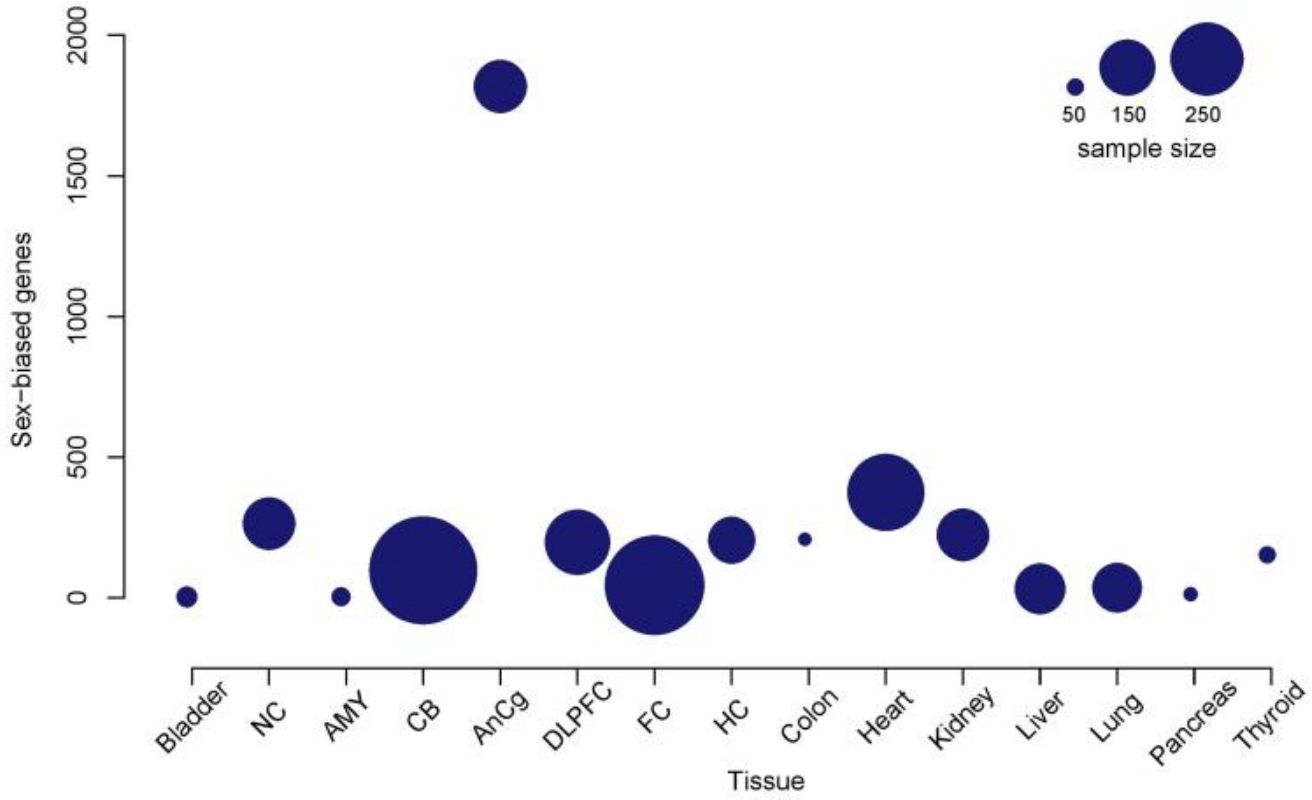


Figure 3-1: Total number of detectable sex-biased genes relative to the sample size in each tissue.

A bubble plot of each tissue where the size of the bubble is proportional to the sample size of the tissue. Bubbles that are higher on the y-axis are tissues that demonstrate a higher number of detectable sex-biased genes. Nucleus accumbens (NC); amygdala (AMY); cerebellum (CB); anterior cingulate cortex (AnCg); dorsolateral frontal cortex (DLPFC); frontal cortex (FC); hippocampus (HC).

3.4.2 Sex-biased gene expression in the human brain

Previous studies have found sex-biased gene expression in the human brain [11-14]. We identified 5 data sets for 7 brain regions and our analyses showed that each region had different numbers of differentially expressed genes (Table 3-1 and 3-2). Our findings were consistent with previous studies [11, 13, 14], whereby the most striking differences in gene expression between the sexes were sex chromosome genes. These comprised most of the sex-biased genes in the amygdala (65%) (AMY) and FC (78%). However, a large proportion of sex-biased genes were autosomal in the nucleus accumbens (91%) (NC), AnCg (95%), dorsolateral prefrontal cortex (91%) (DLPFC), CB (60%) and the hippocampus (89%) (HC). Of the 1690 autosomal sex-biased genes in AnCg, 65% were expressed more highly in males (Figure 3-2A, Supplementary Table 3-1, 3-2 and 3-3). Conversely, we observed a greater proportion of autosomal genes expressed more highly in females in the NC (75%), DLPFC (68%) and the HC (62%). We also found that each brain region was unique in its proportion of sex-biased genes, with as many sex-biased genes in one brain region that were not sex-biased in another (Figure 3-2B).

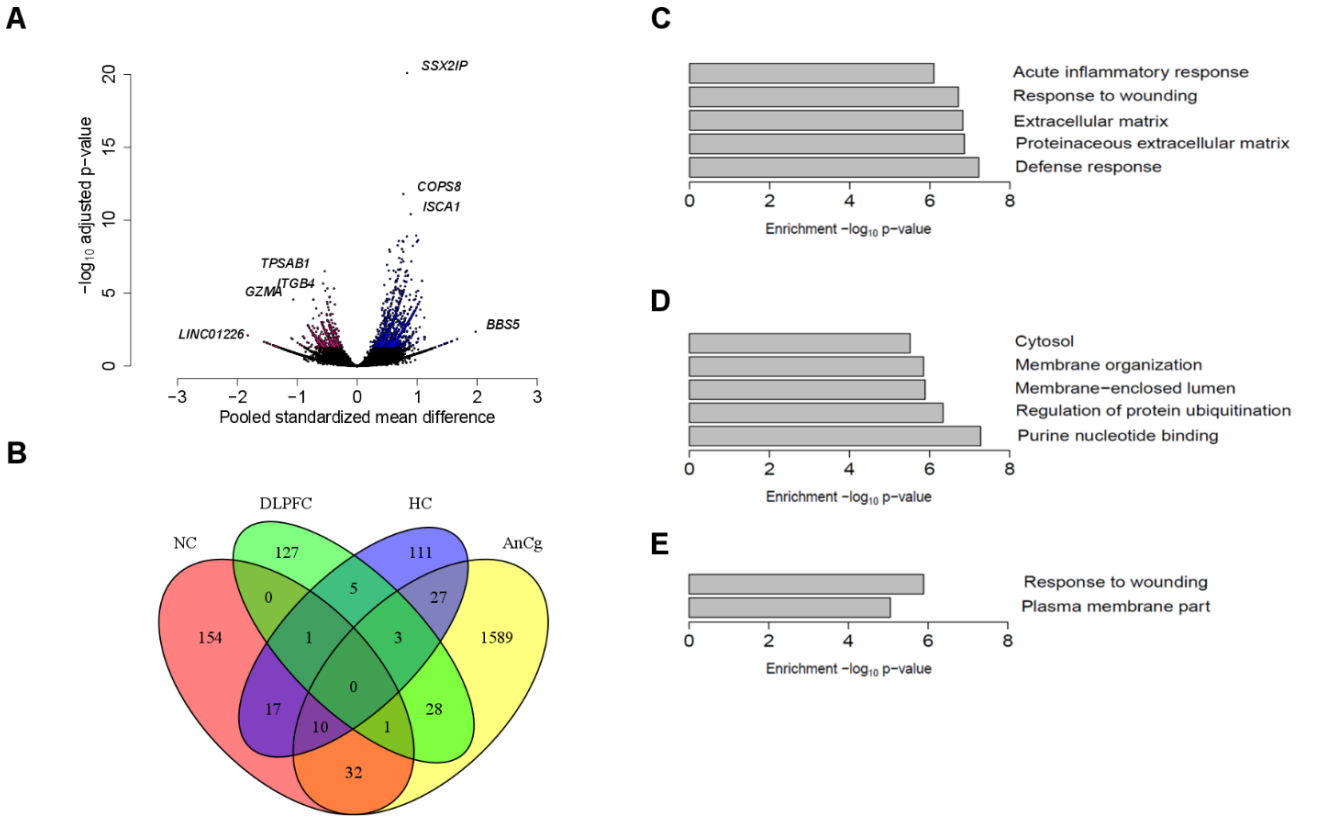


Figure 3-2: Sex differences in autosomal gene expression in the human brain.

(**A**) A volcano plot representing the autosomal genes that were sex-biased in the AnCg. Pink coloured dots represent genes that were significantly expressed more highly in females and blue coloured dots represent genes that were expressed more highly in males. (**B**) A four-way venn diagram showing the overlap of sex-biased autosomal gene expression in different regions of the human brain. Most genes that were found to be sex-biased in one region were not sex-biased in another region. The top GO terms that were enriched for sex-biased genes in (**C**) Nucleus accumbens (NC), (**D**) anterior cingulate cortex (AnCg) and (**E**) hippocampus (HC).

An increase in the expression of heat shock proteins (HSPs) has been shown to have protective roles in pro-inflammatory responses [42]. Consistent with a previous study [43], we found genes that encode for HSPs to have sex-biased expression in the human brain. Our analyses also identified genes that are involved in pro-inflammatory responses, such as those encoding interleukins, that are more highly expressed in females in NC, AnCg, DLPFC and HC tissues (Supplementary Table 3-1, 3-2 and 3-3). By contrast, genes expressed more highly in males within the brain were related to energy production and growth, including ATPase's and insulin-like growth factors in the HC and NC, respectively, and GAPDH in the AnCg. We found sex-biased genes in the NC, AnCg and HC to be enriched for Gene Ontology (GO) as defined by DAVID v6.7 for terms relating to cellular functions, the immune response and energy production (Figure 3-2C, 2D, 2E, Supplementary Table 3-4). We also used g:Profiler [38] for a comparison of GO terms and found similar results to what was found by DAVID v6.7. For example, in the NC, AnCg and HC we found that the gene upregulated in females were enriched for those involved in the immune response (GO:0006955). Whereas, genes upregulated in males were found to be enriched for GO terms such as generation of precursor metabolites and energy (GO:0006091). Overall, varying proportions and types of sex-biased genes were identified within different locations of the brain, suggesting that specific cortical regions may influence sexually dimorphic traits. As mentioned above, the AnCg contained the largest number of genes differentially expressed between males and females. The AnCg is one of the most recently evolved parts of the mammalian brain [44] and also has been shown to regulate behaviour and act in a sex-specific manner [45]. Furthermore, previous studies have identified sex differences in mood disorders and the AnCg is known to have a role in regulating mood [46, 47]. In mice, the AnCg has also been shown to have a critical role in sexual interest of males for females [48] and hence the large number of genes that were differentially expressed between sexes in the AnCg may assist in the explanation for sexual dimorphism in behaviour.

Sex biased gene expression in the brain may potentially contribute to differences in certain neurological diseases between sexes, such as the previously mentioned epilepsy.

Sex differences in gene expression may mediate these differences in susceptibility or comprise part of the mechanistic pathways involved in their pathology. Previously, sex biased gene expression in the brain has been proposed to underlie the sex differences in schizophrenia [10] which has an incidence of 1.4:1 between males and females [49]. We found several genes that have been associated with brain disorders to be sex-biased within specific locations of the brain. For example in the AnCg, NOTCH3, a gene associated with hereditary stroke disorder [50], and ALDH3B1, a gene associated with schizophrenia [51], were more highly expressed in females than males. On the other hand, KCNH3, a gene associated with epilepsy [52], GABRB3, a gene associated with schizophrenia [53], epilepsy [54] and autism [55], SNCA, a gene associated with Parkinson's disease [56], and RGS4, a gene associated with schizophrenia [57], were all expressed more highly in males. Recently, sex-biased gene expression has also been identified during developmental stages of the human brain [58]. Furthermore, genes associated with schizophrenia have been found to be upregulated in male brains as opposed to females across different developmental stages [58]. This demonstrates consistency in sex-biased genes within the human brain across different studies. Taken together, these findings suggest possible mechanisms by which sex-specific prevalence of brain disorders may occur.

3.4.3 The heart and kidney show opposite trends in sex differences in gene expression

Most of the heart gene expression data used in this study are from individuals with an average age of 47 years and we observed many sex differences in expression of genes associated with heart disease. It has been reported in elderly individuals (> 75 years), isolated systolic hypertension can be up to 14% more prevalent in females than males [59]. We found SCN10A, a gene associated with hypertrophic cardiomyopathy [60], and KCNE1, a gene associated with long-QT syndrome [61], to be expressed more highly in hearts from females. Interestingly, 62% of the 334 autosomal sex-biased genes in the heart were expressed more highly in females. The distribution of sex-biased genes across all chromosomes in the heart was similar to that in a previous study [16]. However, we

report a much smaller number of sex-biased genes in the heart (375 genes in 277 samples (Table 3-2, Supplementary Table 3-1) compared to 1800 genes in 102 samples in that study [16]).

Conversely, compared to the heart, we found an opposite trend in the kidneys, with 72% of a total of 194 autosomal genes being expressed more highly in males. We also identified 6 genes located on chromosome 1 that were expressed more highly in females in the heart that were more abundantly expressed in males in the kidney (Figure 3-3). These genes are from the RNA U1 family (RNU1-1, RNU1-2, RNU1-3, RNU1-4, RNVU1-7 and RNU1-18) that includes genes that regulate transcription, elongation and pre-mRNA splicing events [62, 63]. It has been suggested that the expression of these genes is different between tissues to regulate organ specific alternative splicing events [63]. Sex differences in alternative splicing have also previously been detected in the brain, where it has been found to affect 2.5% of expressed genes [10]. Apart from RNA U1 family all other sex-biased genes were only found to be expressed more highly in one sex.

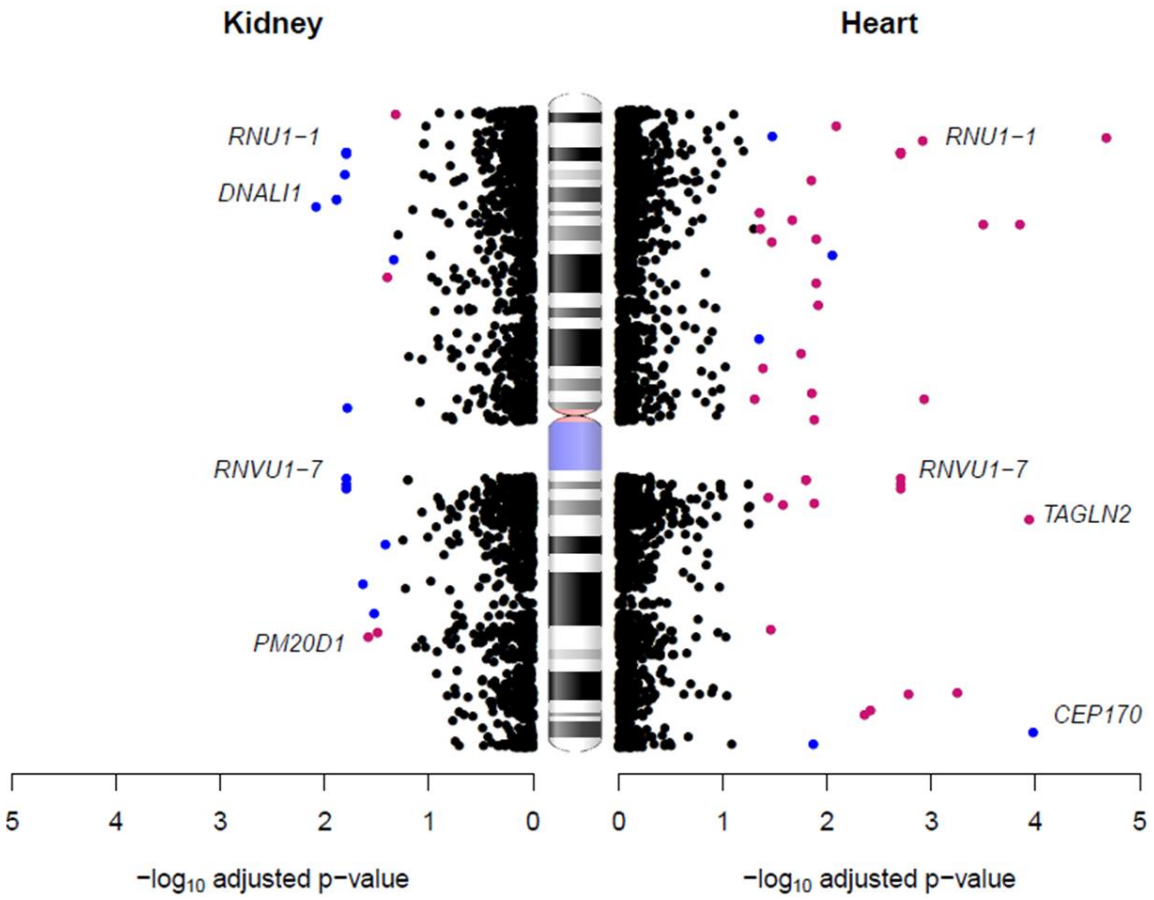


Figure 3-3: Sex-biased gene expression differences on chromosome 1 in the heart and kidney.

Each dot represents a gene, blue dots are genes that were expressed more highly in males and pink dots are those expressed more highly in females. The ideogram of chromosome 1 was obtained from the National Centre for Biotechnology Information (NCBI).

3.4.4 Sex hormones and gene expression

Many of the sex-biased genes we identified encode enzymes that are known to regulate the production of sex hormones. In the AnCg, three genes from the sulfotransferase family that regulates sulphate conjugation in estrogen precursors [64] (SULT2A1, SULT1B1 and SULT1C1) were expressed more highly in females. In addition, we also found STS (a gene involved in the production of estrogen precursors [65]) to be expressed more highly in females in the FC and CB, as well as in the heart and lung. We did not find any major sex differences in gene expression in the bladder, liver, lung or pancreas, apart from genes located on the sex chromosomes and those that are involved in sex hormone production. This can be contradictory to that which has been found in mouse studies where thousands of genes have been found to be sex-biased [66, 67]. This may reflect an evolutionary difference between the species. Apart from the brain, we found the largest number of sex-biased gene expression differences in the heart, kidney, colon and thyroid (Table 3-2). Thyroid hormones are known to regulate sex hormone-binding globulin (SHBG) production, which transports androgens and estrogens through the bloodstream [68]. In the thyroid, 133 autosomal genes were sex-biased, 75% of which were expressed more highly in males. Genes that encode for growth factors and signalling molecules were highly expressed in the thyroid of males, such as CCL28, a growth factor in hematopoietic stem cells [69], CMTM4, a chemokine that regulates the cell cycle [70], and GH1, a gene that encodes for growth hormone [71]. These findings suggest a functional role for the thyroid in influencing sexually dimorphic traits such as metabolism, as well as sex differences in thyroid hormone secretion [72]. There is also evidence to suggest that thyroid hormones significantly influence testosterone levels [73].

To determine if the differentially expressed genes between sexes were regulated by sex hormones, we quantified the number of genes that contained either AREs or EREs. For AREs we downloaded the coordinates of AR binding sites from the JASPAR database [31, 32] and for EREs we used a list of previously reported ERα targets [35]. In total we identified 3014 different genes that were expressed more highly in either sex in at least

one tissue. Of the 3014 genes, 875 contained AREs, 239 contained EREs and 86 contained both. On average 32% of autosomal genes that were sex biased in tissues contained AREs or EREs. Therefore, across all tissues analysed approximately two-thirds of autosomal genes did not contain either AREs or EREs. 489 genes contained AREs within gene bodies such as introns and exons, 216 genes contained AREs upstream and within the promoters and 170 genes contained AREs located downstream of the gene. The precise locations of EREs were unknown as we were using a list of previously defined ERα targets. GO enrichment for genes that contained both AREs and EREs in each individual tissue did not produce any significant enrichment, most likely due to the lists of genes being too small. We therefore found it advantageous to combine the list of genes across different tissues since the list of genes in each tissue were too small to produce any significant results. The genes that contained either or both AREs or EREs and were expressed more highly in females were enriched for GO terms relating to response to wounding and inflammatory response. For example, we found genes related to interleukin signalling and inflammatory processes to be expressed more highly in females such as TNFAIP6, IL10RB, and IFNA2 in the DLPFC, HC and AnCg, respectively. On the other hand genes containing either or both AREs or EREs that were expressed more highly in males were enriched for GO terms relating to mitochondrion and generation of precursor metabolites and energy. As already mentioned, we found a variety of ATPase's to be expressed more highly in males in the AnCg, NC, DLPFC, CB, thyroid, colon and kidney such as ATP5G1, ATP6V1B2, ATP6V0B, ATP6V1C1 and ATP6V1A. These results indicate that sex chromosome genes and sex hormones are key regulators of sex-biased gene expression across a range of tissues. However, our data also suggest a significant number of genes that have sex-biased expression may potentially be independent of direct influence by sex chromosomes or sex hormones.

3.4.5 Sex-biased epigenetic modifications

Genes that are involved in the regulation of transcription and histone modifications also showed sex differences. In the colon, genes expressed more highly in males included those that encode for histones (H3F3A, H3F3AP4, H3F3AP6 and H3F3BP1) and

ribosomal proteins (RPS3A, RPS3AP26, RPS3AP6, RPL13A, RPL4, RPL4P4, RPL13AP5, RPS3AP5, RPS3AP47, RPL7A, RPL7AP6, RPL23AP74, RPL4P5, RPL3P4, RPL13AP20 and RPL13AP25). These genes were also expressed more highly in males in other tissues such as the brain, heart and kidney. It is worth mentioning that we also found other members of the RPL gene family to be more highly expressed in females in other tissues (Supplementary Table 3-1, 3-2 and 3-3). We also found sex bias in some genes that encode for enzymes that regulate histone modifications. For example, SET, a gene that inhibits nucleosome and histone H4 acetylation [74] was expressed more highly in males in the DLPFC, SMYD3, a histone methyltransferase [75], PRMT2, PRMT5 and PRMT8 (histone arginine methyltransferases [76]) were more highly expressed in males in the AnCg and DLPFC (Supplementary Table 3-1, 3-2 and 3-3). Together these findings suggest that sex differences in tissue-specific gene expression extend from sex hormones and into genes that regulate gene expression and translation. Furthermore, our findings of sex bias in genes that encode for histones and histone modifying enzymes in most tissues suggest the possibility that sex-specific epigenetic modifications act on transcription that may result in phenotypic sex differences.

3.4.6 X-linked sex-biased gene expression

As expected, a majority of X-linked, sex-biased genes were expressed more highly in females (Figure 3-4), with the exception of those in the AnCg in which 75% were more abundantly expressed in males. The mechanism by which genes on the single copy X chromosome in males could be expressed more highly than in females with two copies is obviously likely to be associated with XCI but another mechanism is likely to be active and requires investigation. Although we do report Y chromosome genes in our analysis (Table 3-2, Supplementary Table 3-1), we do not consider these genes as differentially expressed between sexes, since females do not have a Y chromosome. We do, however, consider the reported Y chromosome genes as detectable in the analysed tissues and act as a positive control and these genes may have potential roles in the male phenotype in these tissues. Many X-linked genes that were expressed more highly in females have been previously reported to escape XCI [77]. Not surprisingly, we consistently found

XIST and JPX (genes that orchestrate XCI [78, 79]) to be expressed more highly in females and interestingly, many sex-biased X-linked genes that are known to regulate gene expression have been defined previously [80]. For example, we found KDM6A (Figure 3-5), a gene that regulates chromatin modifications, to be expressed more highly in females in the liver, lung, DLPFC, NC, AMY, FC, bladder and CB. In addition, forest plots (Figure 3-5) demonstrate consistency between individual data sets of KDM6A expression showing higher expression in females across different tissues. Furthermore, we also found KDM5C to be expressed more highly in females in the lung, FC, bladder and CB. Genes that are involved in post transcriptional processes and more highly expressed in females in the liver, thyroid, FC and CB, include ZRSR2, DDX3X which are involved in alternative splicing. In addition, we also found translation regulators EIF1AX and RPS4X, to be expressed more highly in females in the lung, pancreas, HC and colon.

Across all tissues, we found a total of 86 different genes on the X chromosome to be more highly expressed in males in at least one tissue. 22 of the 86 X chromosome genes more highly expressed in males have homologous counterparts on the Y chromosome and are located within pseudoautosomal region 1 (PAR1) [81], which may explain the differences in expression. However, not all X chromosome genes that were expressed more highly in males were within PAR1 or had homologous Y chromosome counterparts, such as SMARCA2, an ATPase and chromatin re-modeller [82]. These findings suggest that X-linked sex-biased genes may potentially regulate autosomal gene expression such as the possible case of SMARCA2, through epigenetic modifications and post transcriptional processes.

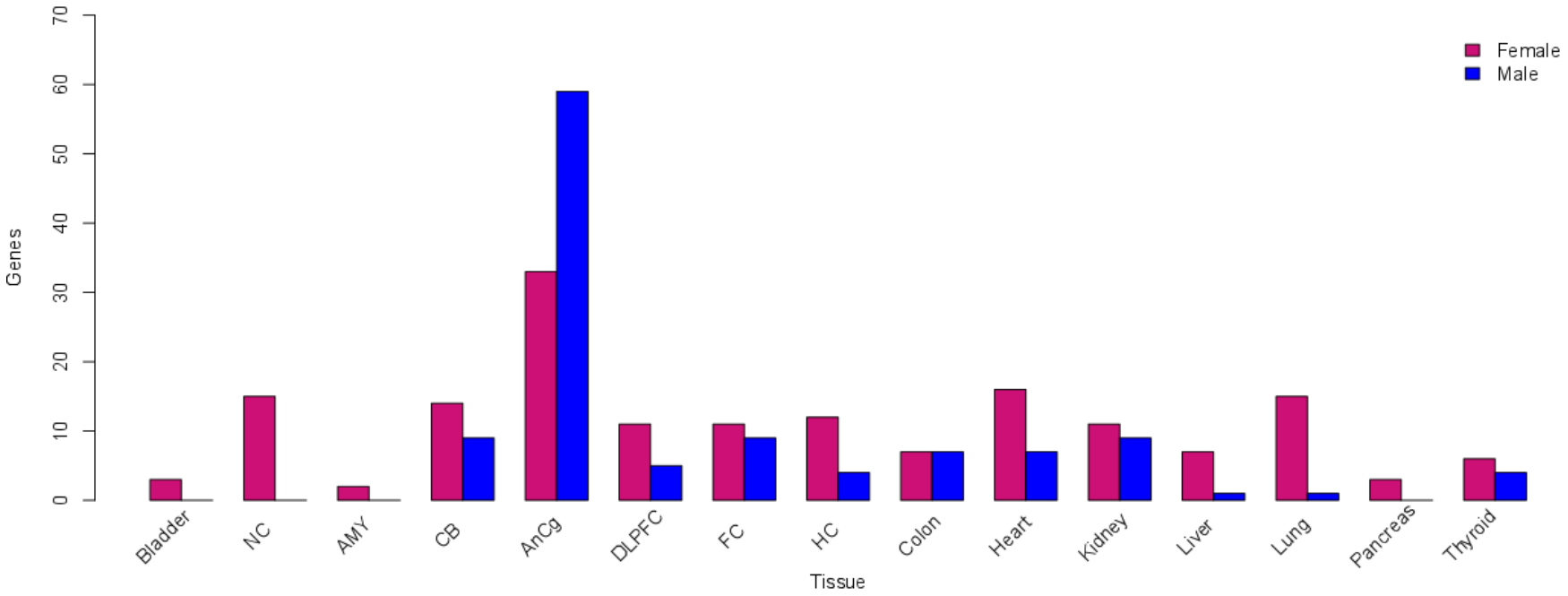


Figure 3-4: X-linked sex-biased gene expression.

The total number of genes located on the X chromosome that were expressed more highly in females (pink) and males (blue) compared to the opposite sex, respectively.

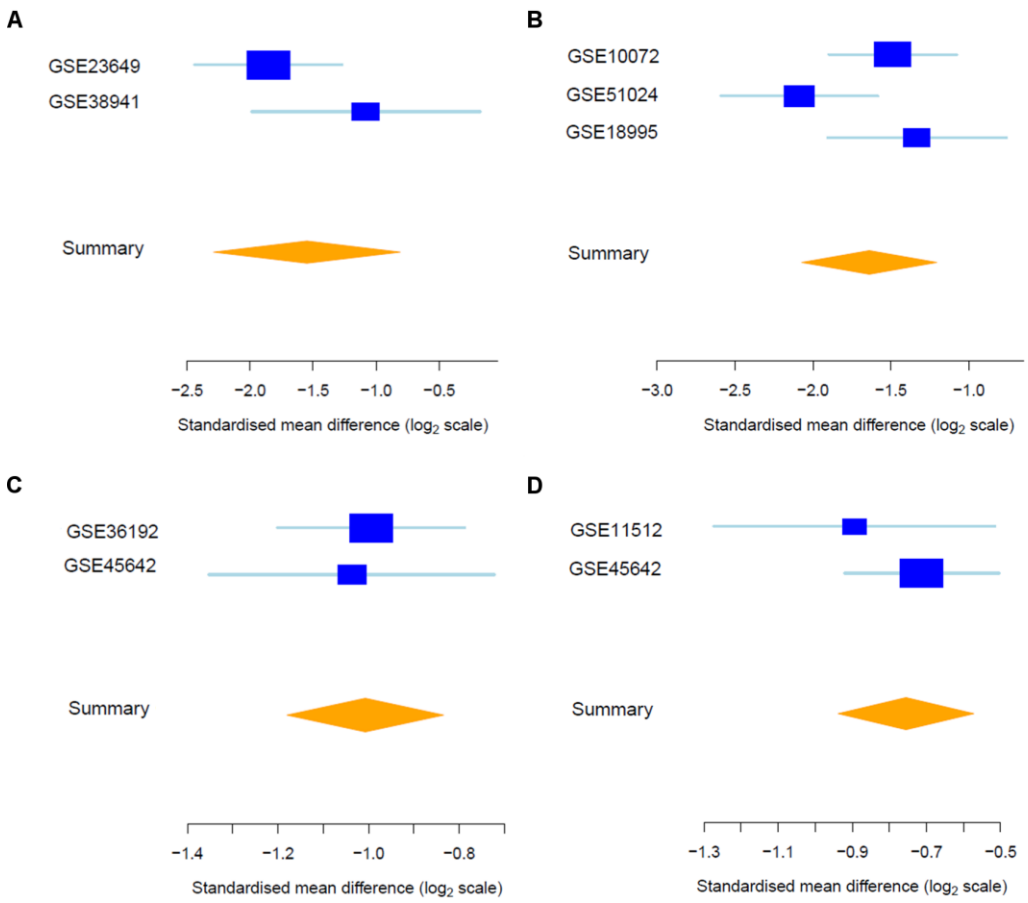


Figure 3-5: Forest plots of the standardised mean difference of KDM6A expression.

Showing higher expression in females in the liver (A), lung (B), CB (C) and DLPFC (D). Each blue box is representative of the study size in each data set and horizontal lines are standard error. The yellow diamond represents the overall gene summary for *KDM6A* in each tissue.

3.4.7 Enriched transcription factors

We next investigated which transcription factors (TFs) were enriched in the sex-biased genes by running a TF binding site (TFBS) enrichment analysis using oPOSSUM-3 and the JASPAR core motifs [31, 36]. Both the Sry-related HMG box (SOX) and the Forkhead-box (FOX) family of TFs were enriched within 10kb of the transcription start site (TSS) of sex-biased genes across all tissues (Supplementary Table 3-5). The SOX TFs are vital for sex determination [83] and the FOX TFs are essential for embryonic development and also have roles in regulating the immune system [84-86]. Sex chromosome derived TFs such as ZFX and SRY were also enriched within 10kb of the TSS. We also found the androgen receptor (AR) as an enriched TF within the AMY, CB, FC, bladder, and lung. In addition, HNF1A and HNF1B were enriched in genes upregulated in both males and females within all tissues apart from the NC and DLPFC. HNF1A and HNF1B are homeobox TFs that are required for expression of specific liver genes [87]. These findings reveal TFs that may have important roles in regulating sexually dimorphic gene expression such as HNF1A and HNF1B, which as mentioned earlier have only previously been shown to be required for expression of specific liver genes [87]. However, the genes that encode for the majority of the TFs that were enriched within sex-biased genes were not themselves differentially expressed between the sexes. Although in this study we focus on gene expression, TFs undergo more processing post transcription and therefore their protein abundance within tissues may differ between sexes.

3.4.8 Sex differences in other tissues

In this study we have analysed sex-biased gene expression in 15 human tissues. However, we must acknowledge other studies that have also analysed sex-biased gene expression. One of the largest studies that has analysed sex-biased gene expression is the Genotype-Tissue Expression (GTEx) project [88]. The GTEx project has used RNA-seq to analyse gene expression in a variety of different human tissues which would give a broader comparison of gene expression differences between tissues. In comparison to GTEx [88]

we have analysed sex-biased gene expression in 5 of the same human tissues which is represented as a Venn diagram (Supplementary Figure 3-1). We found an overlap of sex chromosome genes as being sex-biased between this study and GTEx. However, there were many genes that we found to be sex-biased that were not in GTEx [88]. A possible explanation for the difference between studies is that in GTEx only samples from175 individuals were used [88] as opposed to over 2500 in this study which provides much greater statistical power compared to GTEx [88]. In addition, GTEx also used RNA-seq and were therefore able to quantify the expression of genes for which no probes were available in the microarrays used in this study.

3.4.9 Bias of male samples

To prevent any biases in our analyses we have performed differential gene expression in tissues from all publicly available data to our knowledge. However, since most studies neglect to account for samples sex [7, 8], we unfortunately had a ratio of 2.1:1 males to females on average across all tissues analysed. Therefore, this in itself may create some biases in our analyses. Across all data sets (Table 3-1) the ratio of males to females was skewed towards males apart from one data set containing thyroid samples (GSE33630), where the ratio was 2.5 females for every male.

To determine if the ratio of males to females affects the differential expression analyses we conducted a 10-fold cross validation of the differential gene expression analyses in the tissue where the ratio of males to females was the greatest. The AMY gene expression data had a ratio of 4.5 males to every female. In this analysis we randomly removed male samples from the analysis to make the number of each sex the same and then assessed which genes were differentially expressed between males and females. We performed this analysis 10 times and then compared which genes were consistently identified as sex-biased to our original analysis where we did not sub-set any male samples. In the analysis with the sex chromosomes included we found the sex chromosome genes (XIST, RPS4Y1, DDX3Y, KDM5D, USP9Y, EIF1AY and TTTY15) consistently classified as sex-biased in the 10-fold cross validation. However, in the original analysis we identified 4 autosomal genes to be sex-biased and upregulated in females (Supplementary Table 3-

1). However, these 4 autosomal genes were not found to be sex-biased in the 10-fold cross validation. By performing the 10 fold cross validation, we removed samples which would have decreased our statistical power and therefore increased the magnitude of the adjusted p-value which is what occurred. Therefore caution should be taken when interpreting the results of genes that were found to be sex-biased with an adjusted p-value close to 0.05 and in tissues where there is a large ratio of males to females. However, this analysis does provide reassurance that the sex chromosome genes that were found to be sex-biased in the original analysis were not greatly affected by the bias in male samples.

3.4.10 Strengths and limitations

While our analyses reveal many sex differences in gene expression within a variety of tissues, there are several limitations to this study. Firstly, most tissues (where age was provided) were from individuals who were post reproductive age (average age = 47 years) which may not have captured the true extent of sex-biased gene expression that would otherwise be evident during early adulthood when sex hormones are at their peak production. Thus, using data from older individuals limited our ability to assess sex-biased gene expression in individuals of reproductive age. We also report a number of genes previously associated with diseases and disorders that were differentially expressed between sexes. RNA expression differences do not necessarily cause phenotypic variation, as there are multiple levels of gene and protein regulation that can occur post transcription. Next-generation sequencing, as opposed to microarrays used in this study, would allow a more complete assessment of sex-dependent gene expression differences but there is currently more samples that have been analysed using microarrays and therefore more statistical power can be achieved. Furthermore, on average, 64% of genes differentially expressed between sexes in each tissue had a magnitude log2FC < 1. Most genes that were found to be sex-biased do not have large log2FC apart from genes located on the sex chromosomes. In addition, most genes that were found to be sex-biased across all tissues had a magnitude log2FC < 1.5 (Supplementary Table 3-7). Therefore, future studies would need to be adequately powered to replicate our findings.

Despite these limitations, to our knowledge this is the largest analysis of sex differences in gene expression across a range of human tissues.

Despite the large amount of genomic data that was available for this study it was unfortunate not to consider clinical and lifestyle factors such as age, smoking status, sample heterogeneity and body mass index (BMI) which may potentially have an effect on gene expression. We were unable to correct for these potential confounding factors because, as detailed in Supplementary Table 3-6, most studies provide little or no clinical information about the samples. Furthermore, only 32% of all the samples analysed in this study were from females which may potentially create a bias for genes to be more highly expressed in males. However, by acknowledging this limitation we draw attention to the bias towards using only males in biomedical research. We therefore urge future research in all fields of biomedical science to use an equal sex ratio in study design.

3.5 Conclusions

Our analyses have revealed substantial differences in the transcriptional landscape between sexes across a range of human organs and tissues and highlight possible mechanisms by which gene expression may contribute to sexually dimorphic traits. Improved understanding of these is fundamental to understanding diseases with different prevalence between the sexes. Our data show that sex differences in gene expression vary widely across different tissues. We identified a consistent trend for genes known to regulate the immune system to be more highly expressed in females and those involved in energy production and growth were more highly expressed in males. These may be the result of different evolutionary pressures between the sexes. The brain demonstrates the largest differences in sex-biased gene expression with several sex-biased genes associated with specific brain disorders, providing insight into possible mechanisms for the association of sex-specific prevalence of certain brain disorders.

Our findings also indicate that many sex biased genes within tissues are independent of sex chromosome genes or sex hormones. Approximately 32% of autosomal genes in each

tissue contained an ARE or ERE, which suggests there are other mechanisms that underpin sex differences in gene expression. One potential mechanism is through epigenetic factors, such as chromatin modelling which has been suggested to have sex specific functional roles [89].

Finally, our data demonstrate why it is important to consider sex as a biological confounder in biomedical studies. Future studies should incorporate sex differences in their analyses which will help to provide new insights in health and disease. The sex-biased genes identified in this study provide a basis for determining the mechanism by which sexual dimorphism occurs and potential causal pathways for sexually biased disease susceptibility. More importantly however, they provide potential targets for novel sex specific treatments.

3.6 Acknowledgements

The authors would like to thank the generosity of all individuals who were involved in the data creation of all data sets that were available for public analysis.

3.7 Supporting Information

For Supplementary Table 3-1, 3-2, 3-3, 3-4, and 3-5 please refer to the electronic supporting information.

Supplementary Table 3-1: Differential gene expression analysis between males and females in each tissue, including the sex chromosomes.

A list of differentially expressed genes between sexes in each tissue with all the chromosomes included in the analysis. A fold change > 0 indicates the gene is expressed more highly in males and a fold change < 0 indicates the gene is expressed more highly in females.

Supplementary Table 3-2: Differential gene expression analysis results with genes on the Y chromosome removed from the analysis.

A list of sex-biased genes with the Y chromosome genes removed from the analysis. A fold change > 0 indicates the gene is expressed more highly in males and a fold change < 0 indicates the gene is expressed more highly in females.

Supplementary Table 3-3: Differential gene expression analysis results with genes on the X and Y chromosomes removed from the analysis.

A list of sex-biased genes with the sex chromosome genes removed from the analysis. A fold change > 0 indicates the gene is expressed more highly in males and a fold change < 0 indicates the gene is expressed more highly in females.

Supplementary Table 3-4: Gene ontology results.

This table lists all the GO terms that were found to be enriched within each tissue. Only significant GO terms were found for the NC, AnCg and HC.

Supplementary Table 3-5: Transcription factors that were found to contain enriched motifs with 10kb of the transcription start site of sex-biased genes in each tissue.

A list of enriched transcription factors of the sex-biased genes in each tissue.

Supplementary Table 3-6: Clinical and lifestyle factors supplied by each data set.

A table representing which data set supplied sample information such as age, ethnicity, sex, smoking status and disease status.

GEO Accession	Organ/Tissue	Age	Ethnicity	Smoking	BMI	RIN	Sex	Batches	Diseases	Contact Country
GSE10072	Lung	Yes	NS	Yes	NS	NS	Yes	NA	Cancer	USA
GSE11512	Brain	Yes	NS	NS	NS	NS	Yes	NA	NA	Germany
GSE13507	Bladder	NS	NS	NS	NS	NS	NS	NA	Cancer	South Korea
GSE15471	Pancreas	NS	NS	NS	NS	NS	NS	NA	Cancer	Romania
GSE18995	Liver	NS	NS	NS	NS	NS	NS	NA	NA	South Korea
GSE23649	Liver	NS	NS	NS	NS	NS	NS	NA	NA	Spain
GSE26887	Heart	Yes	NS	NS	NS	NS	Yes	NA	Diabetes	Italy
GSE33630	Thyroid	NS	NS	NS	NS	NS	NS	NS	Cancer	Brussels
GSE36192	Brain	NS	NS	NS	NS	NS	NS	Yes	NA	Bethesda
GSE38941	Liver	NS	NS	NS	NS	NS	NS	NA	HBV	USA
GSE41328	Colon	NS	NS	NS	NS	NS	NS	NA	Cancer	USA
GSE43974	Kidney	NS	NS	NS	NS	NS	NS	NA	NA	Netherlands
GSE44456	Brain	Yes	NS	Yes	NS	NS	Yes	Yes	Cirrhosis	USA
GSE45642	Brain	NS	NS	NS	NS	NS	NS	Yes	NA	USA
GSE50892	Kidney	Yes	Yes	NS	NS	NS	NS	NA	Cirrhosis	USA
GSE51024	Lung	NS	NS	NS	NS	NS	NS	NA	Cancer	USA
GSE54572	Brain	NS	NS	NS	NS	NS	NS	NA	NA	USA
GSE55231	Heart	Yes	NS	NS	NS	NS	Yes	NA	NA	Netherlands
GSE57338	Heart	Yes	NS	NS	NS	NS	Yes	NA	Heart Failure	USA
GSE61276	Liver	NS	NS	NS	NS	NS	NS	Yes	Cancer	Estonia
GSE65144	Thyroid	NS	NS	NS	NS	NS	NS	NA	Cancer	USA
GSE8671	Colon	NS	NS	NS	NS	NS	NS	NA	Cancer	Switzerland

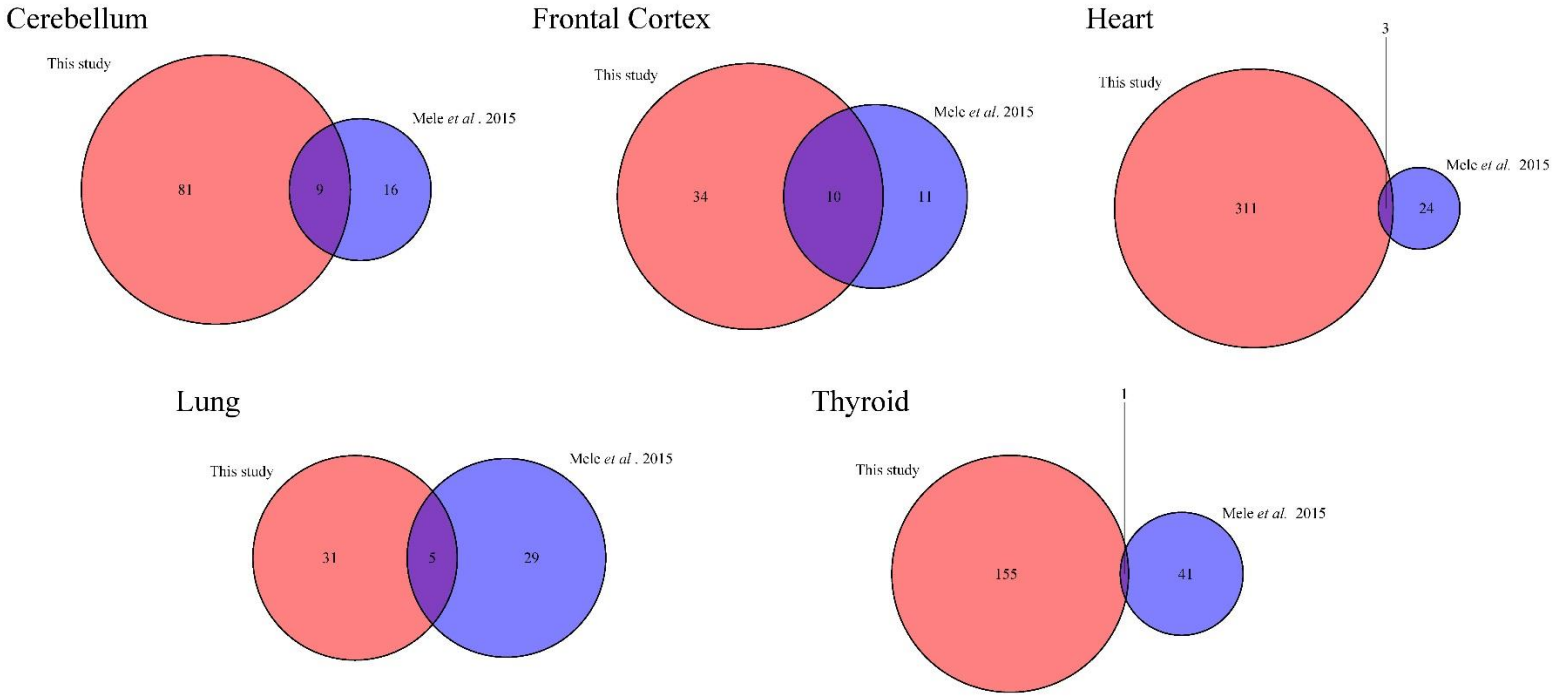
NA: not applicable, NS: not supplied

Supplementary Table 3-7: Total number of sex-biased genes at different log¬2¬FC cut-offs.

A table listing the total number of genes that were found to be sex-biased in each tissue at different log_2FC cut-offs. This analysis was performed with the sex chromosomes included.

	Magnitude Log_2FC cut-offs				
Tissue	**No Cut off**	**FC > 0.5**	**FC > 1**	**FC > 1.5**	**FC > 2**
Bladder	16	13	10	7	6
Brain (Nucleus Accumbens)	264	55	9	4	4
Brain (Amygdala)	17	12	6	4	3
Brain (Cerebellum)	98	32	20	12	8
Brain (Anterior Cingulate Cortex)	1818	1281	54	10	0
Brain (Dorsolateral Prefrontal Cortex)	198	98	7	4	3
Brain (Frontal Cortex)	45	45	45	45	45
Brain (Hippocampus)	205	168	43	12	7
Colon	218	218	175	125	106
Heart	375	251	31	7	2
Kidney	224	59	4	4	3
Liver	32	29	12	5	2
Lung	36	33	19	9	4
Pancreas	22	19	13	11	10
Thyroid	163	162	85	45	27

FC: fold change



Supplementary Figure 3-1: Venn diagrams representing the overlap of defined sex-biased genes between this study and a previous study.

Each Venn diagram represents an individual tissue and the overlap of genes that were found to be sex-biased between studies.

Supplementary File 3-1: Detailed methodology. A description of the precise methods used involved in data collection, data processing, normalisation, batch correction and differential expression.

Overview

In this additional file we supply more detail on the precise methods used to conduct our gene expression meta-analysis.

Data collection

Initially, we searched the Gene Expression Omnibus (GEO) [1] and ArrayExpress [2] for microarray data sets containing human tissues from healthy individuals. We identified 22 microarray data sets containing 15 different human tissues from 9 different organs. Unfortunately, we were unable to consider lifestyle and clinical factors such as smoking, body mass index (BMI), age and ethnicity since most data sets did not supply any of this information (Supplementary Table 3-6). We accept not being able to correct for these factors is a limitation of our study. However, despite the enormous generosity of the researchers who had made their data publicly available we think researchers should be encouraged to collect and supply clinical data with future genomic data sets. Despite a lack of clinical and lifestyle data on individual samples we were able to select samples from data sets where the individual did not have any cancerous tissue in their body. For example, in the data set GSE61276, there were liver tissue samples from fetuses, adult liver tissues from individuals who had met accidental death and liver tissue from individuals who had malignant tumours. Therefore, we selected only adult liver tissue samples from individuals who had met accidental death. Although the precise details of death and other clinical details of each individual in GSE61276 is unknown to the public, by removing samples from individuals who had malignancies reduced the potential for confounders in our differential expression analyses. In addition and as stated in our manuscript we chose data sets that had more than 10 samples for better sample sex determination which is described in more detail in the ‘Identifying sample sex’ section.

Data processing and normalisation

The 22 microarray data sets used in this study used either an Affymetrix or Illumina platform and the pre-processing and normalisation for each platform was tailored by using a variety of Bioconductor packages (http://bioconductor.org/). For studies that used Affymetrix platforms, we firstly downloaded the individual .CEL files from GEO and used simpleaffy [3] to generate a matrix of normalised expression values for each individual data sets. In studies that had used Illumina platforms, we used beadarray [4] which too generated a matrix of normalised expression values for data sets using Illumina platforms.

As detailed in Supplementary Table 3-6, 4 of the 22 data sets contained data obtained in batches. To correct for these batch effects in these three individual data sets we used the Combat function in the SVA package [5].

Differential expression analyses

Differential expression analysis between males and females was performed independently in each tissue. Depending on the number of data sets used in each tissue we used either one of two methods to identify sex-biased genes.

For tissues such as the bladder, where there was only one available data set we used the Empirical Bayes method that is described within the limma package [6]. For tissues such as the heart, where we had more than one data set containing the tissue we used metaGEM package (https://spiral.imperial.ac.uk/handle/10044/1/4217) and used the inverse–variance method as previously described [7]. We chose the metaGEM approach with tissues with more than one data set since the data set have used different microarray platforms. In other words, since each microarray platform has used different probes to annotate for gene expression, their expression values cannot be grouped together. Furthermore, using this approach allows for correction of the data set size using Hedges' g. Using this approach allows the use of more samples that have been analysed on multiple different microarray platforms which is beneficial to gaining sufficient statistical power.

One of the limitations of using these two approaches is that the magnitude difference in gene expression between males and females is not necessarily interchangeable. To overcome this limitation, one approach would be to use a single platform to measure gene expression such as RNA sequencing. However, there are currently many more publicly available samples that have been analysed by microarrays than by RNA-seq. We therefore focused on using microarrays in this study to maximise the statistical power of our analyses.

Supplementary File 3-1 References

1. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al: **NCBI GEO: archive for functional genomics data sets--update.** *Nucleic Acids Res* 2013, **41:**D991-995.
2. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, et al: **ArrayExpress—a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31:**68-71.
3. Wilson CL, Miller CJ: **Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis.** *Bioinformatics* 2005, **21:**3683-3685.
4. Dunning MJ, Smith ML, Ritchie ME, Tavare S: **beadarray: R classes and methods for Illumina bead-based data.** *Bioinformatics* 2007, **23:**2183-2184.
5. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments.** *Bioinformatics* 2012, **28:**882-883.
6. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res* 2015.
7. Ramasamy A, Mondry A, Holmes CC, Altman DG: **Key issues in conducting a meta-analysis of gene expression microarray datasets.** *PLoS Med* 2008, **5:**e184.

References

1. Morrow EH: **The evolution of sex differences in disease.** *Biol Sex Differ* 2015, **6:**5.
2. Möller-Leimkühler AM: **Gender differences in cardiovascular disease and comorbid depression.** *Dialogues Clin Neurosci* 2007, **9:**71-83.
3. Eastwood JA, Doering LV: **Gender differences in coronary artery disease.** *J Cardiovasc Nurs* 2005, **20:**340-351; quiz 352-343.
4. Ochoa S, Usall J, Cobo J, Labad X, Kulkarni J: **Gender differences in schizophrenia and first-episode psychosis: a comprehensive literature review.** *Schizophr Res Treatment* 2012, **2012:**916198.
5. Carlson C, Dugan P, Kirsch HE, Friedman D: **Sex differences in seizure types and symptoms.** *Epilepsy Behav* 2014, **41:**103-108.
6. Qin W, Liu C, Sodhi M, Lu H: **Meta-analysis of sex differences in gene expression in schizophrenia.** *BMC Syst Biol* 2016, **10 Suppl 1:**9.
7. Beery AK, Zucker I: **Sex bias in neuroscience and biomedical research.** *Neurosci Biobehav Rev* 2011, **35:**565-572.
8. Mogil JS, Chanda ML: **The case for the inclusion of female subjects in basic science studies of pain.** *Pain* 2005, **117:**1-5.
9. Zucker I, Beery AK: **Males still dominate animal studies.** *Nature* 2010, **465:**690.
10. Trabzuni D, Ramasamy A, Imran S, Walker R, Smith C, Weale ME, Hardy J, Ryten M: **Widespread sex differences in gene expression and splicing in the adult human brain.** *Nat Commun* 2013, **4:**2771.
11. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, et al: **Spatio-temporal transcriptome of the human brain.** *Nature* 2011, **478:**483-489.
12. Vawter MP, Evans S, Choudary P, Tomita H, Meador-Woodruff J, Molnar M, Li J, Lopez JF, Myers R, Cox D, et al: **Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes.** *Neuropsychopharmacology* 2004, **29:**373-384.
13. Weickert CS, Elashoff M, Richards AB, Sinclair D, Bahn S, Paabo S, Khaitovich P, Webster MJ: **Transcriptome analysis of male-female differences in prefrontal cortical development.** *Mol Psychiatry* 2009, **14:**558-561.
14. Reinius B, Jazin E: **Prenatal sex differences in the human brain.** *Mol Psychiatry* 2009, **14:**987, 988-989.
15. Hall E, Volkov P, Dayeh T, Esguerra JL, Salo S, Eliasson L, Ronn T, Bacos K, Ling C: **Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets.** *Genome Biol* 2014, **15:**522.
16. Fermin DR, Barac A, Lee S, Polster SP, Hannenhalli S, Bergemann TL, Grindle S, Dyke DB, Pagani F, Miller LW, et al: **SEX AND AGE DIMORPHISM OF MYOCARDIAL GENE EXPRESSION IN NONISCHEMIC HUMAN HEART FAILURE.** *Circulation Cardiovascular genetics* 2008, **1:**117-125.

17. Zhang Y, Klein K, Sugathan A, Nassery N, Dombkowski A, Zanger UM, Waxman DJ: **Transcriptional Profiling of Human Liver Identifies Sex-Biased Genes Associated with Polygenic Dyslipidemia and Coronary Artery Disease.** *PLoS One* 2011, **6:**e23506.
18. Carrel L, Willard HF: **X-inactivation profile reveals extensive variability in X-linked gene expression in females.** *Nature* 2005, **434:**400-404.
19. Yang F, Babak T, Shendure J, Disteche CM: **Global survey of escape from X inactivation by RNA-sequencing in mouse.** *Genome Res* 2010, **20:**614-622.
20. Buckberry S, Bianco-Miotto T, Bent SJ, Dekker GA, Roberts CT: **Integrative transcriptome meta-analysis reveals widespread sex-biased gene expression at the human fetal-maternal interface.** *Mol Hum Reprod* 2014, **20:**810-819.
21. Voskuhl RR, Palaszynski K: **Sex hormones in experimental autoimmune encephalomyelitis: implications for multiple sclerosis.** *Neuroscientist* 2001, **7:**258-270.
22. Ebers GC, Sadovnick AD, Dyment DA, Yee IM, Willer CJ, Risch N: **Parent-of-origin effect in multiple sclerosis: observations in half-siblings.** *Lancet* 2004, **363:**1773-1774.
23. Dunning MJ, Smith ML, Ritchie ME, Tavare S: **beadarray: R classes and methods for Illumina bead-based data.** *Bioinformatics* 2007, **23:**2183-2184.
24. Wilson CL, Miller CJ: **Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis.** *Bioinformatics* 2005, **21:**3683-3685.
25. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments.** *Bioinformatics* 2012, **28:**882-883.
26. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics--a bioconductor package for quality assessment of microarray data.** *Bioinformatics* 2009, **25:**415-416.
27. Buckberry S, Bent SJ, Bianco-Miotto T, Roberts CT: **massiR: a method for predicting the sex of samples in gene expression microarray datasets.** *Bioinformatics* 2014, **30:**2084-2085.
28. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc* 2009, **4:**1184-1191.
29. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res* 2015.
30. Ramasamy A, Mondry A, Holmes CC, Altman DG: **Key issues in conducting a meta-analysis of gene expression microarray datasets.** *PLoS Med* 2008, **5:**e184.
31. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, et al: **JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.** *Nucleic Acids Res* 2014, **42:**D142-147.
32. Hu S, Yao G, Guan X, Ni Z, Ma W, Wilson EM, French FS, Liu Q, Zhang Y: **Research resource: Genome-wide mapping of in vivo androgen receptor binding sites in mouse epididymis.** *Mol Endocrinol* 2010, **24:**2392-2405.

33. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA: **Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.** *Int J Epidemiol* 2012, **41:**200-209.
34. BP. CMaM: **TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s). R package version 3.2.2.**
35. Jin VX, Leu Y-W, Liyanarachchi S, Sun H, Fan M, Nephew KP, Huang THM, Davuluri RV: **Identifying estrogen receptor α target genes using integrated computational genomics and chromatin immunoprecipitation microarray.** *Nucleic Acids Res* 2004, **32:**6627-6635.
36. Kwon AT, Arenillas DJ, Worsley Hunt R, Wasserman WW: **oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets.** *G3 (Bethesda)* 2012, **2:**987-1002.
37. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44-57.
38. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J: **g:Profiler-a web server for functional interpretation of gene lists (2016 update).** *Nucleic Acids Res* 2016, **44:**W83-89.
39. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al: **NCBI GEO: archive for functional genomics data sets—update.** *Nucleic Acids Res* 2013, **41:**D991-D995.
40. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, et al: **ArrayExpress—a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31:**68-71.
41. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57:**289-300.
42. Grundtman C, Kreutmayer SB, Almanzar G, Wick MC, Wick G: **Heat Shock Protein 60 and Immune Inflammatory Responses in Atherosclerosis.** *Arterioscler Thromb Vasc Biol* 2011, **31:**960-968.
43. Lin LC, Lewis DA, Sibille E: **A human-mouse conserved sex bias in amygdala gene expression related to circadian clock and energy metabolism.** *Mol Brain* 2011, **4:**18.
44. Allman JM, Hakeem A, Erwin JM, Nimchinsky E, Hof P: **The anterior cingulate cortex. The evolution of an interface between emotion and cognition.** *Ann N Y Acad Sci* 2001, **935:**107-117.
45. Liu J, Zubieta JK, Heitzeg M: **Sex differences in anterior cingulate cortex activation during impulse inhibition and behavioral correlates.** *Psychiatry Res* 2012, **201:**54-62.
46. Seney ML, Sibille E: **Sex differences in mood disorders: perspectives from humans and rodent models.** *Biol Sex Differ* 2014, **5:**17.
47. Yang X, Wang S, Kendrick KM, Wu X, Yao L, Lei D, Kuang W, Bi F, Huang X, He Y, Gong Q: **Sex differences in intrinsic brain functional connectivity underlying human shyness.** *Soc Cogn Affect Neurosci* 2015, **10:**1634-1643.

48. Wu LJ, Kim SS, Li X, Zhang F, Zhuo M: **Sexual attraction enhances glutamate transmission in mammalian anterior cingulate cortex.** *Mol Brain* 2009, **2:**9.
49. Abel KM, Drake R, Goldstein JM: **Sex differences in schizophrenia.** *Int Rev Psychiatry* 2010, **22:**417-428.
50. Joutel A, Corpechot C, Ducros A, Vahedi K, Chabriat H, Mouton P, Alamowitch S, Domenga V, Cecillion M, Marechal E, et al: **Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia.** *Nature* 1996, **383:**707-710.
51. Wang Y, Hu Y, Fang Y, Zhang K, Yang H, Ma J, Xu Q, Shen Y: **Evidence of epistasis between the catechol-O-methyltransferase and aldehyde dehydrogenase 3B1 genes in paranoid schizophrenia.** *Biol Psychiatry* 2009, **65:**1048-1054.
52. Zhang X, Bertaso F, Yoo JW, Baumgartel K, Clancy SM, Lee V, Cienfuegos C, Wilmot C, Avis J, Hunyh T, et al: **Deletion of the potassium channel Kv12.2 causes hippocampal hyperexcitability and epilepsy.** *Nat Neurosci* 2010, **13:**1056-1058.
53. Huang CC, Cheng MC, Tsai HM, Lai CH, Chen CH: **Genetic analysis of GABRB3 at 15q12 as a candidate gene of schizophrenia.** *Psychiatr Genet* 2014, **24:**151-157.
54. Gurba KN, Hernandez CC, Hu N, Macdonald RL: **GABRB3 mutation, G32R, associated with childhood absence epilepsy alters alpha1beta3gamma2L gamma-aminobutyric acid type A (GABAA) receptor expression and channel gating.** *J Biol Chem* 2012, **287:**12083-12097.
55. Buxbaum JD, Silverman JM, Smith CJ, Greenberg DA, Kilifarski M, Reichert J, Cook EH, Jr., Fang Y, Song CY, Vitale R: **Association between a GABRB3 polymorphism and autism.** *Mol Psychiatry* 2002, **7:**311-316.
56. Wang X, Yu S, Li F, Feng T: **Detection of alpha-synuclein oligomers in red blood cells as a potential biomarker of Parkinson's disease.** *Neurosci Lett* 2015, **599:**115-119.
57. Jonsson EG, Saetre P, Nyholm H, Djurovic S, Melle I, Andreassen OA, Skjodt C, Thygesen JH, Werge T, Hall H, et al: **Lack of association between the regulator of G-protein signaling 4 (RGS4) rs951436 polymorphism and schizophrenia.** *Psychiatr Genet* 2012, **22:**263-264.
58. Shi L, Zhang Z, Su B: **Sex Biased Gene Expression Profiling of Human Brains at Major Developmental Stages.** *Sci Rep* 2016, **6:**21181.
59. Maas A, Appelman YEA: **Gender differences in coronary heart disease.** *Neth Heart J* 2010, **18:**598-602.
60. Iio C, Ogimoto A, Nagai T, Suzuki J, Inoue K, Nishimura K, Uetani T, Okayama H, Okura T, Shigematsu Y, et al: **Association Between Genetic Variation in the SCN10A Gene and Cardiac Conduction Abnormalities in Patients With Hypertrophic Cardiomyopathy.** *Int Heart J* 2015, **56:**421-427.
61. Splawski I, Shen J, Timothy KW, Lehmann MH, Priori S, Robinson JL, Moss AJ, Schwartz PJ, Towbin JA, Vincent GM, Keating MT: **Spectrum of mutations in long-QT syndrome genes. KVLQT1, HERG, SCN5A, KCNE1, and KCNE2.** *Circulation* 2000, **102:**1178-1185.

62. O'Reilly D, Dienstbier M, Cowley SA, Vazquez P, Drozdz M, Taylor S, James WS, Murphy S: **Differentially expressed, variant U1 snRNAs regulate gene expression in human cells.** *Genome Res* 2013, **23:**281-291.
63. Guiro J, O'Reilly D: **Insights into the U1 small nuclear ribonucleoprotein complex superfamily.** *Wiley Interdiscip Rev RNA* 2015, **6:**79-92.
64. Adjei AA, Thomae BA, Prondzinski JL, Eckloff BW, Wieben ED, Weinshilboum RM: **Human estrogen sulfotransferase (SULT1E1) pharmacogenomics: gene resequencing and functional genomics.** *Br J Pharmacol* 2003, **139:**1373-1382.
65. Miki Y, Nakata T, Suzuki T, Darnel AD, Moriya T, Kaneko C, Hidaka K, Shiotsu Y, Kusaka H, Sasano H: **Systemic distribution of steroid sulfatase and estrogen sulfotransferase in human adult and fetal tissues.** *J Clin Endocrinol Metab* 2002, **87:**5760-5768.
66. Yang X, Schadt EE, Wang S, Wang H, Arnold AP, Ingram-Drake L, Drake TA, Lusis AJ: **Tissue-specific expression and regulation of sexually dimorphic genes in mice.** *Genome Res* 2006, **16:**995-1004.
67. van Nas A, Guhathakurta D, Wang SS, Yehya N, Horvath S, Zhang B, Ingram-Drake L, Chaudhuri G, Schadt EE, Drake TA, et al: **Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks.** *Endocrinology* 2009, **150:**1235-1249.
68. Selva DM, Hammond GL: **Thyroid hormones act indirectly to increase sex hormone-binding globulin production by liver via hepatocyte nuclear factor-4alpha.** *J Mol Endocrinol* 2009, **43:**19-27.
69. Karlsson C, Baudet A, Miharada N, Soneji S, Gupta R, Magnusson M, Enver T, Karlsson G, Larsson J: **Identification of the chemokine CCL28 as a growth and survival factor for human hematopoietic stem and progenitor cells.** *Blood* 2013, **121:**3838-3842, S3831-3815.
70. Plate M, Li T, Wang Y, Mo X, Zhang Y, Ma D, Han W: **Identification and characterization of CMTM4, a novel gene with inhibitory effects on HeLa cell growth through Inducing G2/M phase accumulation.** *Mol Cells* 2010, **29:**355-361.
71. Vakili H, Jin Y, Cattini PA: **Energy homeostasis targets chromosomal reconfiguration of the human GH1 locus.** *J Clin Invest* 2014, **124:**5002-5012.
72. Ehrenkranz J, Bach PR, Snow GL, Schneider A, Lee JL, Ilstrup S, Bennett ST, Benvenga S: **Circadian and Circannual Rhythms in Thyroid Hormones: Determining the TSH and Free T4 Reference Intervals Based Upon Time of Day, Age, and Sex.** *Thyroid* 2015, **25:**954-961.
73. Meikle AW: **The interrelationships between thyroid dysfunction and hypogonadism in men and boys.** *Thyroid* 2004, **14 Suppl 1:**S17-25.
74. Krajewski WA, Vassiliev OL: **Interaction of SET domains with histones and nucleic acid structures in active chromatin.** *Clin Epigenetics* 2011, **2:**17-25.
75. Hamamoto R, Furukawa Y, Morita M, Iimura Y, Silva FP, Li M, Yagyu R, Nakamura Y: **SMYD3 encodes a histone methyltransferase involved in the proliferation of cancer cells.** *Nat Cell Biol* 2004, **6:**731-740.
76. Lorenzo AD, Bedford MT: **Histone Arginine Methylation.** *FEBS Lett* 2011, **585:**2024-2031.

77. Cotton AM, Price EM, Jones MJ, Balaton BP, Kobor MS, Brown CJ: **Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation.** *Hum Mol Genet* 2014.
78. Augui S, Nora EP, Heard E: **Regulation of X-chromosome inactivation by the X-inactivation centre.** *Nat Rev Genet* 2011, **12:**429-442.
79. Lee JT: **Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control.** *Nat Rev Mol Cell Biol* 2011, **12:**815-826.
80. Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, Koutseva N, Zaghlul S, Graves T, Rock S, et al: **Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators.** *Nature* 2014, **508:**494-499.
81. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, et al: **The DNA sequence of the human X chromosome.** *Nature* 2005, **434:**325-337.
82. Takeshima H, Niwa T, Takahashi T, Wakabayashi M, Yamashita S, Ando T, Inagawa Y, Taniguchi H, Katai H, Sugiyama T, et al: **Frequent involvement of chromatin remodeler alterations in gastric field cancerization.** *Cancer Lett* 2015, **357:**328-338.
83. Huang YH, Jankowski A, Cheah KS, Prabhakar S, Jauch R: **SOXE transcription factors form selective dimers on non-compact DNA motifs through multifaceted interactions between dimerization and high-mobility group domains.** *Sci Rep* 2015, **5:**10398.
84. Lam EWF, Brosens JJ, Gomes AR, Koo C-Y: **Forkhead box proteins: tuning forks for transcriptional harmony.** *Nat Rev Cancer* 2013, **13:**482-495.
85. Jackson BC, Carpenter C, Nebert DW, Vasiliou V: **Update of human and mouse forkhead box (FOX) gene families.** *Hum Genomics* 2010, **4:**345-352.
86. Coffer PJ, Burgering BMT: **Forkhead-box transcription factors and their role in the immune system.** *Nat Rev Immunol* 2004, **4:**889-899.
87. Shih DQ, Bussen M, Sehayek E, Ananthanarayanan M, Shneider BL, Suchy FJ, Shefer S, Bollileni JS, Gonzalez FJ, Breslow JL, Stoffel M: **Hepatocyte nuclear factor-1alpha is an essential regulator of bile acid and plasma cholesterol metabolism.** *Nat Genet* 2001, **27:**375-382.
88. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al: **The human transcriptome across tissues and individuals.** *Science (New York, NY)* 2015, **348:**660-665.
89. Silkaitis K, Lemos B: **Sex-biased chromatin and regulatory cross-talk between sex chromosomes, autosomes, and mitochondria.** *Biol Sex Differ* 2014, **5:**2-2.

Statement of Authorship

Title of Paper	Identification of novel human placenta specific large intergenic non-coding RNAs in deeply sequenced RNAseq from first trimester, term and preeclampitic pregnancies
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Benjamin Mayne
Contribution to the Paper	Designed, conducted the study, analysed and interpreted the data and wrote the manuscript.
Overall percentage (%)	80
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 14-7-17

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Shalem Y Leemaqz
Contribution to the Paper	Analyzed the data, provided intellectual input into the manuscript.
Signature	Date 14 July 2017

Name of Co-Author	Dale McAninch
Contribution to the Paper	Involved in data creation, provided critical discussion and intellectual input into the manuscript.
Signature	Date 14/7/17

Name of Co-Author	Prabha Andraweera		
Contribution to the Paper	Involved in data creation, provided critical discussion and intellectual input into the manuscript.		
Signature		Date	25-7-2017

Name of Co-Author	James Breen		
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.		
Signature		Date	20/7/2017

Name of Co-Author	Claire T Roberts		
Contribution to the Paper	Involved in the study design, provided critical discussion and intellectual input into the manuscript.		
Signature		Date	14.4.17

Name of Co-Author	Tina Bianco-Miotto		
Contribution to the Paper	Involved in the study design, provided critical discussion and intellectual input into the manuscript.		
Signature		Date	26/4/17

4 Identification of novel human placenta specific large intergenic non-coding RNAs using deeply sequenced RNA-seq from first trimester, term and preeclamptic pregnancies

BENJAMIN T MAYNE, SHALEM Y LEEMAQZ, DALE MCANINCH, PRABHA ANDRAWEERA, JAMES BREEN, CLAIRE T ROBERTS, TINA BIANCO-MIOTTO

Abstract

The placenta is the most poorly understood organ within the human body, however its critical role in pregnancy is unquestionable. Most studies that have focused on human placental gene expression have done so at term (≥ 37 weeks' gestation). This makes it difficult to infer gene expression characteristics of the placenta during its development. Moreover, placental expression studies have previously mainly focused on protein-coding genes and known annotated transcripts that have previously been identified. Due to the lack of high-throughput sequencing data carried out on the placenta compared to other tissues, there are still undiscovered "novel" transcripts that may play a crucial role in placental development, and in turn, have an immediate and life-long impact on the health of both mother and baby. In this study, we performed RNA sequencing (RNA-seq) on placental tissue samples from first trimester (n=30), uncomplicated term (n=22) and preeclamptic pregnancies (n=18) to characterise gene expression across gestation and in preeclampsia pathology. Unsupervised clustering revealed distinctively different transcriptomes across gestation time points, with genes expressed during first trimester being enriched for cancer related genes and cell growth, whereas placentas from uncomplicated term pregnancies were enriched for genes relating to inflammation processes. In addition, placentas from preeclamptic pregnancies were enriched for genes relating to

angiogenesis compared to those from uncomplicated term pregnancies. We also performed a de novo transcript analysis which identified 23 large intergenic non-coding RNAs (lincRNAs). The specific splice variants of these transcripts have not previously been annotated in any non-coding RNA databases. By comparing these transcripts with RNA-seq data from the NIH Epigenomics Roadmap, we were able to conclude these lincRNAs were placental tissue specific. Moreover, we compared these transcripts to novel lncRNAs in the FANTOM5 project and found 9 to overlap. However, these 9 transcripts were found to have splice variants variations specific to the placenta. Furthermore, 4 were found to be exclusively expressed during first trimester and 1 at term, 14 were differentially expressed across gestation, and 13 in placentas from preeclamptic pregnancies compared to those from uncomplicated term pregnancies. These analyses suggest these transcripts may potentially have roles in placental development and in preeclampsia. The identification of novel placental specific transcripts highlights the need for further research to completely characterise the placental gene expression profile and that unannotated genes within the placenta may potentially become targets of future research into pregnancy complications.

4.1 Introduction

The placenta remains the most poorly understood organ in the human body [1]. Although it has a short existence, it carries the task of providing protection from the maternal immune system, proper development to the growing fetus and the exchange of nutrients, gases and wastes. Impaired implantation and early placental morphogenesis during the early stages of development can lead to an increased risk of pregnancy complications such as preeclampsia (PE), preterm birth (PTB), fetal growth restriction (FGR) and miscarriage [2]. Many studies have established a disruption of normal placental gene expression in pregnancy complications [3-7]. However, these studies have been focused predominantly on placentas from term pregnancies.

Genome wide transcriptome analysis of the first trimester placenta has only been conducted by microarray studies [8, 9]. These have focused on differential gene expression analyses between gestational time points [8, 9]. It is worth mentioning that differential gene expression (36 genes) has been detected in placentas (10-12 weeks' gestation) from pregnant women who

were destined to develop PE compared to normal pregnancies [8]. Unfortunately, due to low sample sizes (4 PE, 8 controls), it is unlikely that the true extent of differentially expressed genes was captured. Furthermore, studies using RNA sequencing (RNA-seq) of placental tissue have also been limited to pregnancies at term or in late third trimester [3, 4, 10-12]. In addition, there are no reports that examine gene expression differences between first trimester and term pregnancies using a higher order systems level perspective such as a co-expression analysis. Moreover, RNA-seq studies have not sequenced deeply enough to detect many non-coding RNAs. For example, a recent RNA-seq study on placental tissue sequenced to an average depth of 22.3 million reads per sample [3]. However, other projects such as that from the FANTOM5 consortium have sequenced to a depth of approximately 200 million reads per sample in other human tissues, but not in the placenta [13]. Sequencing depth is important as deep sequencing is required to detect the expression of lowly expressed genes.

In this study, we used RNA sequencing (RNA-seq) to assemble the transcriptome of 30 first trimester, 22 term and 18 preeclamptic placentas. Our analyses identify distinct transcriptomes in placentas from first trimester, term and preeclamptic pregnancies. Furthermore, we have also identified novel large intergenic non-coding RNAs (lincRNA) specific to the human placenta.

4.2 Methods

4.2.1 Ethics statement

Written, informed consent was obtained from all patients prior to collection of placental tissue. First trimester placental tissue was collected from women undergoing elective terminations of pregnancy at the Women's and Children's Hospital North Adelaide, whereas term placental tissue samples were collected after delivery at the Lyell McEwin Hospital, South Australia. Collection of first trimester placental tissue was approved by the Women's and Children's Health Network Human Research Ethics Committee (REC2249/2/13) and the University of Adelaide Human Research Ethics Committee (H-137-2006).

Collection of term placental tissue was approved by the Queen Elizabeth Hospital and Lyell McEwin Hospital Human Research Ethics Committee (REC 1712/5/2008 and HREC/12/TQEHLMH/16).

4.2.2 RNA extraction and sequencing

First trimester placental villous tissue was collected from women who were undergoing elective termination (6-11 weeks' gestation). Placental villous tissue from both uncomplicated term and preeclamptic pregnancies were from either elective caesareans or vaginal deliveries (37-42 weeks' gestation). Placental villous tissues were incubated in RNAlater solution (Invitrogen) at 4˚C for 24 h prior to being stored at -80˚C. RNA was extracted from each placental villous tissue using TRIzol following the manufacturer's protocols. Sequencing libraries were prepared using Illumina TruSeq Stranded Total RNA Sample Preparation kits and all the ribosomal RNA was depleted using Ribo-Zero Gold. Sequencing was performed on the Illumina Hi-Seq 2500 using a 100bp pair-end protocol at the Queensland Brain Institute and at the Australian Cancer Genomic Facility in Adelaide.

4.2.3 Differential expression and co-expression analyses

Sequence adapters were trimmed from the fastq files using AdapterRemoval [14]. The trimmed RNA-seq reads were aligned to the human genome (UCSC hg38) using Bowtie 2 v2.1.0 and TopHat v2.0.9 [15]. Aligned read counts for each gene were determined using HTSeq v0.6.0 [16] with UCSC hg38 annotation. All differential expression analyses were performed using edgeR [17] and genes were considered significantly different if the false discovery rate (FDR) < 0.05. WGCNA was performed on all placental tissue samples with all expressed genes using the standard method as previously described [18, 19].

4.2.4 Detection of novel non-coding transcripts

For de novo transcript discovery, we used cufflinks v2.1.1. We selected the top 1% of highly expressed transcripts identified by cufflinks for further validation to reduce the possibility of falsely predicted transcripts. To determine if these transcripts have been previously annotated we compared the transcripts to three additional long non-coding RNA (lncRNA) databases (NONCODE [20], LNCipedia [21] and Gencode 25 [22]) using cuffcompare v2.1.1. As an additional in silico validation we used CPAT [23] to determine the coding potential of the novel transcripts. We also compared our analysis to the most recent FANTOM5 cap analysis [13] which contains data for novel lncRNAs. Overlap between FANTOM5 data and potential novel transcripts found in this study was compared using intersectBed function in BEDTools [24].

4.2.5 Gene Ontology

Genes that were upregulated in a group and gene lists for each module were tested for GO enrichment using DAVID [25]. The genes were compared to all human genes with the DAVID database.

4.3 Results

4.3.1 RNA-Seq data set

We performed RNA-seq (Illumina Hi-Seq 2500) on a total of 70 placentas from first trimester (n=30), uncomplicated term (n=22) and preeclamptic pregnancies (n=18) (Table 4-1). On average, 40.3 million 100-base pair, paired-end reads were aligned to the human genome (UCSC, hg38) with an alignment rate of 91.2%. This is approximately double the sequencing depth of a recent RNA-seq study on placental tissue [3]. This gives this study increased sensitivity to detect the expression of non-coding RNAs. The NIH Epigenomics Roadmap RNA-seq data generated on average between 30-50 million reads per sample [26] which is comparable to this study. This enables comparison of gene expression profiles between placenta and other human tissues. Furthermore, this also makes it a well suited comparison to determine the expression profiles of any potential novel transcripts found in this study. Genes with an expression of < 1 count per million (CPM) were removed from the analysis leaving a total of 16,481 genes from UCSC hg38 annotation. No statically significant differences were found in gestational age of delivery, birth length or birth weight between the uncomplicated term and preeclamptic pregnancy groups.

Table 4-1: Clinical characteristics of the participants in the study.

	First Trimester	Uncomplicated Term	Preeclampsia
N	30	22	18
Maternal Age (years)	NA	23	23
Average Gestation at collection (weeks) (mean, range)	8.4 (6-11)	40.5 (37-42)	38.6 (37-42)
Fetal Sex (F/M)	15/15	11/11	9/9
Birth Weight (g) (mean ± SD)	NA	3573 ± 377	3005 ± 647
Birth Length (cm) (mean ± SD)	NA	50 ± 2	47 ± 3

NA =not available

4.3.2 Clinical comparisons

In order to address potential confounding clinical variables we performed differential expression analyses between different clinical outcomes. Unfortunately, clinical details of women who were undergoing elective terminations were unavailable for this study. Therefore clinical comparisons were performed using placentas from term pregnancies where clinical details were ethically obtained. Previous studies have identified large sex differences in gene expression in human tissues [27] including the placenta [28]. In this study, we had a ratio of 1:1 males to females in all groups. This was done on purpose to reduce to any potential sex differences. We performed differential expression analyses between sexes in each group and only detected differentially expressed X and Y chromosome genes. Here we do not consider the Y chromosome genes as differentially expressed but included them in our analysis to determine if our analyses were correct. We found all differentially expressed Y chromosome genes to be upregulated in placentas from male fetuses. In comparison to a meta-analysis study which contained 303 samples from microarray studies [28], we did not detect large sex differences on autosomes. The only genes found to be differentially expressed between sexes were on the sex chromosomes such as *UTY*, a gene found on the Y chromosome. We also found *XIST*, *KDM5C*, which are on the X chromosome and were found to be upregulated in placentas from females in agreement with our previous study [28]. However, as suggested in a previous study [3], the lack of observable sex differences in gene expression was most likely the result of low sample sizes.

In our uncomplicated term pregnancy group, we tested for differential expression between the delivery modes (vaginal/caesarean section). During a vaginal birth the placenta can be hypoxic for a long period of time due to uterine contractions and time to delivery after detachment from the uterus which can result in gene expression changes compared to elective caesarean delivery [29]. However, we detected no significant differentially expressed genes between groups even at a relaxed threshold (FDR < 0.1). We also tested for gestational age differences in the first trimester group by splitting the

samples into either early (6-8 weeks, n=19) or late (10-11 weeks, n=10) gestation, however we detected no significantly differentially expressed genes between the groups. Rapid placental development and growth occurs during these weeks of gestation and what occurs early on in gestation may have implications later on in pregnancy. A limitation of this study is that the outcome of pregnancy for placentas collected during first trimester is unknown. This therefore may be a confounding factor within our analyses and may contribute to why no gene expression differences were observed between early and late first trimester. Our data suggest that clinical confounding factors such as fetal sex, mode of delivery and gestation differences during first trimester do not have a significant effect on gene expression.

4.3.3 Placental gene expression across gestation

Differential expression analyses were conducted between all three gestation and outcome groups in order to determine differences across gestation and to determine the extent of disrupted placental gene expression in PE. Since maternal age was available for the first trimester group we were unable to account for this in our analyses. This is a limitation of the study as this group may have been potential difference in age between the other groups. We firstly conducted our differential gene expression analysis between placentas from first trimester and uncomplicated term pregnancies. We detected differential expression of 7240 genes (FDR < 0.01) between first trimester and term using UCSC hg38 annotation (Supplementary Table 4-1).

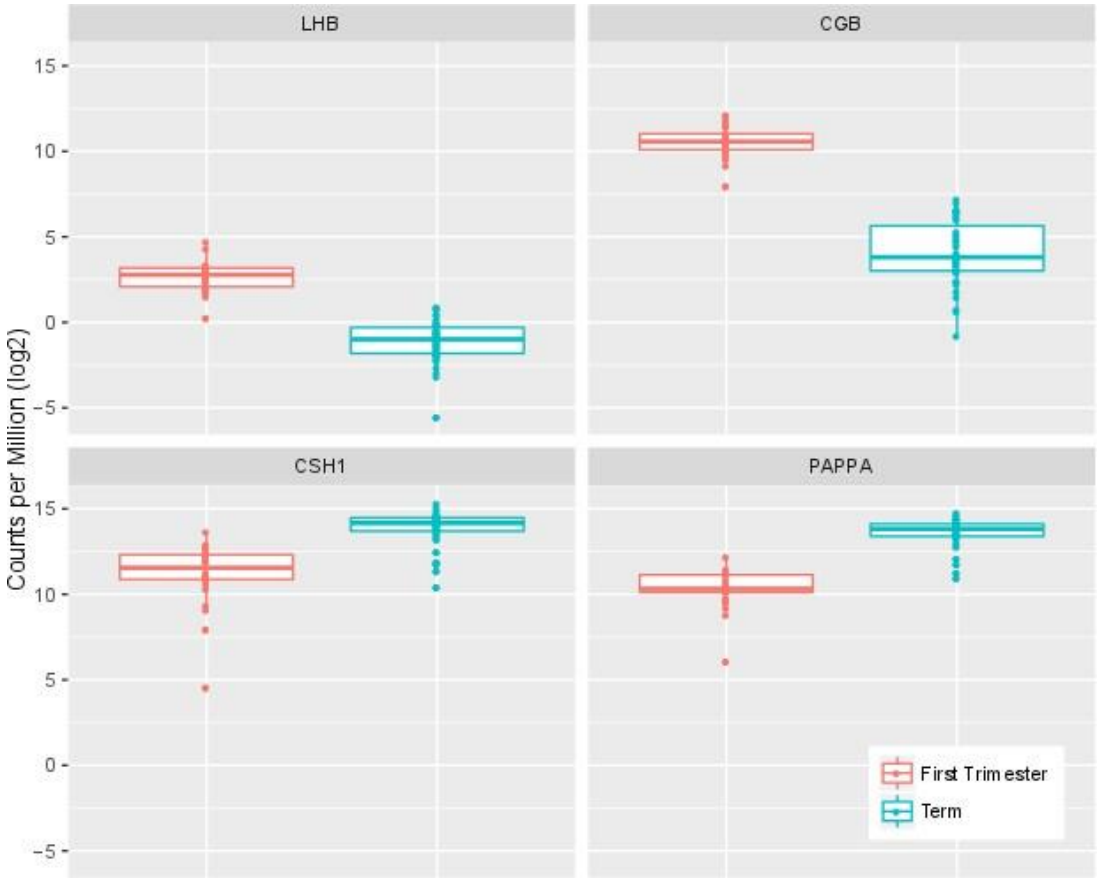


Figure 4-1: Differential placental gene expression of genes that encode for circulating maternal factors.

A. A volcano plot showing genes which were upregulated during first trimester (red dots) or at term (blue dots) in the placenta. **B**. Differential placental gene expression that encode for circulating maternal factors.

Since the placenta secretes factors such as proteins and hormones into the maternal bloodstream, we took a closer look at genes that encode proteins and hormones that are known to circulate in maternal blood to determine if their expression levels correlate with what is observed in the maternal blood. We found genes that encode for subunits for proteins such the beta subunit of luteinizing hormone (LH) and β-human chorionic gonadotrophin (β-hCG) which were LHB and CGB respectively. These genes were found to be more highly expressed during first trimester (Figure 4-1). This matches what is known to occur in the maternal blood where levels of β-hCG are known to decline across gestation [30]. Furthermore, we found placental lactogen (CSH1) and pregnancy associated plasma protein A (PAPPA) to be more highly expressed in placentas from uncomplicated term pregnancies compared to first trimester (Figure 4-1). Circulating levels of CSH1 [31] and PAPPA [32] in maternal blood are known to increase during gestation. Therefore, our placental gene expression data matches the published protein levels found in maternal blood. This provided confidence in the validity of our analyses.

The first trimester placental gene expression profile was distinctly different compared to the uncomplicated term and PE groups (Figure 4-2A). We detected a total of 7240 differentially expressed genes (FDR < 0.01) between first trimester (3481 upregulated) and term (3759 upregulated) but chose to focus on a subset of 703 genes (FDR < 0.01) with magnitude log2FC > 2 (Figure 4-2B), for downstream analyses such as GO term enrichment. We found 321 and 382 of these genes to be more highly expressed in placentas during first trimester and uncomplicated term pregnancies, respectively, compared to placentas from preeclamptic pregnancies. Genes upregulated in first trimester were enriched for GO terms such as mitosis (GO:0007067) and M phase (GO:0000279) similar to a cancer-like expression profile which has previously been shown [33]. For example, we found KIFC1, a kinesin motor protein which is associated with cancer [34, 35], to be upregulated in placentas from first trimester compared to uncomplicated term pregnancies (Supplementary Table 4-1). Conversely, genes upregulated at term were enriched for GO terms involved in inflammation such as the inflammatory response (GO:0006954) and response to wounding (GO:0009611). Some of these genes included IDO2 (Figure 4-2B, Supplementary Table 4-1) a gene associated with a pro-inflammatory response

[36, 37] which was found to be upregulated in placentas from term uncomplicated pregnancies compared to first trimester placenta. This analysis is consistent with the first trimester placenta being highly proliferative. However, it does suggest that term placentas are associated with a pro-inflammatory response. The pro-inflammatory profile from uncomplicated term pregnancies may be the result of the fact that these placentas have undergone stress during labour. This stress would result in an increase in hypoxia which may cause an increase in expression of genes relating to inflammatory processes. The uncomplicated term group was a mix of placentas from both labour and non-laboured pregnancies and which may have implications for this analysis. However, as detail prior in the previous section no differences were found in gene expression between placentas from elective caesareans or vaginal deliveries.

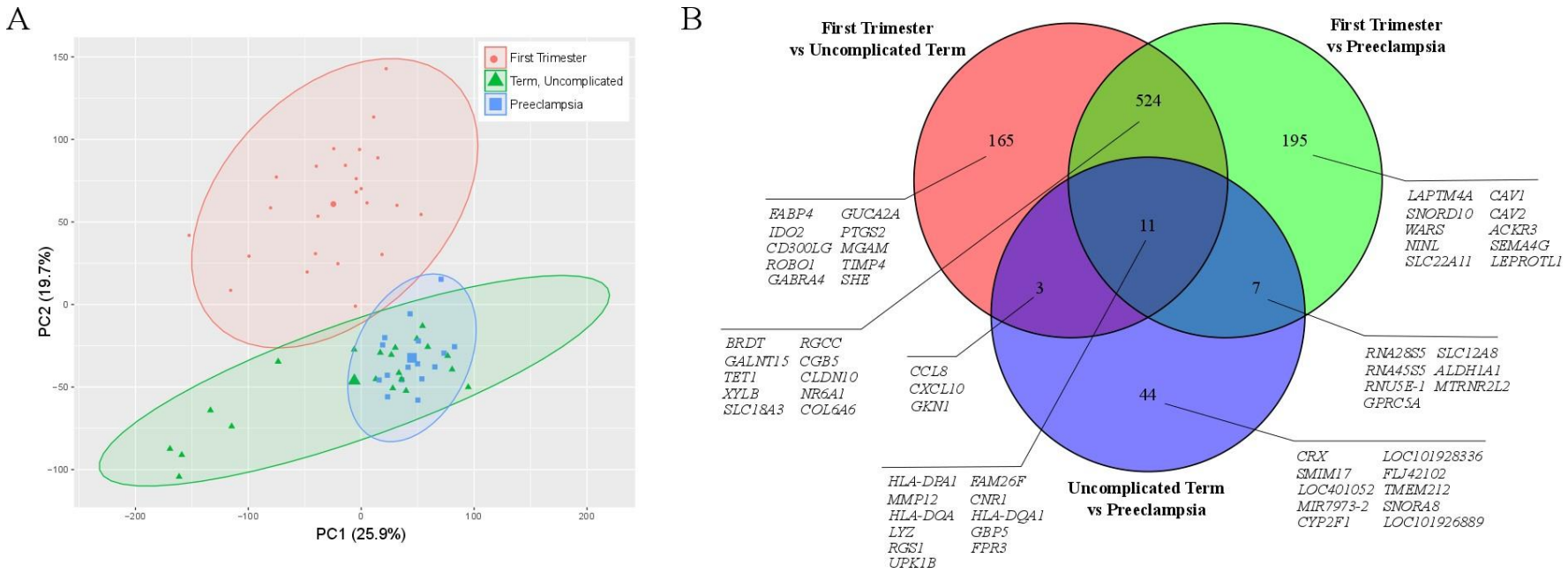


Figure 4-2: Differential gene expression analyses.

A. Principal component analysis of placental transcriptome. First trimester samples are represented by red circles, term uncomplicated samples by green triangles and preeclamptic samples with blue squares. **B.** Venn diagram showing the top significantly differentially expressed genes from the different placenta group comparisons.

4.3.4 Disruption of placental gene expression in preeclamptic pregnancies

Preeclampsia is a pregnancy complication characterised by high maternal blood pressure (> 140/90 mmHg) and proteinuria (24 hour, >0.3g) [38-40]. Previous studies have found that placental gene expression is disrupted in PE compared to uncomplicated term pregnancies [3, 4, 7]. However, no study to date has compared global placental gene expression between first trimester and preeclamptic pregnancies. Since the first trimester placenta grows and differentiates in a hypoxic environment until maternal blood flow is established to it beginning at 10-12 weeks' gestation and the PE placenta is thought to be hypoxic, it may be that the PE placenta is more similar to the first trimester placenta than to that from uncomplicated pregnancies. Furthermore, inadequate spiral artery modelling can result in inconsistent oxygen levels which can induce oxidative stress in PE.

We also compared placentas from PE and term uncomplicated pregnancies and found 4074 genes to be differentially expressed (FDR < 0.01) (Supplementary Table 4-3). We found 2267 and 1807 of these genes to be upregulated in placentas from uncomplicated term pregnancies and preeclamptic pregnancies, respectively. Some of these genes included a micro RNA precursor transcript (miRNA-7393-2) (Figure 4-2B) which suggests a disruption in ncRNA genes in PE. Overall we found genes upregulated in uncomplicated term compared to PE placentas to be enriched for genes relating to nucleotide binding (GO:0000166) and ATP binding (GO:0005524). These GO terms are consistent with the pathology of PE where an increase of oxidative stress can disrupt energy production.

We found 8780 genes (FDR < 0.01) to be differentially expressed between first trimester (5106 upregulated) and PE (3674 upregulated) (Supplementary Table 4-2). Again, we chose to focus on a subset of 737 genes (FDR < 0.01, magnitude log2FC > 2) (Figure 4-2B). 524 of the 737 genes overlapped with the analysis between first trimester and

uncomplicated term. This suggests a strong difference in placental gene expression across gestation which is maintained irrespective of the later gestation diagnosis of preeclampsia or uncomplicated pregnancy. More interestingly, 195 genes were found not to overlap with the previous comparison and were therefore unique to the first trimester and PE comparison (Figure 4-2B). These genes which have been subtracted from the other two analyses can be used to determine the true extent of gene expression differences between first trimester and PE. A closer inspection of these genes found that they were related to GO terms such as collagen trimer (GO:0005581) and extracellular region (GO:0005576). This suggests that there is a difference in genes relating to the cell surface which may be highlighting a difference in the gestational time points of the two groups.

Upregulated transcripts in PE placentas were enriched for genes relating to mitochondrion (GO:0005739) and generation of precursor metabolites and energy (GO:0006091). This suggests there may be a disruption of placental cellular energetics in PE. Defects in implantation and placental morphogenesis in early gestation are known to increase the risk of preeclampsia and other pregnancy complications [41, 42]. Impaired implantation and early placental development may restrict nutrient transfer and placental energetics later thereby disrupting placental gene expression. An increase in oxidative stress in PE may also explain the upregulation of genes relating to energy production processes. PE can develop due to a perfused placenta which can result in an increase of factors such as abnormal gene expression. Therefore the upregulation of genes relating to mitochondria and energy in PE, may be the result of a perfused placenta.

4.3.5 Long non-coding RNA expression in the placenta

To date, other studies that have focused on global placental gene expression have done so on protein coding genes. However, it is important to acknowledge that there are potentially thousands of non-coding RNAs which may be vitally important for proper placental development. To date, no such study has performed a de novo transcript analysis on placental tissue. This has been due to the lack of RNA-seq studies on

placental tissue. Therefore, there is a lack of knowledge surrounding placental lncRNAs which may have functions in regulating placental gene expression.

Here in this study we analysed lncRNA expression in all 3 groups. We used the NONCODE [20] data base to analyse expression, which contains genomic data for 144,157 transcripts. Unfortunately, it does not contain annotated functions for all transcripts since this data base has also performed a de novo transcript analysis using RNA-seq data. We found 34,717 transcripts to be expressed in the placenta and used these genes to perform differential expression analyses. We found 952 transcripts to be differentially expressed (FDR < 0.01, magnitude log2FC > 2) between placentas from first trimester (447 upregulated) and uncomplicated term (505 upregulated) pregnancies (Supplementary Table 4-4). Since there is little knowledge of functions of these transcripts in the literature, we accessed the supplied GO terms relating to each transcript from the NONCODE data base [20]. The potential functions of these transcripts are consistent with the protein coding and more well annotated genes identified above. For example, a transcript on chromosome 14 (NONHSAG015923.1) which was upregulated in first trimester (Supplementary Table 4-4) compared to term, is associated with GO terms such as positive regulation of cell proliferation (GO:0008284). Another transcript on chromosome 1 (NONHSAG003845.2) was more highly expressed in placenta from uncomplicated term pregnancies compared to first trimester (Supplementary Table 4-4) and is associated with the inflammatory response (GO:0006954). Our analysis shows that the expression profiles of lncRNA across gestation have similar functions to the protein coding genes. Therefore, these transcripts may have roles in regulating placental gene expression and development. They may be key regulators of proliferation during first trimester and involved in inflammatory processes later on in gestation.

We also analysed lncRNA expression in placentas from preeclamptic pregnancies. We identified 1154 transcripts to be differentially expressed between first trimester (519 upregulated) and PE (635 upregulated) (Supplementary Table 4-5). Interestingly, 466 of these transcripts were also found to be differentially expressed between the first trimester and uncomplicated term placentas suggesting that there is a strong difference in expression of non-coding RNAs across gestation. This is also consistent with the protein

coding and well annotated genes identified above, in that a large number of genes are differentially expressed across gestation irrespective of pregnancy outcome. However, we also found 688 transcripts differentially expressed between PE and first trimester but not with the uncomplicated term group. One of these genes, NONHSAG002977.2, is a transcript on chromosome 1 that was found to be more highly expressed in placentas from first trimester when compared to preeclamptic placentas (Supplementary Table 4-5). However, this transcript was not differentially expressed between first trimester and uncomplicated term placenta. This gene is associated with blood vessel remodelling (GO:0001974), an essential placental developmental process that proceeds across gestation [43] suggesting this gene may have an important role in regulating blood vessel development throughout gestation. This analysis highlights how proliferative the placenta is during first trimester. Proper spiral artery remodelling is essential for placental development and for successful pregnancy outcome. It is therefore not surprising to find lncRNAs associated with blood vessel development to be more highly expressed during first trimester. However, since it is downregulated in PE when compared to first trimester it may suggest impaired blood vessel remodelling in PE. We also compared placentas from uncomplicated term and preeclamptic pregnancies and found 248 transcripts to be differentially expressed (Supplementary Table 4-6). Of these 248 transcripts 141 were found to be uniquely associated with the uncomplicated and PE comparison. Furthermore, we found other transcripts including NONHSAG078935.1 and NONHSAG093772.1 which are on chromosomes 2 and 6, respectively, also relating to blood vessel development to be downregulated in preeclamptic placentas (Supplementary Table 4-6). This analysis suggests that there may be poor placental blood vessel development in placenta from preeclamptic pregnancies that may be associated with or potentially mediated by lncRNAs.

4.3.6 Identification of novel transcripts specific to the placenta

Since the placental transcriptome is not well characterised, we performed a de novo transcript analysis using all 70 samples from our RNA-seq data set. To reduce false positives, we selected the most highly expressed predicted novel transcripts (2767 transcripts) and compared these to three additional non-coding RNA databases

(NONCODE [20], LNCipedia [21] and Gencode v25 [22]). Thereby, we identified 26 transcripts that were not annotated in any of the non-coding RNA databases. We also performed an in silico validation of the coding potential of the novel transcripts using CPAT. Three of the transcripts were classified as coding and were removed from further downstream analyses, which left 23 remaining novel non-coding RNAs (Supplementary Table 4-7). These 23 novel non-coding RNAs were all >200nt and were within intergenic regions of the genome and therefore would be classified as large intergenic non-coding RNAs (lincRNAs).

We further analysed and made use of publicly available RNA-seq data to determine if the novel lincRNAs were specific to the placenta. We used RNA-seq data generated from the NIH Epigenomics Roadmap Project [26] to determine if the novel lincRNAs were expressed in other human tissues or cells. Using the same methods for alignment and generation of read counts (refer to methods) we analysed data from 198 RNA-seq samples from the NIH Epigenomics Roadmap Project. We did not find expression of any of the novel lincRNAs in any other normal human tissue or cell. This analysis suggest these transcripts are specific to the human placenta.

We combined the 23 novel lincRNAs with our RNA-seq analysis and conducted differential expression analysis to determine if their expression levels alter across gestation and/or in PE. 14 novel lincRNAs were differentially expressed between first trimester and term. In addition, four of the novel lincRNAs were only expressed during first trimester. Moreover, we found one novel lincRNA to be expressed at term but not in first trimester. Some of these transcripts were also found to be in potential clusters (Figure 4-3). For example, Figure 4-3 shows a potential cluster of novel lincRNAs on chromosome 9 which were exclusively expressed during first trimester. A zoomed view of this novel lincRNA (Figure 4-3A) shows it is a bi-exonic placental transcript which spans 20,501bp which is expressed during first trimester (Figure 4-3B). In addition, we confirmed expression of these novel transcripts by qPCR in an independent set of placentas (Figure 4-3C). This extra validation verifies that the novel transcripts found in

this study do exist and can be detected in other placentas from pregnancies outside of the RNA-seq cohort.

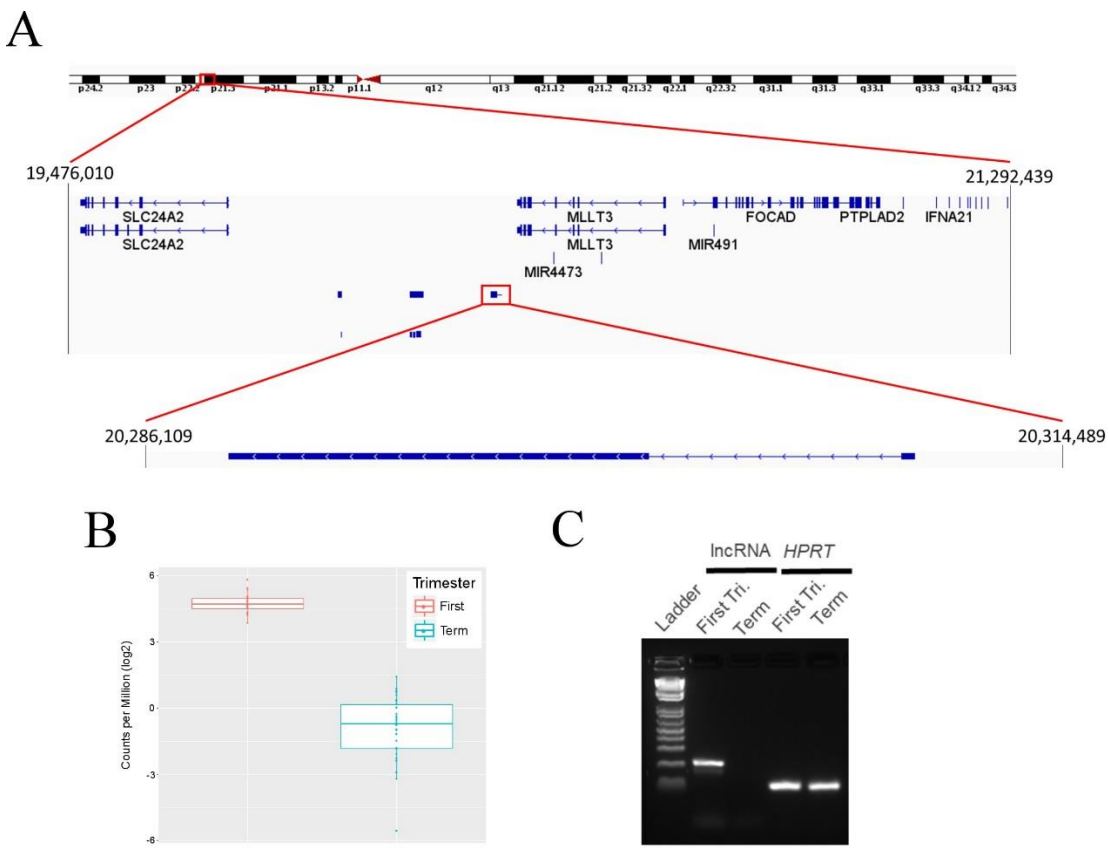


Figure 4-3: A cluster of novel large intergenic non-coding RNAs specific to first trimester placentas on chromosome 9 identified by RNA sequencing.

A. A region of focus on chromosome 9 containing first trimester specific novel large intergenic non-coding RNAs. A bi-exonic transcript which is exclusively expressed in first trimester shown by **B.** RNA sequencing and confirmed by **C.** qPCR. *HPRT* is used as a reference gene.

Since we found a large difference in lncRNA expression in placentas from PE pregnancies, we also performed differential expression analyses of the novel lincRNAs in placentas from preeclamptic pregnancies. We found 14 of the novel lincRNAs to be differentially expressed between placentas from first trimester and PE. Interestingly, these 14 novel lincRNAs were also differentially expressed between the first trimester and uncomplicated term placenta. This suggests the expression profiles of these 14 novel lincRNAs are potentially maintained across gestation irrespective of pregnancy outcome. We also found 13 novel lincRNAs to be differentially expressed between placentas from uncomplicated term and preeclamptic pregnancies. This suggests there is a disruption of their gene expression profiles in PE.

4.3.7 FANTOM5 comparison

The FANTOM5 project is a consortium that has focused on the identification of novel lncRNAs from a diverse range of cell types [13]. This project used RNA-seq to a depth of approximately 200 million reads per sample, which is 5 times deeper than the sequencing used in this project. Therefore, this is a valuable resource to determine if any of the potential novel lincRNAs found in this study are truly placenta specific. Since FANTOM5 only contains genomic data for hg19 annotation, for this comparison only, we converted the coordinates in our study from hg38 to hg19. Using BEDTools we found overlapping genomic coordinates for 9 of the potential 23 novel lincRNAs found in this study (Supplementary Table 4-8). Therefore, these 9 lincRNAs that were found to overlap in some form may be expressed in other human tissues. Since they were not shown to be expressed in any other human tissues in the NIH Epigenomics Roadmap it suggests that these lincRNAs are lowly expressed. The NIH Epigenomics Roadmap RNA-seq data sequenced to a depth between 30-50 million reads per sample [26], which is within the range of that performed in this study. This suggests these 9 lincRNAs that overlapped with the FANTOM5 data may be expressed in tissues other than the placenta but at a much lower level. This shows the importance of sequencing sufficiently deeply to quantify expression of lowly expressed non-coding RNAs.

Moreover, as shown in Supplementary Table 4-8, none of the splice variants of the 9 overlapping lincRNAs were found in the FANTOM5 data. This suggests that these transcript splice variants are placenta specific but other splice variants of these genes are transcribed in other human tissues. Unfortunately, since these lincRNAs are novel and there is currently no functional annotation of these transcripts, it is difficult to draw conclusions regarding their role in the placenta. In addition, the overlap between novel transcripts in this study and FANTOM5 did not contain the full transcript. For example, Figure 4-4 shows the overlap between XLOC_004232 and the first exon of a novel lncRNA identified by FANTOM5. However, as shown in Figure 4-4, it is only the 3' end of XLOC_004232 that overlaps with that found in FANTOM5 and the rest of the transcript was unique to this study. This was the case with most of the overlaps between this study and FANTOM5 (Supplementary Table 4-8) in that it was only a small fraction of the transcripts that overlapped. This highlights the novelty of the transcripts identified in this study and suggests that they are highly likely to be exclusive to placental tissue.

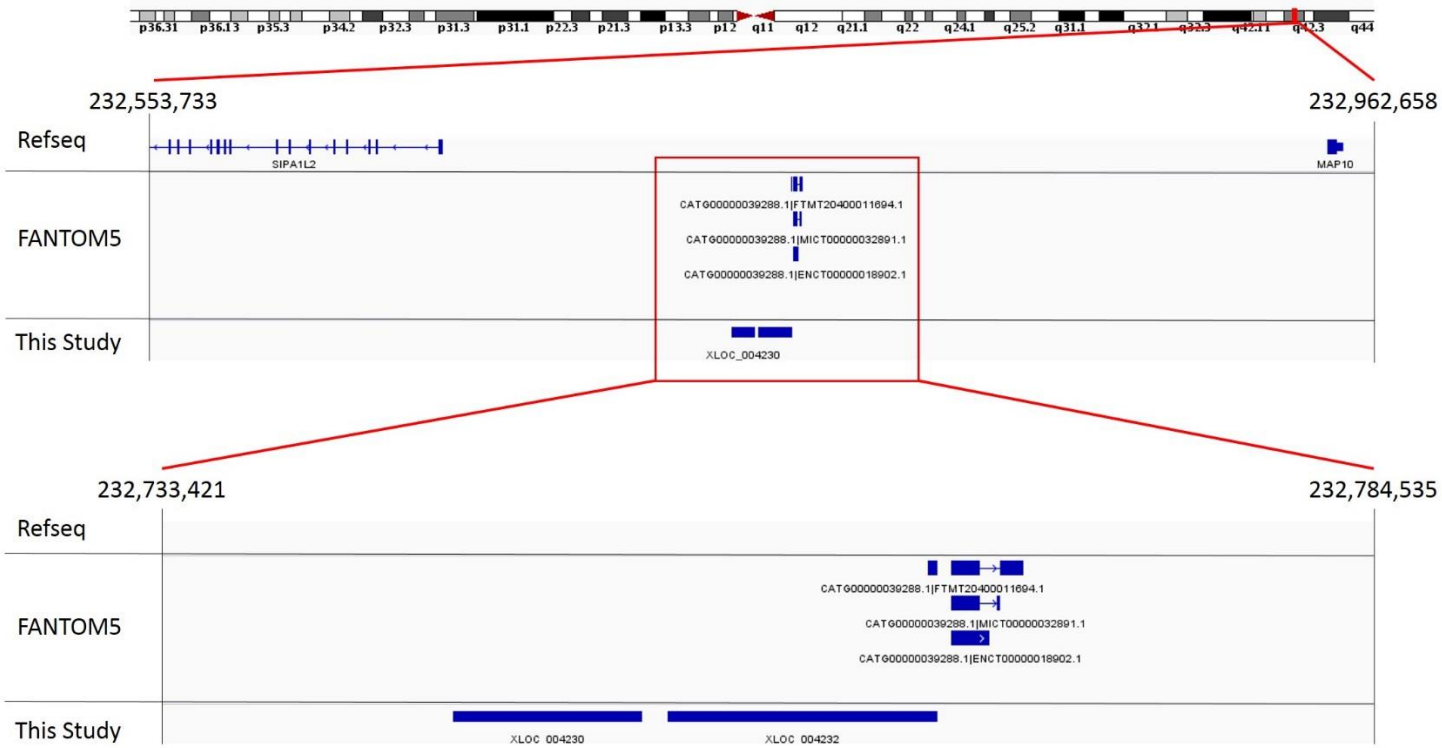


Figure 4-4: Overlap of novel lncRNAs between FANTOM5 and this study.

The 3' end of XLOC_004232, a mono-exonic lincRNA found in this study was found to overlap with the first exon of a novel lncRNA found in FANTOM5. Note these coordinates have been converted to hg19.

4.3.8 Characterisation of novel large intergenic non-coding RNAs and co-expression of lincRNA with other protein-coding genes

Many of the novel lincRNAs identified in this study were either differentially expressed across gestation or had an altered expression in placentas from preeclamptic compared to uncomplicated pregnancy. Therefore, these transcripts are likely to have unique roles in placental development, so we decided to characterise the lincRNAs by in silico analyses to provide insight into their biological roles. We firstly determined if the novel lincRNA locations overlap with any known disease associated single nucleotide polymorphisms (SNPs). We compared the National Human Genome Research Institute (NHGRI) Catalogue of disease associated SNPs from Genome-Wide Association Studies (GWAS) [44] to the locations of the novel lincRNAs. However, none of these overlapped with any of the novel lincRNA locations. We then used GREAT, which is an online tool that returns GO terms based on the genomic location and neighbouring genes. However, this resulted in no enrichment (FDR < 0.05) for GO terms for any of the 23 lincRNAs.

Previous studies have commonly used correlation based techniques to determine the function of lncRNAs such as co-expression based methods [45-47]. In this study, we used a weighted gene co-expression network analysis (WGCNA) [18, 19] which identifies modules of genes with similar expression patterns. This method can be a useful strategy to identify novel transcripts with a high module membership and to provide insight into their possible biological function. Using expression data from all 70 placentas, we identified a total of 18 modules (Figure 4-5A). We summarised the overall expression of each sample then calculating the first principle component of gene expression in each module. GO term enrichment was also performed on each module which simplified analyses by summarising genes into GO terms. We found many co-expression modules to be differentially expressed across gestation (Figure 4-5B). However, these modules mostly contained genes that were found to be differentially expressed in the earlier comparisons. For example, modules containing genes relating to mitosis (GO:0007067) were found to be differentially expressed between placentas from first trimester and uncomplicated term pregnancies. Overall, co-expression analysis reiterated the

differential expression analyses. However, the goal of co-expression analysis here was to identify possible biological pathways involving the novel lincRNAs.

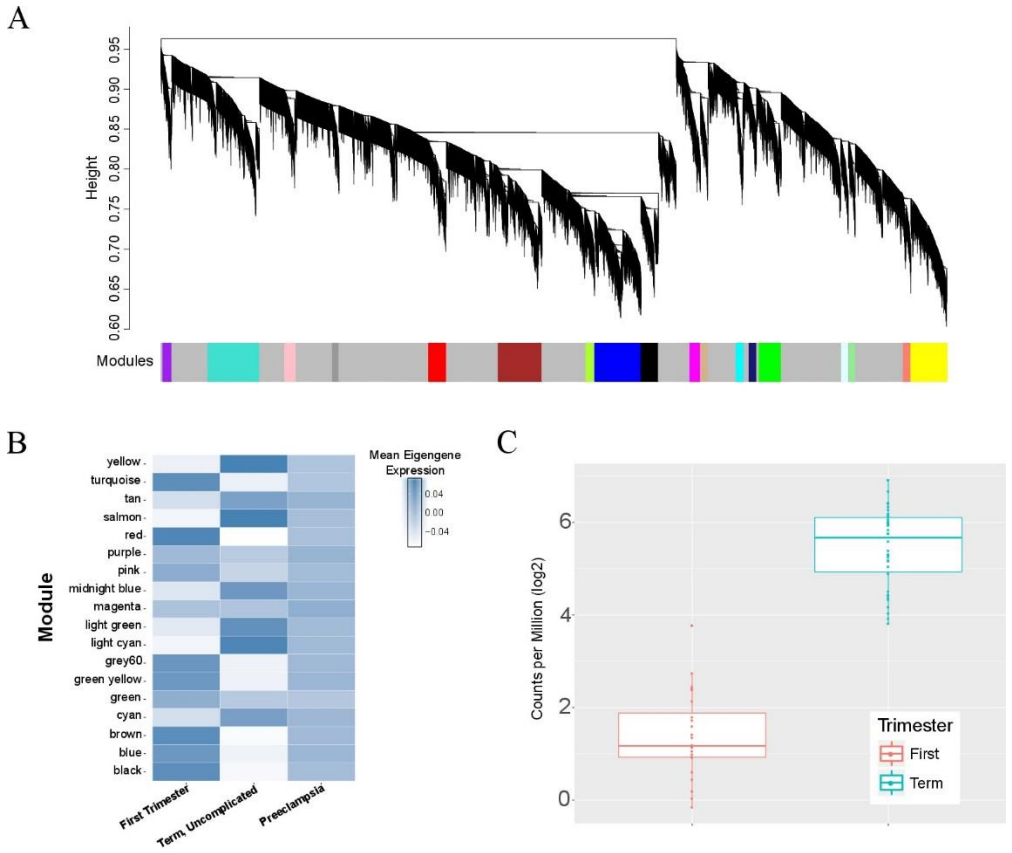


Figure 4-5: Co-expression analysis of the placental transcriptome.

A. Dendrogram of placental gene expression. Each line represents an individual gene expressed in the placenta and these have been grouped into modules. Modules are represented as non-grey colours below the dendrogram. B. Heatmap of the mean eigengene expression of the 18 modules. C. Differential expression of novel transcript on chromosome 7 to be grouped within the midnightblue module.

Unfortunately, 22 of the 23 novel lincRNAs did not share co-expression patterns with other expressed placenta protein-coding genes and therefore were not classified into any of the modules. However, one novel lincRNA located on the p-arm of chromosome 7 was classified into a module that was enriched for GO terms relating to homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156) and angiogenesis (GO:0001525). The transcript was also found to be upregulated in placentas from uncomplicated term pregnancies compared to first trimester (Figure 4-5C). This transcript may have a potential role in placental development by acting through angiogenesis pathways.

4.4 Discussion

Here we have characterised the expression profiles of both protein- and non-coding genes in placentas that were from either first trimester pregnancies or those that had reached term and were uncomplicated or preeclamptic pregnancies. We are the first to use RNA-seq on first trimester placentas, which have been difficult to sample for experimental and ethical reasons. Thus, our study has identified placenta specific novel lincRNAs which likely have potential roles in placental development.

We compared placental gene expression between first trimester and uncomplicated term pregnancies and found placental gene expression consistent with reported protein levels in the maternal circulation. For example, genes within the LHB-CGB cluster on chromosome 19 were expressed more highly during first trimester compared to term. CGB genes encode for βhCG a protein essential for establishment of pregnancy and used as a diagnostic marker of pregnancy [30]. β-hCG is found at its highest levels during first trimester [30] consistent with placental RNA-seq data showing genes in the LHB-CGB cluster to be expressed more highly during first trimester. This was also the case with other genes such as that encoding placental lactogen (CSH1). Placental lactogen is a growth hormone gene family member in the growth hormone locus on chromosome 17 and is required for maternal metabolic adaptation to pregnancy and glucose tolerance to meet fetal energy

demand [48]. Maternal blood levels of placental lactogen steadily increase throughout gestation, which again matched with our observable gene expression data.

Our differential gene expression analysis indicates that, the first trimester placental transcriptome harbours many hallmarks of cancer due to the high number of genes relating to proliferation processes compared to term placenta. Not only did we identify genes upregulated in first trimester to be enriched for GO terms relating to mitosis, but also many genes associated with cancer. For example, INHA, a gene derived from placental trophoblast which is a member of the TGF-β superfamily [49, 50] and a marker of ovarian cancer [51] was upregulated during first trimester compared to term placenta (Supplementary Table 4-1). This analysis suggests that a first trimester placenta is an organ of high proliferation which is required for rapid growth for developing fetus. Using genes that are commonly seen upregulated in a cancer simply shows the high levels of cell growth during the first trimester of pregnancy.

We also performed differential expression analysis on placentas from preeclamptic pregnancies and found 524 differentially expressed genes between first trimester and either uncomplicated term or preeclamptic pregnancies when compared to (Figure 4-2B). This large proportion of genes whose expression is maintained irrespective of pregnancy outcome, suggests strong differences in expression profiles across gestation. Some of these genes included CGB5 (Figure 4-2B, Supplementary Table 4-1, 4-2) which as discussed above is part of the LHB-CGB cluster and is expressed more highly during first trimester. More interestingly, we found 195 genes (Figure 4-2B) the expression of which was not maintained between comparisons. GO term analysis of these genes suggested a disruption in genes involved in cell adhesion in placentas from preeclamptic pregnancies. We also performed differential expression analysis between placentas from either uncomplicated term or preeclamptic pregnancies. GO term analysis of these genes suggested a disruption of energy metabolism pathways in PE. Since impaired implantation and placental trophoblast invasion of the maternal vasculature is associated with PE disruption of expression of genes relating to cell adhesion may potentially provide insight into the

molecular mechanisms associated with PE [41, 42, 52, 53]. Furthermore, since the preeclamptic placenta may not necessarily be obtaining sufficient oxygen and micronutrients from maternal blood, this may lead to disruptions in energy metabolism. In other words, the expression profiles of the preeclamptic placenta shown herein may reflect a placenta that did not undergo proper implantation and morphogenesis during the early stages of development.

This study was also the first, to our knowledge, to characterise global lncRNA expression profiles of placentas from first trimester, uncomplicated term or preeclamptic pregnancies. The transcriptome of placentas in this study included the expression profiles of 34,717 lncRNAs. Our differential expression analyses showed that those differentially expressed between placenta from first trimester and uncomplicated term pregnancies had similar patterns of expression to those of protein coding genes. For example, we found lncRNAs relating to enhanced cell proliferation (Supplementary Table 4-4) were more highly expressed during first trimester. This reiterates the notion that the transcriptome of the first trimester placenta resembles that of a highly proliferative cancer. Compared to first trimester placenta, we found lncRNAs that were more highly expressed in placentas from uncomplicated term pregnancies to be associated with the inflammatory response. This too reiterates the notion that a placenta at term is undergoing an inflammatory response. This also suggests that lncRNAs associate with important cellular and molecular processes in the placenta across gestation. However, unlike protein coding genes, the functions of lncRNAs are yet to be characterised. Therefore, there may be a variety of yet to be determined roles that these lncRNAs play in the placenta. These lncRNAs could be potentially characterised by performing targeted experiments such as knocking them out individually. This would allow determination if they are biologically important in the context of a placenta. Furthermore, by analysing what is dysregulated once knocked out would determine important and associated molecular pathways. In addition, this study has used placental villous tissue, a heterogeneous sample which contains multiple cell types. This is a limitation of the study as it is unknown what specific cell types express the lncRNAs. Future experiments would firstly have to characterise the expression profiles of

different placental cell types to determine which cell types are expressing specific lncRNAs. This would allow for better characterisation of the lncRNA when performing knock out experiments.

We also undertook lncRNA expression analysis using placentas from preeclamptic pregnancies. Over 40% of the lncRNAs found to be differentially expressed between placentas from first trimester and preeclamptic pregnancies were also differentially expressed between first trimester and placenta from uncomplicated pregnancies. Despite a different pregnancy outcome at term these data suggest a strong difference in expression profiles across gestation. However, this also shows that 60% of the lncRNAs were not maintained across comparisons. We had a closer look at these transcripts and found enrichment for GO terms relating to blood vessel development. These lncRNAs specifically were downregulated in placentas from preeclamptic pregnancies compared to first trimester placenta suggesting disrupted expression of lncRNAs related to blood vessel development in preeclampsia. For example, NONHSAG002977.2 is a transcript known to be associated with blood vessel remodelling and was downregulated in PE compared to first trimester. Furthermore, this transcript was enriched for the GO term titled ‘blood vessel endothelial cell proliferation involved in sprouting angiogenesis’. This suggests that this transcript may be important in regulating angiogenesis and possibly spiral artery remodelling. To the best of our knowledge, due to the lack of literature and data surrounding these lncRNAs, this was the only transcript known relating blood vessel remodelling to be differentially expressed. Therefore, future work should focus on characterising other lncRNAs as they may important roles in PE. Spiral artery remodelling is an important process in placental development when insufficient can increase risk of PE [52, 53]. The lncRNAs found to be downregulated herein may underpin some of the mechanisms involved in spiral artery remodelling during placental development and when disrupted they may be associated with poor spiral artery remodelling resulting in PE. This is important as it may potentially identify therapeutic targets for prevention of PE.

In addition, to conducting differential expression analyses we performed a de novo transcript analysis which identified 23 previously unannotated transcripts specific to placental tissue. Four of these transcripts were only expressed during first trimester while 1 was found to be expressed in placentas at term. The remaining 18 transcripts had similar placental expression across gestation and in both uncomplicated and preeclamptic pregnancies. RNA-seq data from the NIH Epigenomics Roadmap from 198 human tissue samples indicated that these transcripts were placenta specific. However, 9 of the 23 transcripts found in this study were shown to overlap with those identified by the FANTOM5 project. However, as shown in Figure 4-4 and Supplementary Table 4-8, the overlap occurred in only a small fraction of the transcripts in this study. Despite FANTOM5 sequencing 5 times deeper than that performed in this study, the full length transcripts were not found suggesting the transcript variants identified in this study are exclusively expressed in placental tissue. Furthermore, 14 were found to be differentially expressed in placenta from first trimester and uncomplicated term pregnancies. In addition, 13 were differentially expressed between placentas from preeclamptic and uncomplicated term pregnancies. Due to their specificity to the placenta and expression changes across gestation and in PE, these novel lincRNAs may have a functional role in placental development. *In silico* analyses on their biological function was limited due to the lack of available data regarding their genomic positions. For example, we found no disease associated SNPs to overlap with genomic regions of the lincRNAs. In addition, GO term analyses based on their surrounding genes returned no significant enrichment terms. The lack of data about their biological function is partly due to their location. This is one of the limitations regarding complex genomic work as undertaken in this study. Although new technology has enabled the identification of these transcripts, further work is required to determine their biological roles and whether or not they have important roles in placental development or PE. Despite not finding much information on their biological function for most of the novel lincRNAs, we did find one that is possibly related to cell adhesion and angiogenesis. Co-expression analysis suggests one of the transcripts located on chromosome 7 to be involved in genes relating to GO terms such as angiogenesis (GO:0001525). Furthermore, this transcript was also upregulated in placenta from uncomplicated term pregnancies compared to during first trimester. This may suggest that

this transcript is involved in angiogenesis pathways which are essential for proper fetal growth and development. However, the precise mechanisms of action for this transcript remain to be elucidated.

The RNA-seq protocol used in this study returned on average 40.3 million reads per sample, which was sufficient to identify 23 previously unannotated transcripts. Some of these transcripts, such as in Figure 4-3, appear to be within clusters. RNA-seq is becoming increasingly more sensitive and able to sequence to a greater depth so future RNA-seq studies on placental tissue will be able to determine if these clusters of novel transcripts are in fact clusters or potentially long transcripts. Hence, this study highlights that continued research in this area is needed. The identification of the novel lincRNAs has potentially only scratched the surface of unannotated transcripts within placental tissue. Using RNA-seq to sequence to a greater depth is likely to reveal more transcripts. This may also uncover transcripts that may become targets in future therapeutic development in pregnancy complication research.

4.5 Conclusion

This study is the first to use RNA-seq on first trimester placental tissue and to characterise lncRNA expression. It has identified unique transcriptomes in placenta from first trimester, term and PE pregnancies. Consequently, we now have a comprehensive list of differentially expressed genes between the three comparisons of placental tissue analysed in this study. Furthermore, our analyses of the placental expression profiles of lncRNAs from first trimester, uncomplicated term and preeclamptic pregnancies has revealed a downregulation of lncRNAs associated with blood vessel development in PE. Apart from analysing known gene expression, this study went a step further to identify previously unannotated transcripts. In total, we identified 23 novel lincRNAs in placental tissue. Using publicly available data from the NIH Epigenomics Roadmap were specific to placental tissue, although comparison with FANTOM5 data suggests some may be very lowly expressed in other tissues. Furthermore, due to the differential expression of these

transcripts between pregnancy group comparisons, these novel transcripts may have potential roles in placental development or the onset of PE.

This study has uncovered new placenta specific transcripts which has identified new targets for placental development research. Future work is required to determine their exact biological roles which may be associated with normal and aberrant placental development. Moreover, RNA-seq analysis to a greater depth is required to potentially identify more novel transcripts which may have biological functions in a successful pregnancy outcome.

4.6 Supporting Information

For Supplementary Table 4-1, 4-2, 4-3, 4-4, 4-5, 4-6, 4-7 and 4-8 please refer to the electronic supporting information.

Supplementary Table 4-1: Genes differentially expressed between first trimester and term placentas.

Supplementary Table 4-2: Genes differentially expressed between first trimester and preeclamptic placentas.

Supplementary Table 4-3: Genes differentially expressed between term and preeclamptic placentas.

Supplementary Table 4-4: Long non-coding RNAs differentially expressed in the placenta between first trimester and uncomplicated term pregnancies.

Supplementary Table 4-5: Long non-coding RNAs differentially expressed in the placenta between first trimester and preeclamptic pregnancies.

Supplementary Table 4-6: Long non-coding RNAs differentially expressed in the placenta between uncomplicated term and preeclamptic pregnancies.

Supplementary Table 4-7: Genomic information of all variations of the 23 novel large intergenic non-coding RNAs.

Supplementary Table 4-8: Overlap between novel lncRNAs found in FANTOM5 and this study.

Note: Genomic coordinates have been converted to hg19 for this analysis.

References

1. Burton GJ, Fowden AL: **The placenta: a multifaceted, transient organ.** *Philos Trans R Soc Lond B Biol Sci* 2015, **370:**20140066.
2. Roberts CT: **IFPA Award in Placentology Lecture: Complicated interactions between genes and the environment in placentation, pregnancy outcome and long term health.** *Placenta* 2010, **31 Suppl:**S47-53.
3. Sober S, Reiman M, Kikas T, Rull K, Inno R, Vaas P, Teesalu P, Marti JM, Mattila P, Laan M: **Extensive shift in placental transcriptome profile in preeclampsia and placental origin of adverse pregnancy outcomes.** *Sci Rep* 2015, **5:**13336.
4. Kaartokallio T, Cervera A, Kyllonen A, Laivuori K: **Gene expression profiling of pre-eclamptic placentae by RNA sequencing.** *Sci Rep* 2015, **5:**14107.
5. Guo L, Tsai SQ, Hardison NE, James AH, Motsinger-Reif AA, Thames B, Stone EA, Deng C, Piedrahita JA: **Differentially expressed microRNAs and affected biological pathways revealed by modulated modularity clustering (MMC) analysis of human preeclamptic and IUGR placentas.** *Placenta* 2013, **34:**599-605.
6. Nishizawa H, Ota S, Suzuki M, Kato T, Sekiya T, Kurahashi H, Udagawa Y: **Comparative gene expression profiling of placentas from patients with severe pre-eclampsia and unexplained fetal growth restriction.** *Reprod Biol Endocrinol* 2011, **9:**107.
7. Sitras V, Paulssen RH, Gronaas H, Leirvik J, Hanssen TA, Vartun A, Acharya G: **Differential placental gene expression in severe preeclampsia.** *Placenta* 2009, **30:**424-433.
8. Founds SA, Conley YP, Lyons-Weiler JF, Jeyabalan A, Hogge WA, Conrad KP: **Altered global gene expression in first trimester placentas of women destined to develop preeclampsia.** *Placenta* 2009, **30:**15-24.
9. Uuskula L, Mannik J, Rull K, Minajeva A, Koks S, Vaas P, Teesalu P, Reimand J, Laan M: **Mid-gestational gene expression profile in placenta and link to pregnancy complications.** *PLoS One* 2012, **7:**e49248.
10. Kim J, Zhao K, Jiang P, Lu Z, Wang J, Murray JC, Xing Y: **Transcriptome landscape of the human placenta.** *BMC Genomics* 2012, **13:**115.
11. Saben J, Zhong Y, McKelvey S, Dajani NK, Andres A, Badger TM, Gomez-Acevedo H, Shankar K: **A comprehensive analysis of the human placenta transcriptome.** *Placenta* 2014, **35:**125-131.
12. Saben J, Lindsey F, Zhong Y, Thakali K, Badger TM, Andres A, Gomez-Acevedo H, Shankar K: **Maternal obesity is associated with a lipotoxic placental environment.** *Placenta* 2014, **35:**171-177.
13. Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al: **An atlas of human long non-coding RNAs with accurate 5' ends.** *Nature* 2017, **543:**199-204.
14. Lindgreen S: **AdapterRemoval: easy cleaning of next-generation sequencing reads.** *BMC Res Notes* 2012, **5:**337.

15. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14:**R36.
16. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31:**166-169.
17. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
18. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Stat Appl Genet Mol Biol* 2005, **4:**Article17.
19. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics* 2008, **9:**559.
20. Bu D, Yu K, Sun S, Xie C, Skogerbø G, Miao R, Xiao H, Liao Q, Luo H, Zhao G, et al: **NONCODE v3.0: integrative annotation of long noncoding RNAs.** *Nucleic Acids Res* 2012, **40:**D210-D215.
21. Volders PJ, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, Mestdagh P: **An update on LNCipedia: a database for annotated human lncRNA sequences.** *Nucleic Acids Res* 2015, **43:**D174-180.
22. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22:**1760-1774.
23. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W: **CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model.** *Nucleic Acids Res* 2013, **41:**e74-e74.
24. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics* 2014, **47:**11.12.11-11.12.34.
25. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44-57.
26. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518:**317-330.
27. Mayne BT, Bianco-Miotto T, Buckberry S, Breen J, Clifton V, Shoubridge C, Roberts CT: **Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans.** *Frontiers in Genetics* 2016, **7**.
28. Buckberry S, Bianco-Miotto T, Bent SJ, Dekker GA, Roberts CT: **Integrative transcriptome meta-analysis reveals widespread sex-biased gene expression at the human fetal-maternal interface.** *Mol Hum Reprod* 2014, **20:**810-819.
29. Janssen AB, Tunster SJ, Savory N, Holmes A, Beasley J, Parveen SAR, Penketh RJA, John RM: **Placental expression of imprinted genes varies with sampling site and mode of delivery.** *Placenta* 2015, **36:**790-795.
30. Korevaar TIM, Steegers EAP, de Rijke YB, Schalekamp-Timmermans S, Visser WE, Hofman A, Jaddoe VWV, Tiemeier H, Visser TJ, Medici M, Peeters RP:

Reference ranges and determinants of total hCG levels during pregnancy: the Generation R Study. *Eur J Epidemiol* 2015, **30:**1057-1066.
31. Chellakooty M, Skibsted L, Skouby SO, Andersson AM, Petersen JH, Main KM, Skakkebaek NE, Juul A: **Longitudinal study of serum placental GH in 455 normal pregnancies: correlation to gestational age, fetal gender, and weight.** *J Clin Endocrinol Metab* 2002, **87:**2734-2739.
32. Kalousova M, Muravska A, Zima T: **Pregnancy-associated plasma protein A (PAPP-A) and preeclampsia.** *Adv Clin Chem* 2014, **63:**169-209.
33. Holtan SG, Creedon DJ, Haluska P, Markovic SN: **Cancer and Pregnancy: Parallels in Growth, Invasion, and Immune Modulation and Implications for Cancer Therapeutic Agents.** *Mayo Clin Proc* 2009, **84:**985-1000.
34. Xiao YX, Yang WX: **KIFC1: a promising chemotherapy target for cancer treatment?** *Oncotarget* 2016, **7:**48656-48670.
35. Li Y, Lu W, Chen D, Boohaker RJ, Zhai L, Padmalayam I, Wennerberg K, Xu B, Zhang W: **KIFC1 is a novel potential therapeutic target for breast cancer.** *Cancer Biol Ther* 2015, **16:**1316-1322.
36. Merlo LMF, Mandik-Nayak L: **IDO2: A Pathogenic Mediator of Inflammatory Autoimmunity.** *Clinical Medicine Insights Pathology* 2016, **9:**21-28.
37. Prendergast GC, Metz R, Muller AJ, Merlo LMF, Mandik-Nayak L: **IDO2 in Immunomodulation and Autoimmune Disease.** *Front Immunol* 2014, **5:**585.
38. Uzan J, Carbonnel M, Piconne O, Asmar R, Ayoubi J-M: **Pre-eclampsia: pathophysiology, diagnosis, and management.** *Vascular Health and Risk Management* 2011, **7:**467-474.
39. Mol BW, Roberts CT, Thangaratinam S, Magee LA, de Groot CJ, Hofmeyr GJ: **Pre-eclampsia.** *Lancet* 2016, **387:**999-1011.
40. Gupte S, Wagh G: **Preeclampsia–Eclampsia.** *J Obstet Gynaecol India* 2014, **64:**4-13.
41. Red-Horse K, Zhou Y, Genbacev O, Prakobphol A, Foulk R, McMaster M, Fisher SJ: **Trophoblast differentiation during embryo implantation and formation of the maternal-fetal interface.** *J Clin Invest* 2004, **114:**744-754.
42. Dekel N, Gnainsky Y, Granot I, Mor G: **Inflammation and Implantation.** *American journal of reproductive immunology (New York, NY : 1989)* 2010, **63:**17-21.
43. Carter AM, Enders AC, Pijnenborg R: **The role of invasive trophoblast in implantation and placentation of primates.** *Philosophical Transactions of the Royal Society B: Biological Sciences* 2015, **370:**20140070.
44. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.** *Nucleic Acids Res* 2014, **42:**D1001-1006.
45. Casero D, Sandoval S, Seet CS, Scholes J, Zhu Y, Ha VL, Luong A, Parekh C, Crooks GM: **Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages.** *Nat Immunol* 2015, **16:**1282-1291.
46. St Laurent G, Wahlestedt C, Kapranov P: **The Landscape of long non-coding RNA classification.** *Trends in genetics : TIG* 2015, **31:**239-251.

47. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al: **The landscape of long noncoding RNAs in the human transcriptome.** *Nat Genet* 2015, **47:**199-208.
48. Newbern D, Freemark M: **Placental hormones and the control of maternal metabolism and fetal growth.** *Curr Opin Endocrinol Diabetes Obes* 2011, **18:**409-416.
49. Florio P, Cobellis L, Luisi S, Ciarmela P, Severi FM, Bocchi C, Petraglia F: **Changes in inhibins and activin secretion in healthy and pathological pregnancies.** *Mol Cell Endocrinol* 2001, **180:**123-130.
50. Louwen F, Muschol-Steinmetz C, Reinhard J, Reitter A, Yuan J: **A lesson for cancer research: placental microarray gene analysis in preeclampsia.** *Oncotarget* 2012, **3:**759-773.
51. Tournier I, Marlin R, Walton K, Charbonnier F, Coutant S, Théry J-C, Charbonnier C, Spurrell C, Vezain M, Ippolito L, et al: **Germline Mutations of Inhibins in Early-Onset Ovarian Epithelial Tumors.** *Hum Mutat* 2014, **35:**294-297.
52. Lyall F, Robson SC, Bulmer JN: **Spiral artery remodeling and trophoblast invasion in preeclampsia and fetal growth restriction: relationship to clinical outcome.** *Hypertension* 2013, **62:**1046-1054.
53. Noris M, Perico N, Remuzzi G: **Mechanisms of Disease: pre-eclampsia.** *Nat Clin Pract Neph* 2005, **1:**98-114.

Statement of Authorship

Title of Paper	Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Mayne BT, Leemaqz SY, Smith AK, Breen J, Roberts CT, Bianco-Miotto T: Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation. *Epigenomics* 2017, 9:279-289.

Principal Author

Name of Principal Author (Candidate)	Benjamin Mayne
Contribution to the Paper	Designed, conducted the study, analysed and interpreted the data, and wrote the manuscript.
Overall percentage (%)	85
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 14-7-17

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Shalem Y Leemaqz
Contribution to the Paper	Analyzed the data, provided intellectual input into the manuscript.
Signature	Date 14 July 2017

Name of Co-Author	Alicia K Smith
Contribution to the Paper	Involved in data creation, provided critical discussion and intellectual input into the manuscript.
Signature	Date 17 July 2017

Name of Co-Author	James Breen		
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.		
Signature		Date	20/7/2017

Name of Co-Author	Claire T Roberts		
Contribution to the Paper	Involved in the study design, provided critical discussion and intellectual input into the manuscript.		
Signature		Date	14.7.17

Name of Co-Author	Tina Blanco-Miotto		
Contribution to the Paper	Involved in the study design, provided critical discussion and intellectual input into the manuscript.		
Signature		Date	26/7/17

5 Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation

BENJAMIN T MAYNE, SHALEM Y LEEMAQZ, ALICIA K SMITH, JAMES BREEN, CLAIRE T ROBERTS and TINA BIANCO-MIOTTO

Abstract

Aim: To determine whether dynamic DNA methylation changes in the human placenta can be used to predict gestational age.

Materials & Methods: Publically available placental DNA methylation data from 12 studies, together with our own data set, using Illumina Infinium Human Methylation BeadChip arrays.

Results & Conclusion: We developed an accurate tool for predicting gestational age of placentas using 62 CpG sites. There was a higher predicted gestational age for placentas from early onset preeclampsia cases, but not term preeclampsia, compared to their chronological age. Therefore, early onset preeclampsia is associated with placental aging. Gestational age acceleration prediction from DNA methylation array data may provide insight into the molecular mechanisms of pregnancy disorders.

5.1 Introduction

DNA methylation is a heritable epigenetic process that can regulate important genetic mechanisms and processes such as gene expression, X chromosome inactivation (XCI) [1], cellular identity [2] and genomic imprinting [3]. DNA methylation is the covalent attachment of a methyl group to a cytosine ring by a DNA methyl transferase (DNMT) [4]. In this article we focus on cytosine-5 DNA methylation within CpG dinucleotides as opposed to other methylated cytosines such as CHG and CHH. Recently, DNA methylation levels for 353 CpG sites have been used to measure the epigenetic age, defined as the predicted age by DNA methylation, of a variety of human tissues, which

has a high correlation (r = 0.92) with the actual age [5]. Moreover, another study found that accelerated aging [5], defined as the difference between the estimated epigenetic age and the actual known age, is associated with maternal characteristics in pregnancy such as smoking, weight, BMI, selenium and cholesterol in peripheral blood [6]. The placenta is unique compared to other tissues, with the exception of cancer tissues [7, 8] and a human fetal fibroblast cell line (IMR90) [9] in that it has been shown that it has low levels of genome wide CpG methylation [10, 11]. Despite this, overall placental genome CpG methylation has been observed to increase during gestation [12]. However, precisely what DNA methylation changes occur during gestation and how these changes relate to pregnancy success is unknown.

Poor placental function, due to impaired placentation has been proposed to be a cause of preeclampsia (PE) [13], which is characterised by high maternal blood pressure and proteinuria [14]. Adversities during pregnancy may cause epigenetic changes and altered fetal development outcomes [15], which may be orchestrated by the placenta [16]. Differential DNA methylation in the placenta has been shown to occur in pregnancy complications [17] including PE [18-21], gestational diabetes mellitus (GDM) [18, 22, 23] and intrauterine growth restriction (IUGR) [24]. DNA methylation is critical for optimal placental and fetal development. For example, genomic imprinting regulates expression of IGF2 and H19 which are both required for proper placental development. Hypomethylation within the imprinting control region of IGF2 has been shown to be associated with reduced fetal growth [25, 26] whereas loss of imprinting of H19 has been reported to be associated with PE and with small for gestational age (SGA) infants [27].

Previous studies have reported altered placental gene expression in human pregnancy complications including PE [28-32], GDM [31] and fetal growth restriction [28, 33]. However, little is known whether changes in DNA methylation and gene expression overlap in the placenta. The relationship between global placental gene expression and DNA methylation is conflicting as one study has reported a general trend between the increase in DNA methylation and decrease in gene expression levels during gestation

[12], whereas another study using matched samples has reported no overlap between gene expression and DNA methylation changes between placentas from PE and uncomplicated pregnancies [21]. In addition, genes within partially methylated domains (PMDs) in the term placenta have been reported to be repressed [10].

In this study, we assembled a large data set of publically available placental tissue DNA methylation data that has been measured using the Illumina Infinium HumanMethylation arrays. We sought to determine the precise changes in DNA methylation that occur in the placenta across gestation and determined whether DNA methylation data can be used to predict the gestational age of a placenta. Finally, we investigated what happens to the predicted gestational age in placentas from preeclamptic pregnancies. In this study, we hypothesised that the gestational age of the placenta can be estimated by its DNA methylation levels and that pregnancy complications such as PE would be characterised by accelerated placental aging. Our computational analysis of DNA methylation data reveal accurate predictions of the gestational age of the placenta. Furthermore, our findings of gestational age acceleration in early onset preeclamptic placentas suggest a novel method to assess biological aging in the placenta. Moreover, gestational age prediction of the placenta may identify novel mechanisms in pregnancy complications, such as those associated with placental aging.

5.2 Materials and methods

5.2.1 Quantification of the DNA methylation levels

Quantification of the DNA methylation level of each CpG site that is annotated in either the Illumina 27k or 450k was performed using standard techniques. The publicly available human placental data sets (Table 5-1) were obtained from the Gene Expression Omnibus (GEO) using the GEOquery Bioconductor package. Briefly, DNA methylation levels of each CpG site from each data set was quantified by the β value, which is the ratio of the methylated and unmethylated alleles. The β value is calculated by taking the ratio of the two fluorescent signals (methylated and unmethylated signals) in studies that provided the two fluorescent signals.

Table 5-1: A description of the DNA methylation data sets containing placental tissue used in this study.

GEO Accession	Data set summary	Platform	No. of placental tissue samples	Gestational age range (weeks)
GSE31781	Placental tissue samples from three gestational age time points.	27k	18 first trimester, 10 second trimester, 14 term	8-42
GSE36642	Cord blood and placentas from monozygotic (MZ) and dizygotic (DZ) twins.	27k	28 third trimester	32-38
GSE36829	Placental tissue from term pregnancies.	27k	48 term	37-42
GSE59274	Placental tissue samples from women with PE or uncomplicated pregnancies.	27k	24 uncomplicated, 24 PE	28-41
GSE46573	Epigenome-wide association study	450k	2 term	NA
GSE52576	Genome wide human imprinting analysis of different healthy human tissue.	450k	4 term	NA
GSE54399	Whole cord blood and placental tissue from normal pregnancies.	450k	24 uncomplicated	NA
GSE57767	Placental tissue samples from normal term, preterm PE and term PE women.	450k	14 uncomplicated, 12 preterm PE, 19 term PE	NA
GSE44667	Placental tissue from women with EOPET and gestational age matched controls.	450k	20 third trimester, 20 PE	25-37
GSE66210	First trimester chorionic villous samples from normal and trisomy pregnancies.	450k	12 normal, 12 trisomy 21, 12 trisomy 18, 6 trisomy 13	NA
GSE73375	Placental tissue samples from preeclamptic and normotensive women.	450k	17 uncomplicated, 19 PE	NA
GSE74738	Placental tissue from uncomplicated pregnancies.	450k	28 term	36-42
This study	Placental tissue from uncomplicated elective caesarean pregnancies.	27k	22 term	35-40

DZ: dizygotic, EOPET: early onset preeclampsia, MZ: monozygotc, NA: Not available, PE: preeclampsia, 27k: Illumina Infinium HumanMethylation27, 450k: Illumina Infinium HumanMethylation450

5.2.2 Differential methylation analysis

Differential methylation analysis was performed using the 450k data sets for which we had data for 469,017 probes. Normalisation was performed using the BMIQ method [34], which corrects for the probe design bias in the Illumina Infinium HumanMethylation450 BeadChip followed by quantile normalisation. Since the data were from multiple data sets, batch effects were corrected using the Combat function in the ChAMP Bioconductor package which corrects for biases as a result of batch effects [35, 36]. Multidimensional scaling plots of the 1000 most variable probes of the data were used to check for outliers. Sample sex was identified using the minfi package in which the median value of the β values for probes that mapped uniquely for the X and Y chromosome, respectively, were determined [37]. Differentially methylated CpG sites were identified using empirical Bayesian variance method in limma which was used to test for differential methylation differences [38]. The Bumphunter Bioconductor package was used to identify differentially methylated regions by running 1000 permutations of the data [39]. We selected data sets (GSE44667, GSE46573, GSE52576, GSE54399, GSE57767 and GSE73375) that had used the Illumina Infinium HumanMethylation450 BeadChip array to assess sex differences in placental DNA methylation and differential methylation between PE and uncomplicated pregnancies (Table 5-1). In total, we had placental DNA methylation data for 70 preeclamptic and 62 uncomplicated pregnancies.

5.2.3 Gestational age prediction

We selected data sets for placental tissue samples that were publically available and which contained individual gestational age information and were from healthy singleton pregnancies (GSE31781, GSE36829, GSE59274, GSE44667 and GSE74738) (Table 5-1). We also combined these data sets with our own data set for 22 term placentas from uncomplicated pregnancies. Probes that were present in both the 27k and 450k array were kept and probes that were annotated to the sex chromosomes were removed. Every sample in the training data set contained β values for 18,437 probes. Normalisation was performed as described by Horvath [5], using a modified version of the Beta MIxture

Quantile dilation (BMIQ) [34] method. This modified version of BMIQ rescales the Infinium II probes to the mean β value of each probe in the largest data set (GSE36829) prior to quantile normalisation. Using the R package glmnet [40], we regressed gestational age with the 18,437 probes. The elastic net regression or the alpha parameter of glmnet was 0.5 and the minimum lambda value based on the training data was 0.6807.The elastic net model automatically selected 62 CpG sites, such that given the level of methylation of the 62 CpG sites the gestational age of a placenta can be calculated.

5.2.4 Gestational age acceleration heritability

To determine the heritability of gestational age acceleration, which is defined as the difference between the chronological and predicted gestational age of a placenta we used Falconer's formula (H2 = 2(cor(MZ)-cor(DZ))). Falconer's formula was used to determine the broad sense heritability which is the proportion of variance of gestational age acceleration as a result of genetic factors. The broad sense heritability was determined by using a data set (GSE36642) that contained monozygotic and dizygotic twins of the same sex. Firstly, the predicted gestational age of the twin placentas were determined as described above. The twin data set was split into either monozygotic or dizygotic twins and for each twin pair the gestational age acceleration was calculated. For each twin pair, each sibling was randomly selected as either twin 1 or twin 2 and the pearson correlation of gestational age acceleration was determined in both monozygotic and dizygotic twins. The correlation of both monozygotic and dizygotic twins were inputted into Falconer's formula to determine the broad sense heritability.

5.2.5 Annotation of CpG sites

Annotation of all the CpG sites within the analysis of this study for the Illumina Infinium HumanMethylation27 and HumanMethylation450 BeadChip was performed using two annotation Bioconductor packages [41, 42].

5.3 Results

5.3.1 Differential methylation in placentas from preeclamptic pregnancies

We first searched the GEO [43] for DNA methylation data sets containing placental tissue that were measured on either the Illumina Infinium HumanMethylation27 BeadChip or Infinium HumanMethylation450 BeadChip. In total, we identified 387 placental tissue samples from 12 different data sets (Table 5-1). We selected only placentas assessed using the Illumina DNA methylation arrays, the most commonly used platform to quantify DNA methylation in the placenta. Here, we used publically available placental DNA methylation data (Table 5-1) to assess differential methylation between placentas from PE and uncomplicated pregnancies. We selected only datasets that used the Infinium HumanMethylation450 BeadChip for differential methylation analysis since most studies involving placental samples from women with preeclampsia had used this platform (Table 5-1) and it contains the largest number of CpG sites available for analysis. When comparing 70 placentas from preeclamptic pregnancies and 62 placentas from term uncomplicated pregnancies we found a total of 741 CpG sites (FDR < 0.01) were differentially methylated (Supplementary Table 5-1). We then tested for differentially methylated regions (DMRs) and identified 3 DMRs in placentas from preeclamptic pregnancies which overlapped the 5' region of MARC2, FAM3B and TP53TG1.

5.3.2 Sex differences in DNA methylation

To identify whether sex differences also occur in the placental DNA methylome, we identified a total of 2,898 differentially methylated CpG sites (FDR < 0.01) between 35 male and 27 female placentas from uncomplicated singleton term pregnancies that had been analysed using the Infinium HumanMethylation450 BeadChip (Supplementary Table 5-2). Of the 2,898 CpG sites, 420 were located on autosomes, 2,464 on the X chromosome and 14 on the Y chromosome. Although we are reporting Y chromosome CpG sites, we do not consider these CpG sites as differentially methylated between sexes. In addition, upon removing the Y chromosome from our analysis we did not observe a

difference in the total number of differentially methylated X chromosome or autosomal CpG sites. We also identified a total of 396 DMRs between sexes (Supplementary Table 5-3). All of these were located on the X chromosome with 311 and 85 being hypermethylated in females and males, respectively. The 85 hypermethylated DMRs on the X chromosome in males did not overlap with any reported sex biased genes [44], with the exception of XIST which is upregulated in females and is well-known for its role in XCI in female mammalian somatic cells [45].The most statistically significant differentially methylated autosomal CpG site as defined by the empirical Bayesian variance method in limma between sexes was hypermethylated in females and was within exon 1 of TLE1 which is a marker of synovial sarcoma [46]. Although TLE1 is expressed in the placenta, it has not been reported to be differentially expressed between fetal sexes in the placenta [44, 47].

5.3.3 Gestational age calculator training data set

Although a multi-tissue age predictor using DNA methylation data has been previously developed [5], we set out to determine if DNA methylation can be used to predict the gestational age of the placenta. To develop our placental gestational age calculator, only placentas from healthy singleton pregnancies with individual gestational age information were included. Placental tissue samples from PE pregnancies were excluded from the training data set to reduce potential confounding factors caused by the disease. In total, we used 170 placental tissue samples that had individual gestational age information to generate the gestational age calculator. The 170 placental tissue samples were taken from five publically available data sets, along with our own generated data analysing 22 term placentas (Table 5-1). Four of the six data sets were obtained using the Infinium HumanMethylation27 BeadChip and the other two on the Infinium HumanMethylation450 BeadChip and the included datasets contained placental tissue samples that spanned 8-42 weeks gestation. We selected probes that were present in all 6 data sets and removed probes that were found on sex chromosomes, leaving a total of 18,437 probes (no missing data). We randomly assigned half of the 170 placental tissue samples to a training data set, leaving the other 85 samples for validation.

5.3.4 Identifying and validating the gestational age calculator

Briefly, we firstly normalised the training data using the modified version of the Beta MIxture Quantile dilation (BMIQ) [5, 34]. The mean beta value of each probe in the largest data set (GSE3829) was used as the gold standard of the probes, similarly as previously described [5], in the normalisation step (Supplementary Table 5-4). A gold standard of the probes was used for normalisation since the data sets were from two different microarray platforms and it rescaled the probes that were present in both microarray platforms. After normalisation we regressed the chronological gestational age against the 18,437 CpG sites using an elastic net penalised regression model [40]. The model automatically selected 62 CpG sites (Supplementary Table 5-5) to predict the gestational age of a placenta. In the training data set we found an extremely high correlation ($r = 0.99$, $p < 2.2e-16$) between the chronological and the predicted gestation age (Figure 5-1A). In addition, the median absolute difference between the predicted and chronological gestational age in the training data set was found to be 0.23 weeks. We then tested these 62 CpG sites in the validation data set (Figure 5-1B) and also found a high correlation between the chronological and predicted gestational age ($r = 0.95$, $p < 2.2e-16$). The median absolute difference in the validation data set was 1.47 weeks and the root mean square error was 2.3. The heatmap (Figure 5-1C) allows visualisation of the CpG sites and shows changes in DNA methylation across gestation. Furthermore, the lack of vertical lines in the heatmap suggests that the CpG sites are robust against data set effects. In order to further validate the 62 CpG sites, we predicted the gestational age of all remaining publically available placental tissue samples from uncomplicated pregnancies that did not have individual gestational age information (Supplementary Table 5-6). Although these remaining samples did not have individual gestational age information, we sought to determine if the predicted gestational age matched the labelled trimester of pregnancy for each sample. We found a concordance between the predicted gestational age and labelled trimester of pregnancy, which assured us that it is an accurate predictor of gestational age. Here in this study we refer to the predicted gestational age of each placenta as the DNA methylation gestational age (DNAm GA).

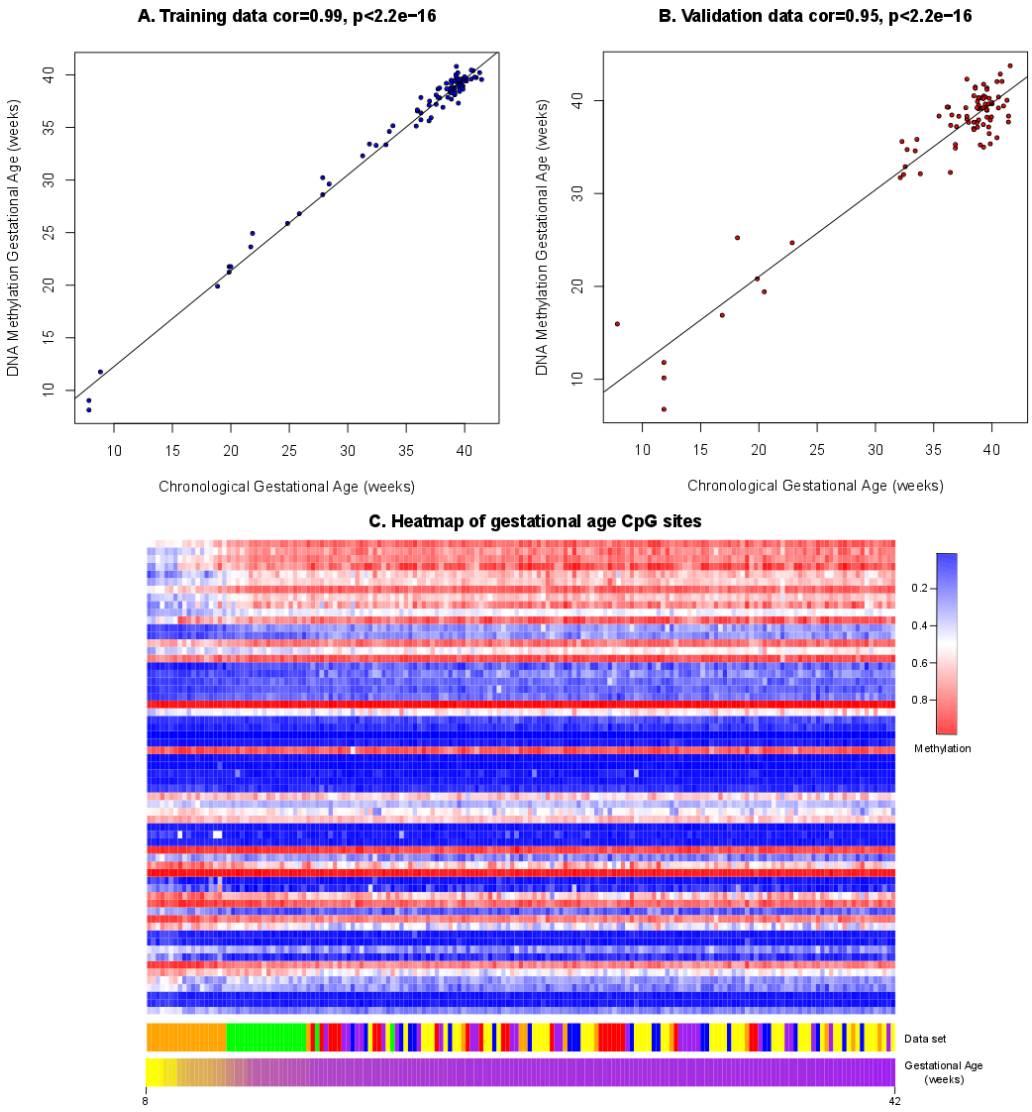


Figure 5-1: The correlation between the chronological and DNA methylation gestational age of each placenta.

(**A**) Training data set and the (**B**) validation data set. (**C**) A heatmap visualising the gradual changes in DNA methylation in each of the 62 CpG sites (rows) across gestation in all samples (columns). The samples have been ordered by increasing gestational age and the probes have been ordered by the increasing magnitude of correlation with gestational age. The Data Set heatmap represents the origin of each sample.

5.3.5 The 62 gestational clock CpG sites

The 62 gestational clock CpG sites can be characterised into two groups depending on the direction of their correlation with gestational age. 27 CpG sites were found to positively correlate and become hypermethylated with increasing gestational age whereas, the other 35 CpG sites negatively correlated and became hypomethylated with increasing gestational age.

5.3.6 Gestational age acceleration in placentas from preeclamptic pregnancies

Since differential DNA methylation occurs in placentas from PE compared to uncomplicated pregnancies [17], we investigated if accurate prediction of gestation age can also be achieved in placentas from PE pregnancies. Two data sets (GSE44667 and GSE59274) contained individual gestational age information for placental tissue samples from PE pregnancies (26 early onset PE and 18 late onset PE). We then compared the chronological gestational age with the DNAm GA (Figure 5-2) and found that placentas from early onset PE pregnancies (< 34 weeks gestation) had a higher DNAm GA compared to their chronological gestational age ($p = 3.44e-6$, two-tailed, paired t-test). However, late onset placentas from PE pregnancies (≥ 34 weeks gestation) did not show any significant difference between their chronological and DNAm GA ($p = 0.38$) indicating that late onset PE does not affect placental aging. From here on, we refer to the difference between chronological and DNAm GA as gestational age acceleration, similar to what has been defined previously [5]. One data set (GSE36642) contained placentas from monozygotic (MZ) and dizygotic (DZ) twins with individual gestational age information, allowing the determination of gestational age acceleration heritability by calculating the broad sense heritability using Falconer's formula ($H2 = 2(cor(MZ)-cor(DZ))$). We conducted our analysis on gestational age acceleration heritability on twin samples of the same sex. The broad sense heritability was used to determine the proportion of variance of gestational age acceleration as a result of genetic variation. Despite having a small sample size, we calculated the gestational age acceleration for

each sample and determined the broad sense heritability to be 57.2% in MZ and DZ twin pairs (Supplementary Figure 5-1).

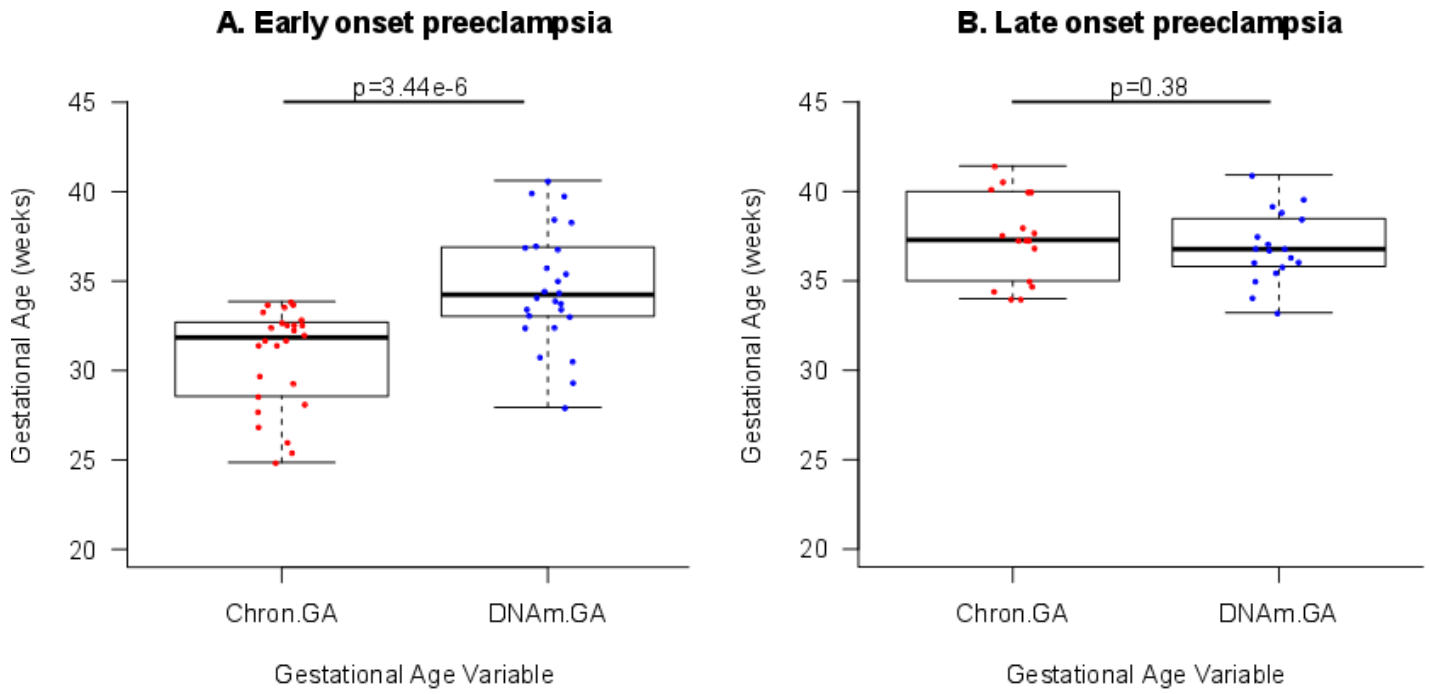


Figure 5-2: Gestational age acceleration in early onset PE.

Placentas from pregnancies complicated by early onset PE (**A**) show a higher DNA methylation gestational age (DNAm.GA) compared to their chronological gestational age (Chron.GA). (**B**) Placentas from late onset PE do not show any difference between the chronological and DNA methylation gestational age. Error bars are represented as standard deviations.

5.4 Discussion

In this study, we investigated DNA methylation differences in the placenta across gestation and in different pregnancy outcomes. We found 34 genes that have been reported to be differentially expressed in placenta from PE compared to uncomplicated pregnancies [30] to contain at least one differentially methylated CpG site (Supplementary Table 5-1). For example, there was a differentially methylated CpG site in the promoter of RAC1 (ras-related C3 botulinum toxin substrate 1) (Supplementary Table 5-1), a member of the RAS-superfamily [48], which has been found to be upregulated in placenta from preeclamptic pregnancies [30]. We also tested for DMRs in placentas from preeclamptic and uncomplicated pregnancies. We identified three DMRs which overlapped the 5' region of MARC2, FAM3B and TP53TG1. However, to our knowledge these genes have not been reported to be differentially expressed in placenta from women with preeclampsia.

Sex differences in pregnancy outcomes have been reported, for example women bearing a male fetus are at a 20% higher risk of preterm birth (PTB) [49-51]. The fetus and placenta are genetically identical [52]. The placenta has been implicated in a number of pregnancy complications, including PTB [53]. Therefore it is reasonable to suggest that sex differences in outcomes may potentially be orchestrated by the placenta. Sex differences in placental gene expression have been previously reported [44, 47] and in comparison to the placental sex biased gene expression meta-analysis [44], 20 genes (located on the X chromosome and upregulated in females) were found to contain at least one CpG site or DMR that was hypermethylated in females. However, this may be the result of XCI in females.

It is unclear why 85 X chromosome DMRs should be hypermethylated in males considering XCI occurs in females. Despite being hypermethylated in males, there are no reports of sex differences in expression of the genes in which these DMRs occur.

Potentially they may regulate other sex-specific differentially expressed RNAs including ncRNAs or affect TF binding. Further research is required to elucidate the precise mechanisms by which the 85 X chromosome hypermethylated DMRs in males act.

We therefore, found a lack of overlap with our differential DNA methylation analysis and two gene expression meta-analyses [30, 44]. These findings were consistent with a previous study in which changes in gene expression and DNA methylation did not overlap from matched placental samples from PE and uncomplicated pregnancies [21]. Furthermore, changes in DNA methylation within the promoters of genes does not always alter gene expression as reported for colon cancer [54]. We also found little overlap between genes that have previously been reported to be sex biased in the placenta [44] and our DNA methylation results. Interestingly, we did find 420 differentially methylated autosomal CpG sites and 85 X chromosome hypermethylated DMRs in males. However, due to the poor overlap with gene expression and the effect of XCI it is difficult to draw any conclusions on what the effect of the differences in DNA methylation that were observed had on gene expression.

The lack of overlap between DNA methylation and reported gene expression changes have several limitations. Firstly, in comparison to the two gene expression meta-analyses, we analysed far fewer samples and therefore may not have been statistically powered to detect small differences. Secondly, the Illumina DNA Methylation BeadChip arrays only assess approximately 2% of the CpG sites in the human genome and do not assess other methylated cytosine sites such as CHG and CHH. Therefore, we may not have captured the true landscape of DNA methylation in the placenta. Finally, the comparison between DNA methylation and gene expression was not in the same samples and therefore there may have been too much biological variability to detect any overlap. Despite these limitations, our findings suggest CpG methylation in the placenta may not be a key regulator of gene expression and therefore may be more dependent on other epigenetic factors such as histone modifications, small RNA regulation or non-coding RNA changes. It has been reported that the placenta with the exception of the brain, has high

levels of non-CpG methylation compared to other human tissues [55] which may have a bigger influence on gene expression levels.

In this study, we identified 62 CpG sites, which together can be used to determine the gestational age of a placenta. Several limitations of our study for predicting gestational age do require discussion. Firstly, the training data set consisted of placentas from 8-42 weeks gestation with a bias of samples being from late third trimester. First and second trimester samples comprised only 11% and 9.5% of the training and validation data, respectively. This may have created some biases in the CpG sites chosen and may cause some inaccuracy in identifying the gestational age of placentas from first and second trimester. Secondly, in relation to gestational age acceleration heritability the twin data set only contained 14 twin pairs. Therefore we may have not captured the true extent of gestational age heritability within this study. Future studies with large sample sizes of twin pairs are required to determine the true extent of the heritability of gestational age acceleration. In addition to gestational age prediction, we used first and second trimester placentas from terminated pregnancies. One limitation is that some of these placentas may have been from women destined to develop a pregnancy complication which may have implications for our gestational age prediction. A possible approach to overcome this limitation is to use placental villi from chorionic villus sampling in ongoing pregnancies. Thereby, samples that were from complicated pregnancies could be excluded from the analysis. Unfortunately, to our knowledge there is no publically available DNA methylation data on such samples that we could use to test our gestational age prediction.

Placentas from early onset preeclamptic pregnancies were found to have a higher DNAm GA compared to their chronological gestational age. Using a twin data set we were able to determine the broad sense heritability of gestational age acceleration to be 57.2%. This finding suggests environmental factors also have an influence together with genetic factors on gestational age acceleration. Maternal lifestyle factors such as smoking [56, 57] are known to alter DNA methylation levels in the placenta. Therefore, the maternal environment can affect the intrauterine environment, and thereby could influence

gestational age acceleration in the placenta. Furthermore, other lifestyle factors such as BMI have been found to increase the epigenetic age in certain tissues such as the liver [58]. It would therefore be important to investigate the effect of maternal lifestyle factors on gestational age acceleration as they may have implications for pregnancy success. Unfortunately these data are often not recorded for publically available data sets. This limitation also applies to the exact nature of the twin placentas included in the publically available data. For example, we do not know if the twin pregnancies had fused placentas or not. Although each individual twin is listed as having a separate placenta, it may be possible that some of the twin pregnancies had fused placentas and therefore the exact sampling sites could confound the data on heritability. Therefore, we suggest some caution with respect to the heritability of gestational age acceleration analysis as we do not have full clinical details for the publicly available twin data set (GSE36642).

Using our gestational age tool, we also provide an estimation of the gestational age of all remaining publicly available placenta samples from uncomplicated pregnancies for which gestational age information is not recorded as a resource to the scientific community (Supplementary Table 5-6). It has also been reported that the placental transcriptome is clearly distinct in PE compared to other pregnancy complications [31]. It would therefore be of interest to determine if the placental gestational age acceleration observed in early onset PE pregnancies also occurs in other pregnancy complications. Likewise, the gestational age acceleration observed in placentas from early but not late onset PE highlights potential differences in the etiology of the two diseases. Furthermore, gestational age acceleration may potentially reveal potential mechanisms in the development of early-onset PE and other pregnancy complications. First trimester chorionic villus samples could potentially be used to determine when accelerated placental aging first occurs. This may provide insight into the mechanisms and the association of placental aging and pregnancy complications such as early onset PE. However, chorionic villus sampling is only used in some high risk pregnancies but has a miscarriage risk so tissue availability is limited.

5.5 Conclusion

In conclusion, we have identified 62 CpG sites that can be used to determine the DNAm GA of a placenta. Furthermore, we found evidence of placental aging in placentas from early onset preeclampsia. Future studies are required to determine if gestational age acceleration is unique to early onset preeclampsia or is common to other pregnancy complications. In addition, future research should also determine if gestational age acceleration or placental aging could be detected perhaps in maternal blood early in pregnancy in women who are destined to develop a pregnancy complication. Although, we found little overlap between DNA methylation and gene expression changes, further studies involving matched samples are required to confirm these findings.

5.6 Supporting Information

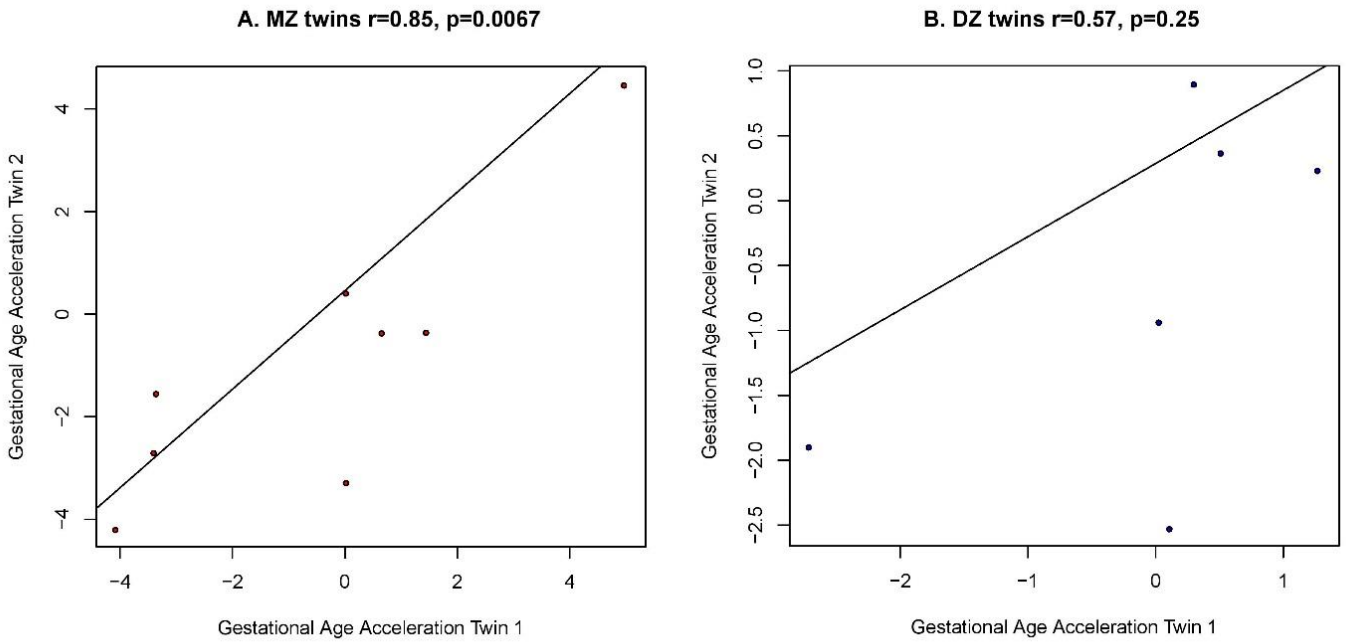
For Supplementary Table 5-1, 5-2, 5-3, 5-4, 5-5, and 5-6 please refer to the electronic supporting information.

Supplementary Table 5-1: CpG sites that were detected to be differentially methylated between placentas from preeclampsia and uncomplicated pregnancies.
The position and the gene that the CpG sites is associated with is supplied. Genes that have also been detected to be differentially expressed in placentas from preeclampsia and uncomplicated pregnancies [30] are highlighted in bold.
Supplementary Table 5-2: CpG sites that were found to be detected to be differentially methylated between male and females placentas from uncomplicated pregnancies.
Supplementary Table 5-3: Differentially methylated regions between male and females pregnancies from uncomplicated pregnancies.
Regions with a p.value and/or fwer = 0 are values with p.value and/or fwer < 0.001. Bumphunter was unable to determine the exact number by running 1000 permutations on the data.
Supplementary Table 5-4: The gold mean of each probe used in the normalisation step in determining the gestational age of a placenta.

The mean of methylation of each probe in the largest data set (GSE36829) was used along with the modified BMIQ normalisation step [5] to rescale and adjust probes with the two different chemistries as previously described [5].

Supplementary Table 5-5: The coefficient values and locations of the 62 CpG sites used to determine the gestational age of a placenta.

Supplementary Table 5-6: The predicted gestational age of all remaining placenta samples that are publicly available that do not have individual gestational age information.



Supplementary Figure 5-1: Twin data set (GSE36642) used to calculate the broad sense of heritability of accelerated gestational aging.

Each point represents a twin pair from (A) monozygotic (MZ) and (B) dizygotic (DZ) placental tissue samples. There was a high correlation of gestational age acceleration between twin pairs for MZ but not for DZ twins.

References

1. Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, Antonarakis SE: **DNA methylation profiles of human active and inactive X chromosomes.** *Genome Res* 2011, **21:**1592-1600.
2. Novak P, Stampfer MR, Munoz-Rodriguez JL, Garbe JC, Ehrich M, Futscher BW, Jensen TJ: **Cell-type specific DNA methylation patterns define human breast cellular identity.** *PLoS One* 2012, **7:**e52299.
3. Paulsen M, Ferguson-Smith AC: **DNA methylation in genomic imprinting, development, and disease.** *J Pathol* 2001, **195:**97-110.
4. Robertson KD: **DNA methylation and human disease.** *Nat Rev Genet* 2005, **6:**597-610.
5. Horvath S: **DNA methylation age of human tissues and cell types.** *Genome Biol* 2013, **14:**R115.
6. Simpkin AJ, Hemani G, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, Ring SM, Sharp GC, Tilling K, Horvath S, et al: **Prenatal and early life influences on epigenetic age in children: a study of mother-offspring pairs from two cohort studies.** *Hum Mol Genet* 2016, **25:**191-201.
7. Hon GC, Hawkins RD, Caballero OL, Lo C, Lister R, Pelizzola M, Valsesia A, Ye Z, Kuan S, Edsall LE, et al: **Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer.** *Genome Res* 2012, **22:**246-258.
8. Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CP, van Dijk CM, Tollenaar RA, et al: **Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains.** *Nat Genet* 2012, **44:**40-46.
9. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, et al: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462:**315-322.
10. Schroeder DI, Blair JD, Lott P, Yu HO, Hong D, Crary F, Ashwood P, Walker C, Korf I, Robinson WP, LaSalle JM: **The human placenta methylome.** *Proc Natl Acad Sci U S A* 2013, **110:**6037-6042.
11. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, et al: **Human body epigenome maps reveal noncanonical DNA methylation variation.** *Nature* 2015, **523:**212-216.
12. Novakovic B, Yuen RK, Gordon L, Penaherrera MS, Sharkey A, Moffett A, Craig JM, Robinson WP, Saffery R: **Evidence for widespread changes in promoter methylation profile in human placenta in response to increasing gestational age and environmental/stochastic factors.** *BMC Genomics* 2011, **12:**529.
13. Redman CW, Sargent IL, Staff AC: **IFPA Senior Award Lecture: making sense of pre-eclampsia - two placental causes of preeclampsia?** *Placenta* 2014, **35 Suppl:**S20-25.
14. Williams D, Craft N: **Pre-eclampsia.** *BMJ* 2012, **345:**e4437.

15. Monk C, Spicer J, Champagne FA: **Linking Prenatal Maternal Adversity to Developmental Outcomes in Infants: The Role of Epigenetic Pathways.** *Dev Psychopathol* 2012, **24:**1361-1376.
16. Novakovic B, Saffery R: **The ever growing complexity of placental epigenetics - role in adverse pregnancy outcomes and fetal programming.** *Placenta* 2012, **33:**959-970.
17. Bianco-Miotto T, Mayne B, Buckberry S, Breen J, Rodriguez Lopez C, Roberts C: **Recent progress towards understanding the role of DNA methylation in human placental development.** *Reproduction* 2016.
18. Liu L, Zhang X, Rong C, Rui C, Ji H, Qian YJ, Jia R, Sun L: **Distinct DNA methylomes of human placentas between pre-eclampsia and gestational diabetes mellitus.** *Cell Physiol Biochem* 2014, **34:**1877-1889.
19. Anton L, Brown AG, Bartolomei MS, Elovitz MA: **Differential methylation of genes associated with cell adhesion in preeclamptic placentas.** *PLoS One* 2014, **9:**e100148.
20. Blair JD, Yuen RK, Lim BK, McFadden DE, von Dadelszen P, Robinson WP: **Widespread DNA hypomethylation at gene enhancer regions in placentas associated with early-onset pre-eclampsia.** *Mol Hum Reprod* 2013, **19:**697-708.
21. Chu T, Bunce K, Shaw P, Shridhar V, Althouse A, Hubel C, Peters D: **Comprehensive Analysis of Preeclampsia-Associated DNA Methylation in the Placenta.** *PLoS One* 2014, **9**.
22. Finer S, Mathews C, Lowe R, Smart M, Hillman S, Foo L, Sinha A, Williams D, Rakyan VK, Hitman GA: **Maternal gestational diabetes is associated with genome-wide DNA methylation variation in placenta and cord blood of exposed offspring.** *Hum Mol Genet* 2015.
23. Ruchat SM, Houde AA, Voisin G, St-Pierre J, Perron P, Baillargeon JP, Gaudet D, Hivert MF, Brisson D, Bouchard L: **Gestational diabetes mellitus epigenetically affects genes predominantly involved in metabolic diseases.** *Epigenetics* 2013, **8:**935-943.
24. Hillman SL, Finer S, Smart MC, Mathews C, Lowe R, Rakyan VK, Hitman GA, Williams DJ: **Novel DNA methylation profiles associated with key gene regulation and transcription pathways in blood and placenta of growth-restricted neonates.** *Epigenetics* 2015, **10:**50-61.
25. Jacob KJ, Robinson WP, Lefebvre L: **Beckwith-Wiedemann and Silver-Russell syndromes: opposite developmental imbalances in imprinted regulators of placental function and embryonic growth.** *Clin Genet* 2013, **84:**326-334.
26. Fowden AL, Sibley C, Reik W, Constancia M: **Imprinted genes, placental development and fetal growth.** *Horm Res* 2006, **65 Suppl 3:**50-58.
27. Yu L, Chen M, Zhao D, Yi P, Lu L, Han J, Zheng X, Zhou Y, Li L: **The H19 gene imprinting in normal pregnancy and pre-eclampsia.** *Placenta* 2009, **30:**443-447.
28. Guo L, Tsai SQ, Hardison NE, James AH, Motsinger-Reif AA, Thames B, Stone EA, Deng C, Piedrahita JA: **Differentially expressed microRNAs and affected biological pathways revealed by modulated modularity clustering (MMC) analysis of human preeclamptic and IUGR placentas.** *Placenta* 2013, **34:**599-605.

29. Nishizawa H, Pryor-Koishi K, Kato T, Kowa H, Kurahashi H, Udagawa Y: **Microarray analysis of differentially expressed fetal genes in placental tissue derived from early and late onset severe pre-eclampsia.** *Placenta* 2007, **28:**487-497.
30. van Uitert M, Moerland PD, Enquobahrie DA, Laivuori H, van der Post JA, Ris-Stalpers C, Afink GB: **Meta-Analysis of Placental Transcriptome Data Identifies a Novel Molecular Pathway Related to Preeclampsia.** *PLoS One* 2015, **10:**e0132468.
31. Sober S, Reiman M, Kikas T, Rull K, Inno R, Vaas P, Teesalu P, Marti JM, Mattila P, Laan M: **Extensive shift in placental transcriptome profile in preeclampsia and placental origin of adverse pregnancy outcomes.** *Sci Rep* 2015, **5:**13336.
32. Kaartokallio T, Cervera A, Kyllonen A, Laivuori K: **Gene expression profiling of pre-eclamptic placentae by RNA sequencing.** *Sci Rep* 2015, **5:**14107.
33. Nishizawa H, Ota S, Suzuki M, Kato T, Sekiya T, Kurahashi H, Udagawa Y: **Comparative gene expression profiling of placentas from patients with severe pre-eclampsia and unexplained fetal growth restriction.** *Reprod Biol Endocrinol* 2011, **9:**107.
34. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S: **A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data.** *Bioinformatics* 2013, **29:**189-196.
35. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments.** *Bioinformatics* 2012, **28:**882-883.
36. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S: **ChAMP: 450k Chip Analysis Methylation Pipeline.** *Bioinformatics* 2014, **30:**428-430.
37. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA: **Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.** *Bioinformatics* 2014, **30:**1363-1369.
38. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res* 2015.
39. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA: **Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.** *Int J Epidemiol* 2012, **41:**200-209.
40. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *J Stat Softw* 2010, **33:**1-22.
41. S. D: **IlluminaHumanMethylation27k.db: Illumina Illumina Human Methylation 27k annotation data (chip IlluminaHumanMethylation27k). R package version 1.4.8.**
42. Jr.. TTa: **IlluminaHumanMethylation450k.db: Illumina Human Methylation 450k annotation data. R package version 2.0.9.**
43. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al: **NCBI GEO: archive**

for functional genomics data sets--update. *Nucleic Acids Res* 2013, **41:**D991-995.
44. Buckberry S, Bianco-Miotto T, Bent SJ, Dekker GA, Roberts CT: **Integrative transcriptome meta-analysis reveals widespread sex-biased gene expression at the human fetal-maternal interface.** *Mol Hum Reprod* 2014, **20:**810-819.
45. Augui S, Nora EP, Heard E: **Regulation of X-chromosome inactivation by the X-inactivation centre.** *Nat Rev Genet* 2011, **12:**429-442.
46. Terry J, Saito T, Subramanian S, Ruttan C, Antonescu CR, Goldblum JR, Downs-Kelly E, Corless CL, Rubin BP, van de Rijn M, et al: **TLE1 as a diagnostic immunohistochemical marker for synovial sarcoma emerging from gene expression profiling studies.** *Am J Surg Pathol* 2007, **31:**240-246.
47. Sood R, Zehnder JL, Druzin ML, Brown PO: **Gene expression patterns in human placenta.** *Proc Natl Acad Sci U S A* 2006, **103:**5478-5483.
48. Bai Y, Xiang X, Liang C, Shi L: **Regulating Rac in the nervous system: molecular function and disease implication of Rac GEFs and GAPs.** *Biomed Res Int* 2015, **2015:**632450.
49. Clifton VL: **Review: Sex and the human placenta: mediating differential strategies of fetal growth and survival.** *Placenta* 2010, **31 Suppl:**S33-39.
50. Vatten LJ, Skjaerven R: **Offspring sex and pregnancy outcome by length of gestation.** *Early Hum Dev* 2004, **76:**47-54.
51. Verburg PE, Tucker G, Scheil W, Erwich JJ, Dekker GA, Roberts CT: **Sexual Dimorphism in Adverse Pregnancy Outcomes - A Retrospective Australian Population Study 1981-2011.** *PLoS One* 2016, **11:**e0158807.
52. Gude NM, Roberts CT, Kalionis B, King RG: **Growth and function of the normal human placenta.** *Thromb Res* 2004, **114:**397-407.
53. Roberts CT: **IFPA Award in Placentology Lecture: Complicated interactions between genes and the environment in placentation, pregnancy outcome and long term health.** *Placenta* 2010, **31 Suppl:**S47-53.
54. Moarii M, Boeva V, Vert J-P, Reyal F: **Changes in correlation between promoter methylation and gene expression in cancer.** *BMC Genomics* 2015, **16:**873.
55. Guo JU, Su Y, Shin JH, Shin J, Li H, Xie B, Zhong C, Hu S, Le T, Fan G, et al: **Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain.** *Nat Neurosci* 2014, **17:**215-222.
56. Suter M, Ma J, Harris A, Patterson L, Brown KA, Shope C, Showalter L, Abramovici A, Aagaard-Tillery KM: **Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression.** *Epigenetics* 2011, **6:**1284-1294.
57. Maccani JZ, Koestler DC, Houseman EA, Marsit CJ, Kelsey KT: **Placental DNA methylation alterations associated with maternal tobacco smoking at the RUNX3 gene are also associated with gestational age.** *Epigenomics* 2013, **5:**619-630.
58. Horvath S, Erhart W, Brosch M, Ammerpohl O, von Schonfels W, Ahrens M, Heits N, Bell JT, Tsai PC, Spector TD, et al: **Obesity accelerates epigenetic aging of human liver.** *Proc Natl Acad Sci U S A* 2014, **111:**15538-15543.

Statement of Authorship

Title of Paper	msgbsR: an R package for analysing methylation-sensitive genotyping-by-sequencing data
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Benjamin Mayne
Contribution to the Paper	BTM conceived, designed and made the R package, analysed and interpreted the data and wrote the manuscript.
Overall percentage (%)	85
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 14-7-17

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Shalem Leemaqz
Contribution to the Paper	Involved in study design provided critical discussion and intellectual input into the manuscript.
Signature	Date 14 July 2017

Name of Co-Author	Sam Buckberry
Contribution to the Paper	Involved in study design provided critical discussion and intellectual input into the manuscript.
Signature	Date 16 July 2017

Name of Co-Author	Carlos Rodriguez Lopez
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.
Signature	Date 25/7/2017

Name of Co-Author	Claire Roberts
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.
Signature	Date 14.7.17

Name of Co-Author	Tina Blanco-Miotto
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.
Signature	Date 26/7/17

Name of Co-Author	James Breen
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.
Signature	Date 20/7/2017

6 msgbsR: an R package for analysing methylation-sensitive Genotyping-by-sequencing data

BENJAMIN T MAYNE, SHALEM LEEMAQZ, SAM BUCKBERRY, CARLOS RODRIGUEZ LOPEZ , CLAIRE T ROBERTS, TINA BIANCO-MIOTTO, JAMES BREEN

Abstract

Genotyping-by-sequencing (GBS) is a practical and cost effective method for analysing large genomes from high diversity species. This method of sequencing, coupled with methylation-sensitive enzymes, is an effective tool to study DNA methylation in parts of the genome that are inaccessible in other sequencing techniques or are not annotated in microarrays technologies. Current software tools do not fulfil all experimental GBS assays such as those using methylation-sensitive restriction enzymes for determining differences in DNA methylation between samples. Here we present *msgbsR*, an R package that contains tools for the analysis of methylation-sensitive genotyping-by-sequencing (msGBS) experiments. *msgbsR* can be used to identify and quantify read counts at methylated sites directly from alignment files (BAM files). The package also enables verification of restriction enzyme cut sites with the correct recognition sequence of the individual enzyme. In addition, *msgbsR* allows the analysis of differential DNA methylation and facilitates the creation of genomic plots of cut site locations. *msgbsR* assesses DNA methylation based on read coverage (like RNAseq experiments) rather than methylation proportion, and is a useful tool in analysing differential methylation on large populations. The package is fully documented and available freely online as a Bioconductor package (https://bioconductor.org/packages/release/bioc/html/msgbsR.html).

6.1 Introduction

Genotyping-by-sequencing (GBS) is a cost effective next generation sequencing method to analyse high diversity in large genome species. Reducing genome complexity with restriction enzymes (REs) can be advantageous as it may reach parts of the genome

inaccessible to sequence capture approaches [1]. However, current GBS data analysis tools do not satisfy all experimental designs. For example, using methylation-sensitive REs in a GBS experiment, which is known as methylation-sensitive GBS (msGBS) [2] or methylation-sensitive restriction enzyme digestion followed by sequencing (MRE-seq) [3] is an effective way to identify differentially methylated sites that may not be annotated or accessible in other technologies, such as microarrays. Other GBS data analysis tools such as Stacks [4] and TASSEL [5] focus on association mapping and do not supply methods on identifying, annotating or testing for differential methylated sites from a msGBS experiment.

The cost of NGS has declined dramatically in recent years due to the increased throughput of current sequencing machines such as Illumina HiSeq X Ten and NovaSeq platforms. As a consequence, it is now feasible to determine DNA methylation on a large population using msGBS making it advantageous compared to array and other sequencing techniques. For example, in humans the Illumina HumanMethylation450 BeadChip is a popular method for epigenome studies [6] but relies on prior knowledge of individual CpG sites that are contained on the array. This contrasts with msGBS which only relies on the action of restriction enzyme cut sites, and therefore offers an unbiased approach to annotating methylated sites as opposed to microarrays has the ability to annotate more methylated sites in the human genome. Furthermore, msGBS also offers an unbiased approach to annotating methylated sites as opposed to microarrays. Compared to other sequencing approaches designed to identify methylation such as whole-genome bisulfite sequencing (WGBS) or methylation-capture techniques, msGBS infers methylation via read coverage and does not require additional sample treatment to convert methylated cytosines (i.e. Sodium bisulfite treatment). Determining read coverage, as opposed to WGBS and array methods, enables a library preparation step that avoids treatment with sodium bisulfite, which damages and fragments input DNA [7].

While msGBS protocols have a wide-variety of uses, current analysis approaches lack the ability to quantify methylation on a methylation-sensitive panel. Here in this paper, we

present msgbsR, an R package for the analysis of data obtained from msGBS experiments. Our pipeline allows researchers to conduct analyses of msGBS experiments, in order to identify differentially methylated sites. msgbsR includes tools such as reading the read counts from a sorted and indexed genome alignments or BAM file(s) directly into the R environment, checking that the cut sites match the RE sequence, identifying differential methylated sites, and seamless annotation using available reference genomes in the R/Bioconductor framework. To demonstrate the utility of the msGBS approach, and the msgbsR analysis package, we analysed a population of rats (control vs treatment) for differential DNA methylation (Rattus norvegicus), and two publicly available agricultural crop datasets from barley (Hordeum vulgare) and maize (Zea mays) to show the extensive potential applications in epigenetic research.

6.2 Results

6.2.1 Generating the table of read counts

Using a reference genome, alignment of the sequencing data produced from an msGBS experiment results in reads that begin at RE cut sites. As a result reads will generally begin at a defined RE cut site, producing a pileup of reads at those genomic positions (Figure 6-1A). Thus, it is possible to count the total number of reads that mapped to these RE cut site positions. The msgbsR analysis pipeline firstly starts with functions allowing the import and verification of raw read counts at the RE cut sites that were produced in a msGBS experiment by reading sorted and indexed BAM file(s) (Figure 6-1B). These sorted and indexed BAM files can be the output from an alignment tool such as Bowtie2 [8] and SAMtools [9]. The rawCounts function takes a list of sorted and indexed BAM files and imports the raw read counts into the R environment. The msgbsR package utilises other Bioconductor packages to ease data analysis, with rawCounts being directly applicable to the data format of a RangedSummarizedExperiment from the Bioconductor package SummarizedExperiment [10]. The resulted RangedSummarizedExperiment data object contains a table of read counts of potential cut site locations with their genomic coordinates, such as the chromosome, position and strand information.

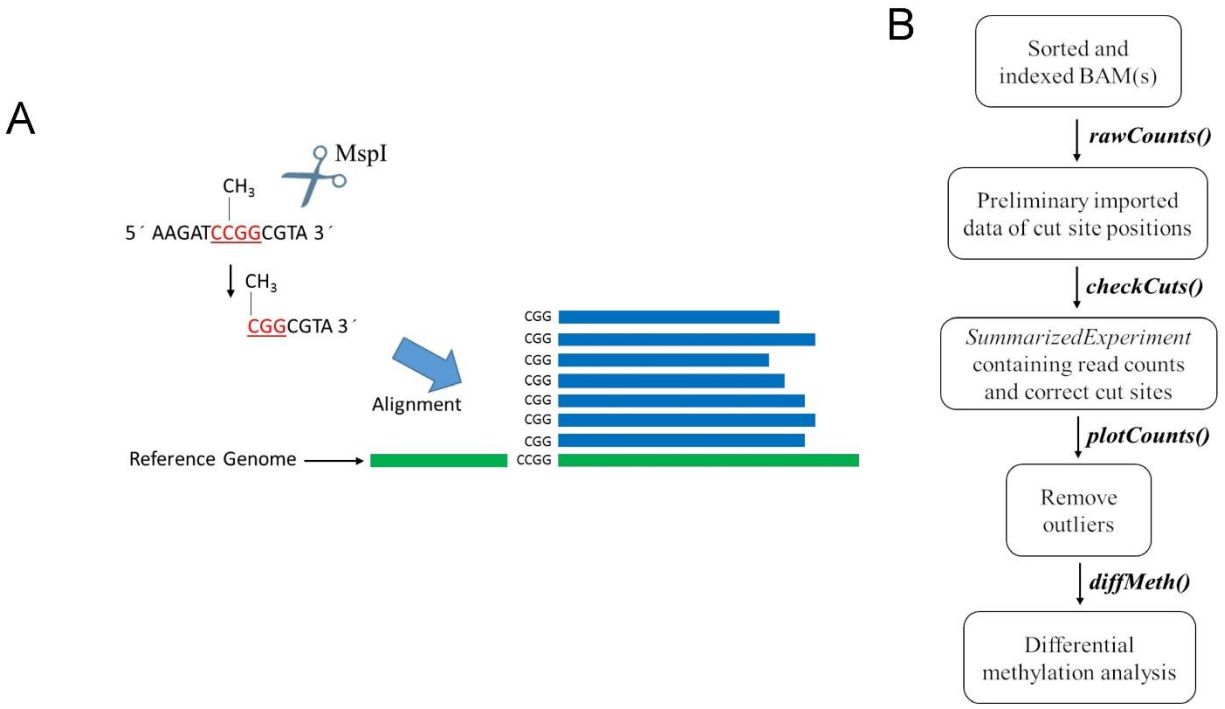


Figure 6-1: A simplified schematic of methylation-sensitive genotyping-by-sequencing (msGBS) and the msgbsR pipeline.

(**A**) An example of msGBS using the restriction enzyme, *MspI*, which cleaves DNA at the recognition sequence C^CGG if the internal cytosine is methylated. However, *MspI* does not cut at the recognition site if both cytosines are methylated or the external cytosine is methylated. (**B**) The data analysis pipeline represented by a flowchart which highlights the names of the main functions in the msgbsR package.

The output of the rawCounts function uses the start position of all mapped reads in a BAM file. However, there may be incorrectly mapped reads that do not correspond to a specified RE recognition sequence. Incorrectly mapped reads can be filtered out of the analysis prior to any downstream analyses using the checkCuts function, which takes a GRanges data object that contains the positioning of the potential cut sites and the recognition sequence of the RE. The checkCuts function then uses a reference genome in the format of a BSgenome which is obtainable from Bioconductor. However, if a BSgenome is unavailable, a user-defined FASTA file can also be used to determine if the recognition sequence matches in the reference genome. checkCuts then returns a GRanges object with the correct positions of the cut sites while incorrectly mapped reads can then be filtered out of the RangedSummarizedExperiment.

6.2.2 Package Validation

We performed the msgbsR pipeline on our own msGBS data set. Our msGBS data set focused on the prostates tissue from the offspring of rats who were either fed a control (n = 26) or experiment high fat maternal diet (n = 18). This experiment focused on using the methylation sensitive RE, MspI, which cleaves at the recognition site C^CGG (Figure 6-1A). Initially, after mapping there were a potential of 1,616,611 MspI, sites. However, after running checkCuts, this was reduced to 1,252,042 MspI sites. The incorrectly mapped reads were unique to an individual sample. In other words the same incorrect site did not occur in multiple samples. We therefore found it advantageous to have the function checkCuts, as it can remove incorrectly mapped reads which may have been introduced in an earlier step prior to running the msgbsR pipeline. By running the checkCuts function ensures there are no incorrect mapped reads within any downstream analyses. This is important as incorrect sites can impact downstream analyses such returning sites that are differentially methylated but are in fact false positives.

We also used msgbsR on a publicly available msGBS experiment focusing on barley and maize (SRP004282.1) leaf samples [2]. This experiment used ApeKI, a methylation-sensitive endonuclease that recognizes the 5 bp sequence GCWGC (W = A or T). Firstly,

we mapped the barley and maize samples to their respective available reference genomes (see methods) and used the rawCounts function on the resulted sorted and indexed BAM files to determine count numbers. Initially, this resulted in a total of 4,081,975 and 1,155,762 potential ApeKI sites for the maize and barley data set respectively. However, after running the checkCuts function this was reduced to 3,791,316 and 1,032,360 cut sites for the maize and barley data set respectively. This was potentially due to incorrectly mapped reads and ensured all downstream analyses were performed using sites that were correctly mapped.

6.2.3 Visualisation

msGBS experiments can produce varying numbers of cut sites and reads depending on the DNA methylation state and the efficiency of the library preparation step for each individual sample. A way to overcome false positives associated with differences in read numbers between samples is to remove samples that produced a low number of reads and/or cut sites. This can be done before performing differential methylation analysis using the plotCounts function incorporated in the msgbsR package. Figure 6-2A and 6-2B were generated using the plotCounts function and show the total number of reads compared to the total number of cut sites produced for each individual sample from the publicly available data set described above [2]. We also performed this function on our own data set focusing on prostates from rat offspring from either a control or experimental high fat maternal diet. As shown in Figure 6-3A, there are potential outliers which can be removed prior to performing any differential methylation analyses. For example, there were 7 samples that produced < 1 million MspI cut sites and 4 samples that produced > 6 million reads. Ideally, msGBS sequencing should be performed multiple times to produce technical replicates enabling us to determine if outliers were introduced as a result during sequencing. For the purpose of this experiment, sequencing was performed once and we therefore removed these outliers prior to differential methylation analysis.

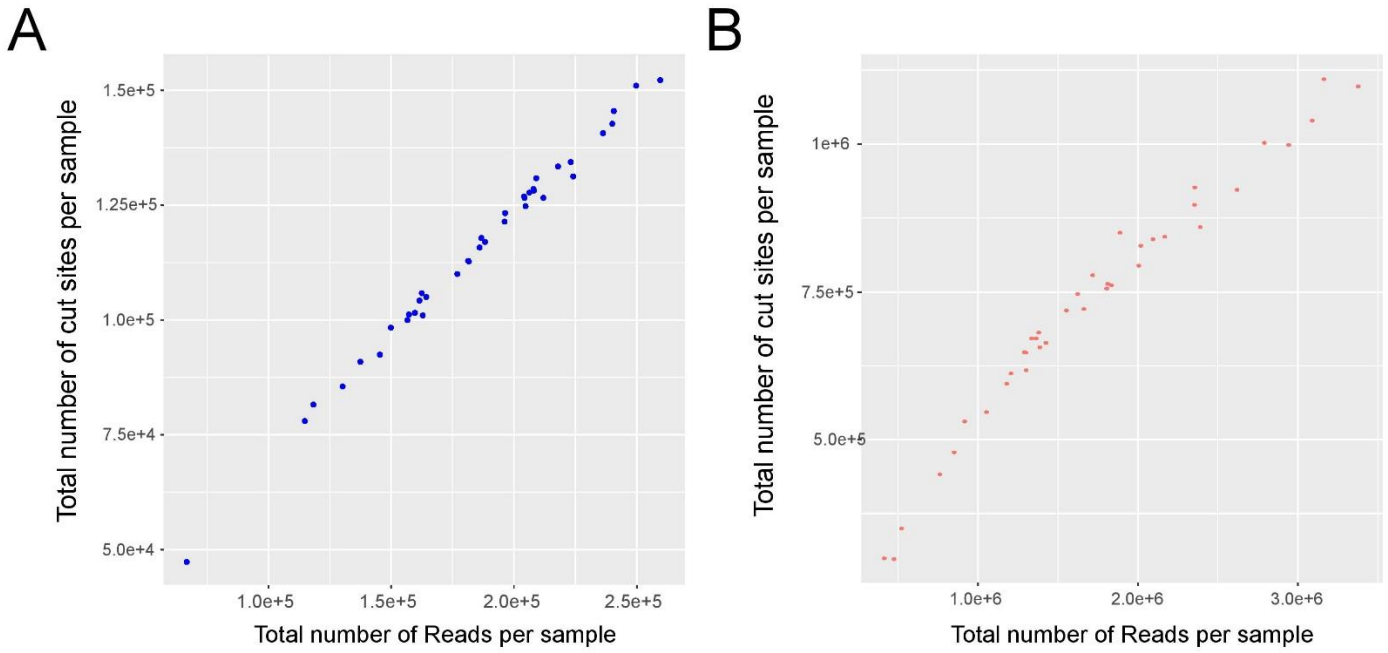


Figure 6-2: The output of the plotCounts function.

The distribution of the library size compared to the total number of *ApeKI* cut sites produced for each sample from either the (**A**) barley or (**B**) maize data set. Each individual point represents a unique sample.

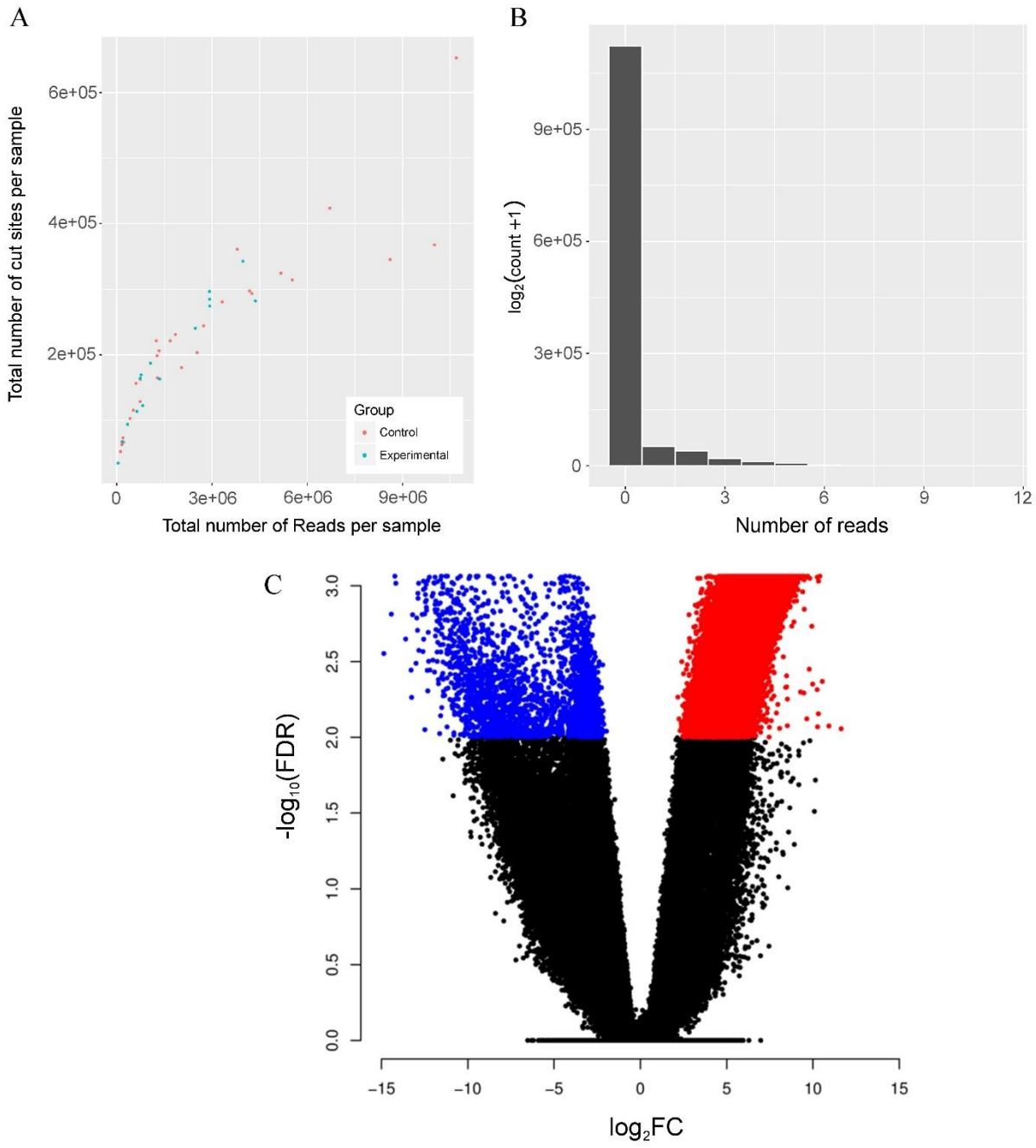


Figure 6-3: The msgbsR pipeline on our rat prostate msGBS data.

(**A**) Output of the *plotCounts* function showing the distribution of the total number of reads and cut sites per sample. Samples are coloured depending on their diet group. (**B**) A histogram of reads for a control sample showing a negative distribution. (**C**) A volcano plot showing differentially methylated sites (FDR < 0.01) between the control diet (blue dots) and the experimental diet (red dots).

6.2.4 Differential methylation analysis

One of the advantages of msGBS experiments is the ability to sequence hundreds of samples from different groups or conditions, and thus increase statistical power in analyses. However, when experiments have numerous groups or conditions it can become time consuming to perform multiple analyses. The msgbsR package contains a function that automates normalisation and determines differentially methylated sites between groups. Since the data generated from a msGBS experiment is in the form of read counts, we can take advantage of tools typically used in RNA-seq analyses [11]. The diffMeth function uses edgeR [12] tools to automate splitting the data, perform normalisation and identify differentially methylated sites. We choose to use edgeR, since it works on the assumption that the read count data distribution is negative binomial. msGBS data is too negative binomially distributed (Figure 6-2B) which is represented in a random control sample from our own msGBS data. We performed differential methylation analysis using the diffMeth function, and found 31,768 sites to be differentially methylated (Figure 6-2C). In other msGBS experiments there is generally more than one comparison. Thus, this function can quickly perform differential methylation analyses from an experiment where there is a large number of groups or conditions.

6.3 Discussion

The advancement of high throughput technologies has enabled varying sequencing techniques. However, there is a limited number of bioinformatics tools available for the analysis of all the available sequencing protocols. msGBS is a reduced representation of whole genome sequencing which can be used to study DNA methylation and parts of the genome that are normally inaccessible in other sequencing technologies [1]. However, there is a current lack of bioinformatics tools that are tailor made for the analysis of msGBS experiments within the literature. Here in this study we outline msgbsR, an R package which can be used in part of the pipeline in analysing msGBS experiments. Our package works by identifying methylated sites and read counts directly from sorted and indexed BAM files into the R environment and can verify if the reads have mapped

correctly to the recognition site of the RE by using a reference genome in the format of either a BSgenome or FASTA file.

Reduced representation sequencing conducted in msGBS enable a larger number of samples to be sequenced, making this a more desireable methylation analysis platform compared to high-resolution protocols such WGBS. For example, in an agricultural setting it may be used to assess both genetic and epigenetic variation over mapping populations [13] or for assessing the epigenetic impact of breeding populations in new environments [14]. In a medical setting the DNA methylation data can be used to make group comparisons [15]. Furthermore, single nucleotide polymorphisms (SNPs) data can also be obtained from msGBS data, thereby making this approach essential for genome-wide and epigenome-wide associations studies at the same time. msgbsR can also be used with non-methylation sensitive GBS to verify reads have been mapped correctly and to determine if there are any differences in read counts between groups, allowing it to be used in conjunction with other Bioconductor packages for assessing genetic variation, such as GWAStools [16].

Differential methylation can be performed using msgbsR which contains a wrapper function using edgeR [12]. We choose to make a wrapper function of edgeR since msGBS experiments typically contain samples from multiple groups. Performing differential methylation analyses can become time consuming especially when there are multiple comparisons to consider. Our wrapper function uses the recommend trimmed mean of M-values (TMM) normalisation method suggested by edgeR [17]. However, we do acknowledge users may wish to use other bioinformatics tools to perform differential methylation analyses. Users may want to perform other normalisation methods or use other downstream packages such as methylSig [18] a package designed to identify differentially methylated sites and regions. There are also other Bioconductor packages such as BiSeq [19] or DSS [20] which have been too can be used identify differentially methylated sites and regions. However, these tools have been primarily designed to work with whole genome bisulphite (WGBS) sequencing whereby methylation is determined through the proportion of methylated and un-methylated reads, and may not necessarily

fulfil the user requirements when working with msGBS data. The output of msgbsR after reading the sorted and indexed BAM file(s) is SummarizedExperiment, a data format which is compatible with other Bioconductor packages. Therefore, users can then use other packages within their own msGBS pipeline after msgbsR has read the data into the R environment.

The msgbsR package contains a variety of functions to automate the data analysis of a msGBS experiment. The input and output of the functions used in the msgbsR package are compatible with Bioconductor packages such as BSgenomes, and edgeR [12]. Furthermore, msgbsR is fully documented, contains a tutorial data set and is freely available from Bioconductor.

6.4 Methods

6.4.1 Library preparation and sequencing of rat msGBS

DNA was extracted from prostates and then digested with EcoRI and MspI using the MSAP technique [21, 22]. EcoRI is a restriction enzyme and recognises the sequence G^AATTC and is not methylation sensitive. Illumina sequencing primer adapters were ligated to the digested genomic DNA. Using a technique as previously described [23, 24], cycling was performed using a BioRad 100 thermocycler at 37˚C for 2 hours followed by enzyme inactivation for 20 min at 65˚C. Barcoded adapters were designed with an MspI overhang and a common Y adapter with an EcoRI overhang using the script by Thomas P. van Gurp (www.deenabio.com/services/gbs-adapters) and were ligated as previously described [23]. T4 ligase (200U) and T4 ligase buffer (NEB T4 DNA Ligase #M0202) along with 0.1 ρmol and 15 ρmol of the barcoded MspI adapter and EcoRI adapter respectively. The reaction mixture was incubated at 22˚C for 2 hours and then 65˚C for 20 mins for enzyme inactivation. 5µL from each ligation reaction were pooled together and then divided into equal volumes for column clean-up using the PureLink PCR Purification Kit (Life Technologies). Samples were then pooled back together for a total of 60µL in molecular biology grade water. PCR reactions were performed in a 25µL volume with 10µL of digested DNA, 5µL of 5x NEB MasterMix, 2µL of 10µM Forward

and Reverse primers at 10µM. PCR cycle reactions (Solexa) were performed at 98˚C for 30 seconds, followed by 16 cycles of 98˚C for 30 seconds, 62˚C for 20 seconds and 72C for 30 seconds and finally 72˚C for 5min. Size selection of fragments was performed using Ampure XP magnetic beads (Beckman). Fragments were captured and eluted into 30µL of water. Samples were sequenced using an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA, USA) at the Queensland Brain Institute (QBI).

6.4.2 Publicly available data set

The publicly available data set (SRP004282.1) used to demonstrate several functions of msgbsR was firstly obtained from the Sequence Read Archive (SRA) [25]. SRA files were then converted to FASTQ files using the SRA tool kit [26]. This study contained two data sets containing samples from either barley or maize leaves [2]. Both data sets were demultiplexed using specific barcodes provided within the study [2] and GBSX [27]. This resulted in each individual sample from each data set in a FASTQ format.

6.4.3 Processing of sequencing data

Alignment of reads was performed using bowtie2 v2.2.3 [8] to each respective reference genome. We used the latest barley reference genome (ASM32608v1) which was obtained from the plant Ensembl website (plants.ensembl.org/Hordeum_vulgare/). For maize, we used the Ensembl release (AGPv4) which we obtained from the Illumina iGenomes website. For the Rat data we used UCSC latest release (rn6) which was obtained from the Illumina iGenomes website. Alignment with bowtie2 resulted in BAM files which were then sorted and indexed using SAMtools v1.2 [9]. The sorted and indexed BAM files were then directly read into the R environment using the rawCounts function within the msgbsR package enabling downstream analyses with msgbsR. Since the offspring were from some of the same mothers, differential methylation was performed using the mother as a blocking factor.

References

1. He J, Zhao X, Laroche A, Lu Z-X, Liu H, Li Z: **Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding.** *Frontiers in Plant Science* 2014, **5:**484.
2. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES: **A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.** *PLoS One* 2011, **6**.
3. Li D, Zhang B, Xing X, Wang T: **Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation.** *Methods* 2015, **72:**29-40.
4. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA: **Stacks: an analysis tool set for population genomics.** *Mol Ecol* 2013, **22**.
5. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES: **TASSEL: software for association mapping of complex traits in diverse samples.** *Bioinformatics* 2007, **23:**2633-2635.
6. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Djik S, Muhlhausler B, Stirzaker C, Clark SJ: **Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling.** *Genome Biol* 2016, **17:**208.
7. Plongthongkum N, Diep DH, Zhang K: **Advances in the profiling of DNA modifications: cytosine methylation and beyond.** *Nat Rev Genet* 2014, **15:**647-661.
8. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.
9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
10. Morgan M OV, Hester J and Pagès H: **SummarizedExperiment: SummarizedExperiment container. R package version 1.6.0.** 2017.
11. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A: **A survey of best practices for RNA-seq data analysis.** *Genome Biol* 2016, **17:**13.
12. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
13. Xiong W, Li X, Fu D, Mei J, Li Q, Lu G, Qian L, Fu Y, Disi JO, Li J, Qian W: **DNA Methylation Alterations at 5′-CCGG Sites in the Interspecific and Intraspecific Hybridizations Derived from Brassica rapa and B. napus.** *PLoS One* 2013, **8:**e65946.
14. Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A: **Genotyping by sequencing for genomic prediction in a soybean breeding population.** *BMC Genomics* 2014, **15:**740.
15. Nazarenko MS, Markov AV, Lebedev IN, Freidin MB, Sleptcov AA, Koroleva IA, Frolov AV, Popov VA, Barbarash OL, Puzyrev VP: **A Comparison of Genome-Wide DNA Methylation Patterns between Different Vascular Tissues from Patients with Coronary Heart Disease.** *PLoS One* 2015, **10:**e0122601.

16. Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, Zheng X, Crosslin DR, Levine D, Lumley T, et al: **GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies.** *Bioinformatics* 2012, **28:**3329-3331.
17. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
18. Park Y, Figueroa ME, Rozek LS, Sartor MA: **MethylSig: a whole genome DNA methylation analysis pipeline.** *Bioinformatics* 2014, **30:**2414-2422.
19. Hebestreit K, Dugas M, Klein HU: **Detection of significantly differentially methylated regions in targeted bisulfite sequencing data.** *Bioinformatics* 2013, **29:**1647-1653.
20. Feng H, Conneely KN, Wu H: **A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data.** *Nucleic Acids Res* 2014, **42:**e69.
21. Reyna-Lopez GE, Simpson J, Ruiz-Herrera J: **Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi by amplification of restriction polymorphisms.** *Mol Gen Genet* 1997, **253:**703-710.
22. Rodríguez López CM, Morán P, Lago F, Espiñeira M, Beckmann M, Consuegra S: **Detection and quantification of tissue of origin in salmon and veal products using methylation sensitive AFLPs.** *Food Chem* 2012, **131:**1493-1498.
23. Poland JA, Brown PJ, Sorrells ME, Jannink JL: **Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach.** *PLoS One* 2012, **7:**e32253.
24. Xia Z, Zou M, Zhang S, Feng B, Wang W: **AFSM sequencing approach: a simple and rapid method for genome-wide SNP and methylation site discovery and genetic mapping.** *Sci Rep* 2014, **4:**7300.
25. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database C: **The Sequence Read Archive.** *Nucleic Acids Res* 2011, **39:**D19-D21.
26. **SRA Knowledge Base [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK56551/.**
27. Herten K, Hestand MS, Vermeesch JR, Van Houdt JK: **GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments.** *BMC Bioinformatics* 2015, **16:**1-6.

Statement of Authorship

Title of Paper	A prognostic DNA methylation signature for pregnancy complications
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Benjamin Mayne
Contribution to the Paper	Analysed and interpreted the data and wrote the manuscript.
Overall percentage (%)	80
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 14-7-17

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Shalem Leemaqz
Contribution to the Paper	Analyzed the data, provided intellectual input into the manuscript.
Signature	Date 14 July 2017

Name of Co-Author	Claire Roberts
Contribution to the Paper	Conceived the initial part of the study, involved in the study design and provided intellectual input into the manuscript.
Signature	Date 14.4.17

Name of Co-Author	James Breen
Contribution to the Paper	Provided critical discussion and intellectual input into the manuscript.
Signature	Date 20/7/2017

Name of Co-Author	Carlos Rodriguez Lopez
Contribution to the Paper	Conceived the initial part of the study, involved in the study design and provided intellectual input into the manuscript.
Signature	Date 25/07/2017

Name of Co-Author	Tina Blanco-Miotto
Contribution to the Paper	Conceived the initial part of the study, involved in the study design and provided intellectual input into the manuscript.
Signature	Date 26/7/17

7 A prognostic DNA methylation signature for pregnancy complications

BENJAMIN T MAYNE, SHALEM LEEMAQZ, CLAIRE T ROBERTS, JAMES BREEN, CARLOS RODRIGUEZ LOPEZ, TINA BIANCO-MIOTTO

Abstract

DNA methylation is an epigenetic modification well known for regulating gene expression. DNA methylation changes provide good prognostic and diagnostic biomarkers for a range of health issues, as well as being used to determine an individual's age and probability of cancer relapse. In this study, we focused on DNA methylation as a prognostic tool to identify women who subsequently develop a pregnancy complication. We assessed DNA methylation using a methylation-sensitive genotyping by sequencing (msGBS) assay using DNA extracted from circulating maternal leukocytes at 15 weeks' gestation from women who went on to have an uncomplicated healthy pregnancy or developed gestational diabetes mellitus (GDM), preeclampsia (PE), preterm birth (PTB) or delivered a small for gestational age (SGA) baby. msGBS provided DNA methylation data for 2,610,160 sites from 402 women. Using a recursive feature-elimination method, we built binary classification models for each pregnancy complication using uncomplicated pregnancies as a reference group. In total, we identified 84 methylated sites that can distinguish women having either a GDM, PE, PTB or SGA pregnancy. We were also able to build other predictive models of other clinical characteristics such as maternal smoking and BMI. This study highlights the potential of DNA methylation from maternal blood as a non-invasive, prognostic tool in predicting pregnancy complications.

7.1 Introduction

Pregnancy complications such as gestational diabetes mellitus (GDM), preeclampsia (PE), preterm birth (PTB) and small for gestational age (SGA) are major causes of maternal and perinatal morbidity and mortality. Pregnancy complication rates vary worldwide, the most common, GDM, affects up to 20% of women worldwide [1]. PE is a

hypertensive disorder of pregnancy which occurs after 20 weeks' gestation and affects 2-10% of pregnancies worldwide [2]. PTB occurs when infants are born before 37 weeks' gestation and affects up to 11% of pregnancies worldwide [3]. IUGR, defined as an infant born < 5th percentile of birth weights affects up to 3-7% pregnancies worldwide [4].

Prediction of a pregnancy outcome prior to the onset of a complication would be clinically very useful to reduce or possibly prevent any maternal or perinatal morbidity or mortality. Previous prediction models have included the use of Doppler ultrasound approaches, maternal serum levels of beta human chorion gonadotrophin (β-hCG), inhibin A and pregnancy-associated plasma protein A (PAPPA), placental growth factor [5-11]. However, in other biomedical settings analysis of DNA methylation is proving to be an appropriate candidate for disease outcome prediction. In this study, we focused on the most widely studied epigenetic modification, 5' methylcytosine DNA methylation, where DNA is methylated via the addition of a methyl group to a cytosine. DNA methylation is essential for development [12] and is influenced by both genetic [13] and environmental factors [14]. Epigenetic marks, including DNA methylation, can contribute to identification of complex phenotypes and diseases.

DNA methylation has proven to be an effective biomarker for disease prediction and clinical characteristics. It has been used to predict age [15, 16], cardiovascular disease [17], cancer [18, 19], mortality [20] and other clinical characteristics such as obesity and smoking [21]. In a reproductive setting it has also been used to predict gestational age from cord blood [22] and placental tissue [23]. However, no study has used DNA methylation as a biomarker for the prediction of pregnancy complications.

In this study, we used a methylation sensitive genotyping by sequencing (msGBS) approach to firstly assess DNA methylation in leukocytes isolated from maternal blood. Extracting leukocytes from the maternal blood, which are in a high abundance to be profiled properly, is a non-invasive method to profile pregnant women. The latter was obtained from women who participated in the Screening fOr Pregnancy Endpoints (SCOPE) study at 15 weeks' gestation [24]. Using this DNA methylation data set we were able to build predictive models of pregnancy complications, including, GDM, PE,

PTB and SGA. Our approach has identified selected methylated sites that can predict pregnancy outcome and other clinical characteristics.

7.2 Methods

7.2.1 Pregnancy data set

Ethics approval was obtained from the local ethics committee (REC 1712/5/2008) and all women provided written informed consent. Women in this study were recruited from the Lyell McEwin Hospital, South Australia. Women were recruited into the SCOPE study from November 2004 to February 2011 and were nulliparous and were having singleton pregnancies. In total, blood collected at 15 weeks' gestation from 402 women was used in this study. Clinical outcomes and measurements of the participants (Table 7-1) were collected as part of the SCOPE study. Body mass index (BMI) was collected at 15 weeks gestation, the same time as the collection of the blood. GDM was defined using a Glucose Tolerance Test with a reading of 5.5 mM or higher when fasting or a 2 h level of 8 mM or a random glucose level of > 11 mM. PE was defined as systolic blood pressure ≥140 mmHg and/or diastolic blood pressure ≥90 mmHg post 20 weeks' gestation occurring at least twice 4 h apart with proteinuria (24 h urinary protein ≥ 300 mg). PTB was defined as birth occurring prior to 37 weeks' gestation. SGA was defined as birthweight <10th customised centile.

Table 7-1: Characteristics of participants used for the final DNA methylation data set after outliers were removed. Characteristics are represented as mean and standard deviation unless detailed otherwise.

	Uncomplicated	**Gestational Diabetes Mellitus**	**Preeclampsia**	**Preterm Birth**	**Small for Gestational Age**
N	178	31	54	48	74
Maternal Age (years)	23.5 ± 5.2	28.5 ± 4.7	23.3 ± 4.2	24.1 ± 6.0	24.3 ± 5.7
Gestational Age at delivery (weeks)	40.1 ± 1.1	39.1 ± 1.1	38.9 ± 1.9	32.9 ± 4.5	40.0 ± 1.4
Fetal Sex (F/M)	93/85	17/14	29/25	21/27	36/38
Birth Weight (g)	3570 ± 387	3316 ± 444	3417 ± 536	2164 ± 851	2711 ± 356
Birth Length (cm)	50.1 ± 1.8	48.8 ± 2.0	49.5 ± 2.5	42.1 ± 9.5	47.3 ± 2.1
Smoking (%)	21.9	9.6	5.6	37.5	51.4

7.2.2 DNA isolation

Buffy coats were isolated from EDTA blood samples by adding proteinase K (20µg/µL) and inverting, followed by the addition of 20% Sodium Dodecyl Sulphate (SDS). The resuspended solution in TES solution was then incubated for 24 h at 37˚C followed by the addition of an equal volume of 3M NaCl and mixed vigorously. The solution was precipitated on ice and then centrifuged at 13,000 rpm for 10 mins. The supernatant was removed and combined with 2x the volume of 100% ethanol. The DNA was then precipitated and washed with 70% ethanol. The pellet of DNA was then dried and re-dissolved in Tris-EDTA (TE) buffer. DNA was assessed and quantified on a Thermo Fisher Scientific NanoDropTM 1000 Spectrophotometer.

7.2.3 Library Preparation and sequencing

Using a modified methylation sensitive amplified polymorphisms (MSAP) technique [25, 26], DNA was digested with EcoRI and MspI (Figure 7-1). MspI is a restriction enzyme which cleaves DNA at C^CGG, only when the internal cytosine is methylated. EcoRI is a restriction enzyme and recognises the sequence G^AATTC. Illumina sequencing primer adapters were ligated to the digested genomic DNA. Using a technique previously described [27, 28], reactions were performed in a 96 well plate (95 samples, 1 water as a control). Cycling was performed using a BioRad 100 thermocycler at 37˚C for 2 h followed by enzyme inactivation for 20 min at 65˚C. 96 barcoded adapters (Supplementary Table 7-1) were designed with an MspI overhang and a common Y adapter with an EcoRI overhang using the script by Thomas P. van Gurp (www.deenabio.com/services/gbs-adapters). A 40µL ligation reaction as described by [27], was carried out using the same 96 well plate. T4 ligase (200U) and T4 ligase buffer (NEB T4 DNA Ligase #M0202) along with 0.1 ρmol and 15 ρmol of the barcoded MspI adapter and EcoRI adapter respectively. The reaction mixture was incubated at 22˚C for 2 h and then 65˚C for 20 mins for enzyme inactivation. 5µL from each ligation reaction were pooled together and then divided into equal volumes for column clean-up using the PureLink PCR Purification Kit (Life Technologies). Samples were re-pooled for a total of 60µL in molecular biology grade water. PCR reactions were performed in a 25µL volume

with 10µL of digested DNA, 5µL of 5x NEB MasterMix, 2µL of 10µM Forward and Reverse primers at 10µM. PCR cycle reactions (Solexa) were performed at 98˚C for 30 s, followed by 16 cycles of 98˚C for 30 s, 62˚C for 20 s and 72˚C for 30 s and finally 72˚C for 5min. Size selection of fragments was performed using Ampure XP magnetic beads (Beckman). Fragments were captured and eluted into 30µL of water. Samples were sequenced using an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA, USA) at the Queensland Brain Institute (QBI). To achieve a high sequencing depth and coverage, samples were sequenced three times and sequencing data were pooled together for each sample.

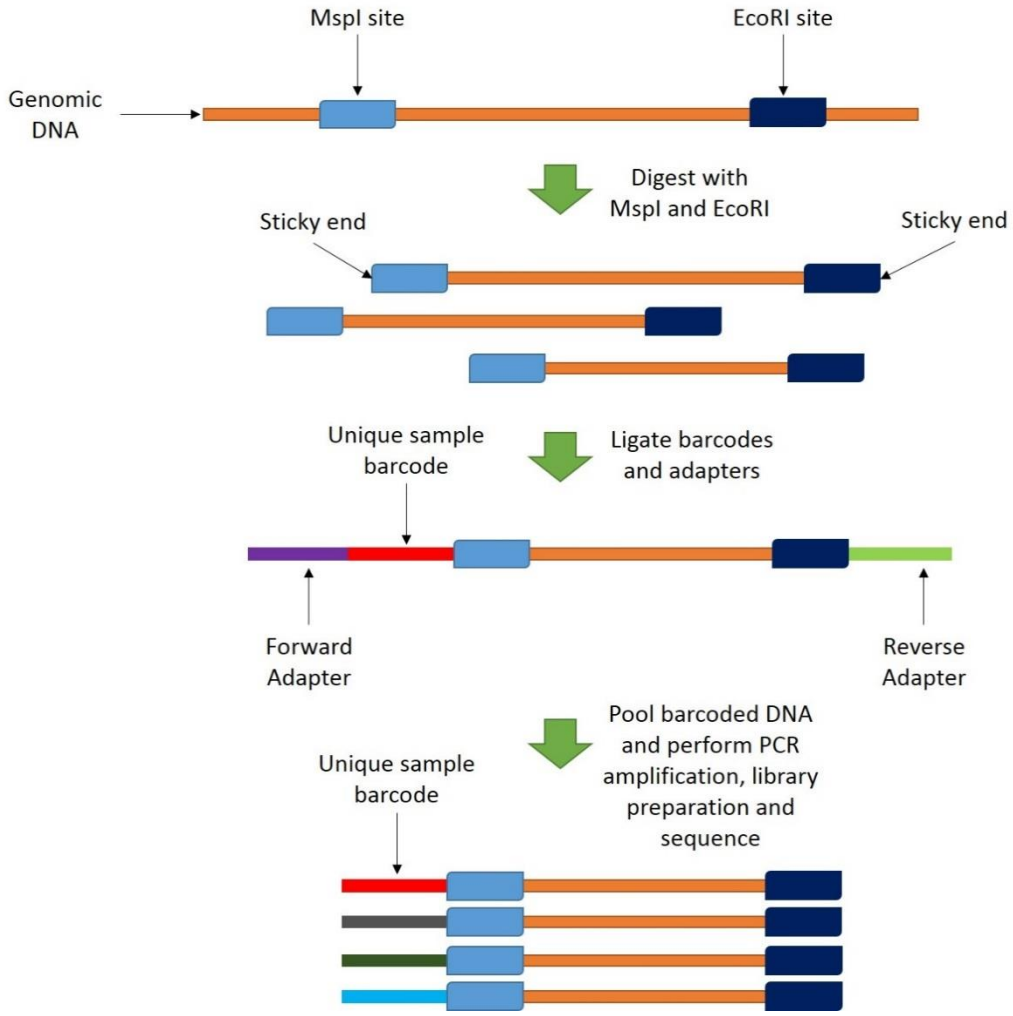


Figure 7-1: Library preparation of msGBS prior to sequencing and analysis.

7.2.4 Processing raw sequencing data

Samples were demultiplexed using specific barcodes for each sample using GBSX [29]. Paired end samples were merged together using BBMerge [30] with a minimum overlap of 25bp. Reads were aligned to the human reference genome (UCSC hg38) using bowtie2 v2.2.3 [31]. The MspI cut sites generated from this experiment were located and confirmed using the msgbsR package [32]. The MspI cut sites were annotated using a gff3 file of the human genome which was obtained from Ensembl [33].

7.2.5 Single nucleotide polymorphism overlap

Since single nucleotide polymorphisms (SNPs) can disrupt the recognition site by altering the sequence, we determined the overlap of MspI cut sites produced compared to common SNPs within the human population using dbSNP (SNP147) [34]. This database contains approximately 150 million SNPs that are present within at least 1% of the human population [34]. Overlap of SNPs with MspI cut sites was performed using the R package GenomicFeatures [35].

7.2.6 Principal component linear discriminant analysis plots and feature selection

Principal component linear discriminant analysis (PC-LDA) plots were generated by firstly performing a linear discriminate analysis on all MspI cut sites using the FIEmspro R package [36]. We performed feature selection to reduce the total number of MspI cut sites in order to identify markers of clinical outcomes and characteristics. Feature selection was performed using the FIEmspro R package where the total number of MspI cut sites selected was equal to 1% of the total number of samples for each comparison, thereby selecting the most important sites and reducing overfitting. We then used these reduced forms of the data to build predictive models of clinical outcomes and characteristics.

7.2.7 Building predictive models of categorical clinical outcomes and characteristics

In order to develop predictive models we firstly divided the data into a training and testing data set to build and train our models (Figure 7-2). 70% of the samples, with each specific pregnancy outcome and from the uncomplicated term group, were randomly assigned to a training data set, with the other remaining samples being assigned to a testing data set. In order to identify which MspI cut sites can be used to predict clinical outcomes and characteristics, we used a recursive feature elimination (RFE) method implemented in the caret R package [37]. The RFE method firstly fits a model with all predictors (MspI cut sites), then after multiple iterations returns the model with the best performance and the minimum number of predictors required. We also performed the RFE method using a repeated cross validation and with different functions including linear discriminant analysis, random forests, naïve Bayes and bagged trees. We then selected the best performing function, which we defined with the highest accuracy in the testing data set after 10-fold cross validation. This resulted in a model with the minimum MspI cut sites required to predict a clinical outcome or characteristic. The ROCR R package [38] was used to calculate area under the curve (AUC) values for receiver operating characteristic (ROC) curves.

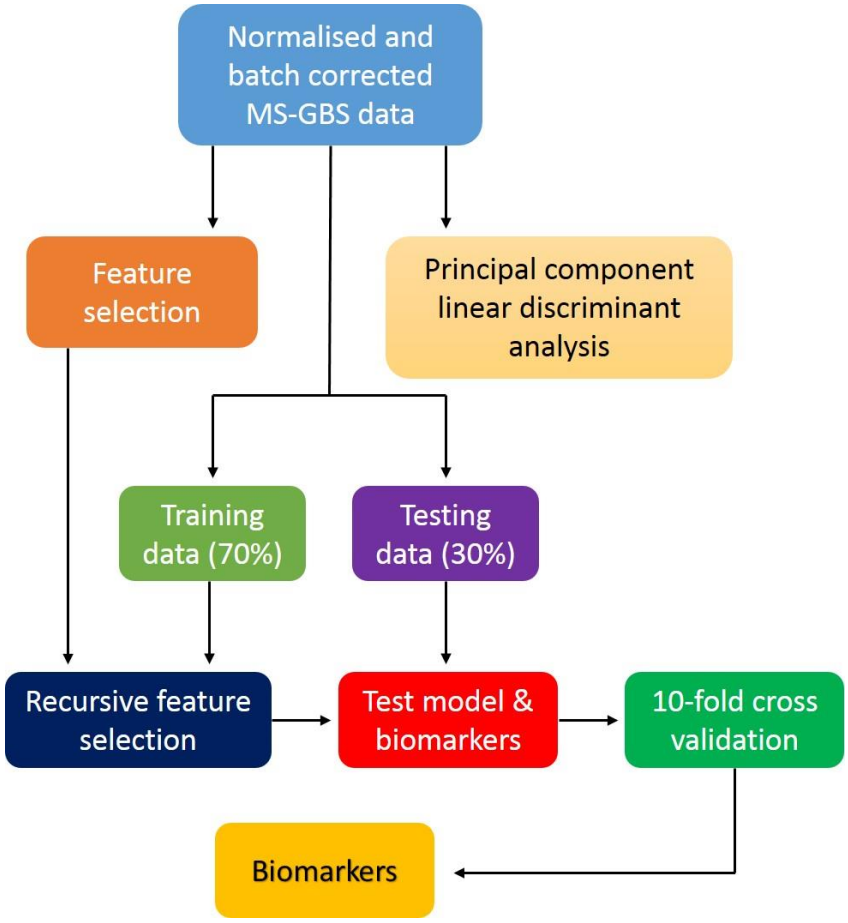


Figure 7-2: Flow diagram of model construction to identify biomarkers of pregnancy outcome and other clinical characteristics or outcomes.

7.2.8 Building predictive models of continuous characteristics

In order to build models that can assess a continuous characteristic such as maternal age, we used the glmnet R package [39]. The data was firstly divided into training (70%) and testing (30%) data sets and using an elastic net regression method, the normalised and batch corrected methylation data was regressed against a continuous characteristic. The elastic net regression model automatically selects the minimum number of sites that are required to accurately predict an outcome. These sites were then used in the testing data set to assess a model's performance.

7.2.9 Functional Annotation

Functional annotation was conducted using the closest gene to a cut site which was then used to identify enriched Gene Ontology (GO) terms. Annotation was performed using the hg38 reference genome obtained from Ensembl [40]. GO enrichment was conducted using all human genes in the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 [41] and g:Profiler [42].

7.3 Results and Discussion

7.3.1 DNA methylation data set

On average, 1,418,372 reads were aligned to the human reference genome per sample with an average alignment rate of 82.6%. The resulting table of read counts contained a total of 2,610,160 MspI cut sites across all the autosomal chromosomes and the X chromosome (Supplementary Figure 7-1). We removed outliers that generated >20 million reads (1 sample) and <100,000 and >600,000 cut sites (13 samples). Ideally, these samples would then have to be sequenced again. This resulted in a DNA methylation data set of 388 samples (Uncomplicated = 178, GDM = 31, PE = 54, PTB = 48, SGA = 74, Multiple Complications = 3), which was used for downstream analyses. On average, 303,052 methylated sites were produced per sample. The average (mean) distance from the start site of a gene was 500,511 bp and the total number of sites per chromosome are supplied in Supplementary Table 7-2.

7.3.2 Removal of potential SNP-driven differences

SNPs can often disrupt restriction enzyme recognition sites, so we set out to remove MspI cut sites which overlapped with a SNP that was present in at least 1% of the population. We obtained the full database of common SNPs (see methods) within the human population and determined the overlap with the MspI cut sites generated in our DNA methylation dataset. In total, 32% of all MspI cut sites overlapped with a common SNP. These were all removed from the data. We then constructed models to predict pregnancy complications and other clinical characteristics using this methylation data set.

7.3.3 DNA methylation at 15 weeks' gestation is a biomarker of subsequent pregnancy outcome

After the removal of confounding SNP sites, we initially built predictive models of pregnancy complications using all remaining MspI cut sites generated from all chromosomes. However, due to high variability of DNA methylation in X chromosome inactivation [43-45], we found it advantageous to remove the X chromosome sites from the analysis. This is because X inactivation is not consistent within each sample and is random. This left a total of 1,665,641 MspI cut sites on autosomal chromosomes to build predictive models of pregnancy complications.

PC-LDA revealed pregnancy outcome can be separated by DNA methylation (Figure 7-3A). This analysis suggests that there are selected methylated sites that are responsible for distinct clustering. As described in the methods, we applied a machine learning method to identify sites which are able to distinguish defined pregnancy complications and build predictive models of future pregnancy complications. Samples from the uncomplicated pregnancy group comprised the reference group and binary classification models were built for each pregnancy complication. Hierarchical clustering (Figure 7-3B) indicates that GDM, PE and PTB are distinctively different in comparison to the uncomplicated group. It also suggests that the SGA group is the most similar to the uncomplicated group. This was most likely due to SGA being defined as babies born <10th customised centile. Some of these babies may have been growth restricted and others constitutionally small but normal. The GDM, PE and PTB may be distinctively

different from the SGA and uncomplicated group due to unknown factors influencing their methylation profiles. Potential factors could have included maternal characteristics such as smoking which is known to impact fetal growth [46-48]. These pregnancy complications are influenced by a combination of genetic and environmental factors. These factors can also impact on methylation profiles which may potentially distinguish between different pregnancy complications.

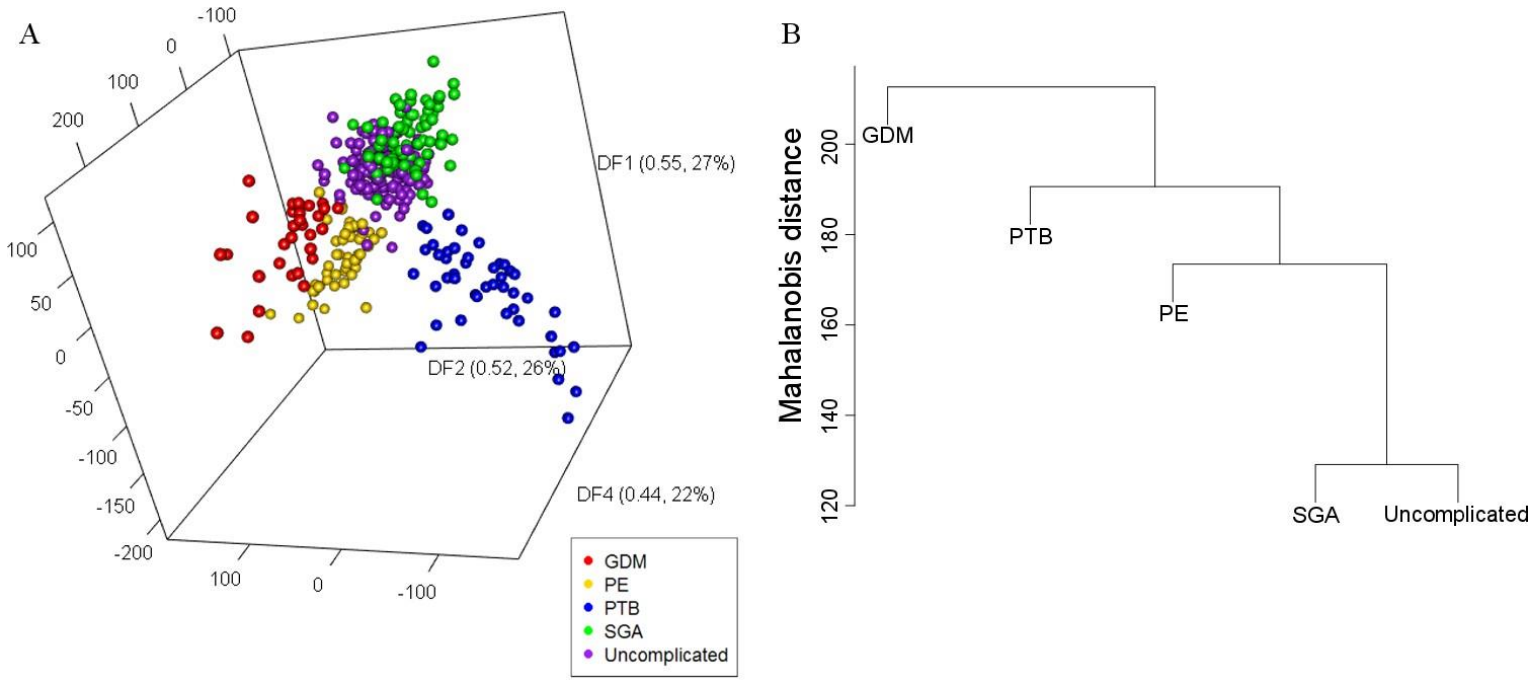


Figure 7-3: Separation of pregnancy outcomes by DNA methylation

A. PC-LDA plot of the three discriminant functions highlighting differences in DNA methylation patterns between pregnancy outcomes. **B.** Hierarchical clustering based on PC-LDA of pregnancy outcome. Small for gestational age (SGA) and the uncomplicated group are similar based on DNA methylation, whereas gestational diabetes mellitus (GDM), preeclampsia (PE) and preterm birth (PTB) are distinctively different.

Our 10-fold cross validation on the testing data set returned an excellent AUC ≥ 0.9 for our GDM and PTB in ROC analyses (Figure 7-4, Table 7-2). A good AUC for our PE and SGA models (Figure 7-4, Table 7-2) and a moderate sensitivity and high specificity rates across all of our models (Table 7-2) were achieved. Our GDM model required a total of 20 sites, whereas the PE, PTB required 21 sites each and the SGA model required 24 sites (Table 7-2, Figure 7-5). Although, not marginally different in the total number of sites in each model, the more separated group, GDM, did require the least number of sites. Potentially, the more distinct groups from the reference or uncomplicated group required fewer sites for accurate prediction. Two chosen MspI sites (chr11:43579222 and chr18:76001386) were common to both the GDM and SGA models (Figure 7-5). These sites were 27bp and 6732bp upstream, respectively, from the start site of two long intergenic non-coding RNAs (lincRNA) (Supplementary Table 7-3). These markers may have been chosen by the model as they are markers of uncomplicated term pregnancies. This would explain why there is an overlap of markers between the two models despite the two pregnancy complications clustering separately in hierarchical clustering (Figure 7-3). However, the focus of this study was to identify biomarkers and therefore any association with particular genes is outside the scope of this study. The models were also built independently of clinical characteristics such as smoking and BMI. This was to determine if such biomarkers can be found that are strong enough to predict the pregnancy outcome regardless of clinical characteristics. Furthermore, it is difficult to determine or incorporate clinical characteristics into the pregnancy outcome models since there some of the groups had a low percentage of these characteristics (Table 7-1). For example, 9.6% and 5.6% of GDM and PE samples were smokers, respectively. Due to a low sample size of smokers and statistical power in these groups, characteristics such as smoking were unable to be incorporated into the models.

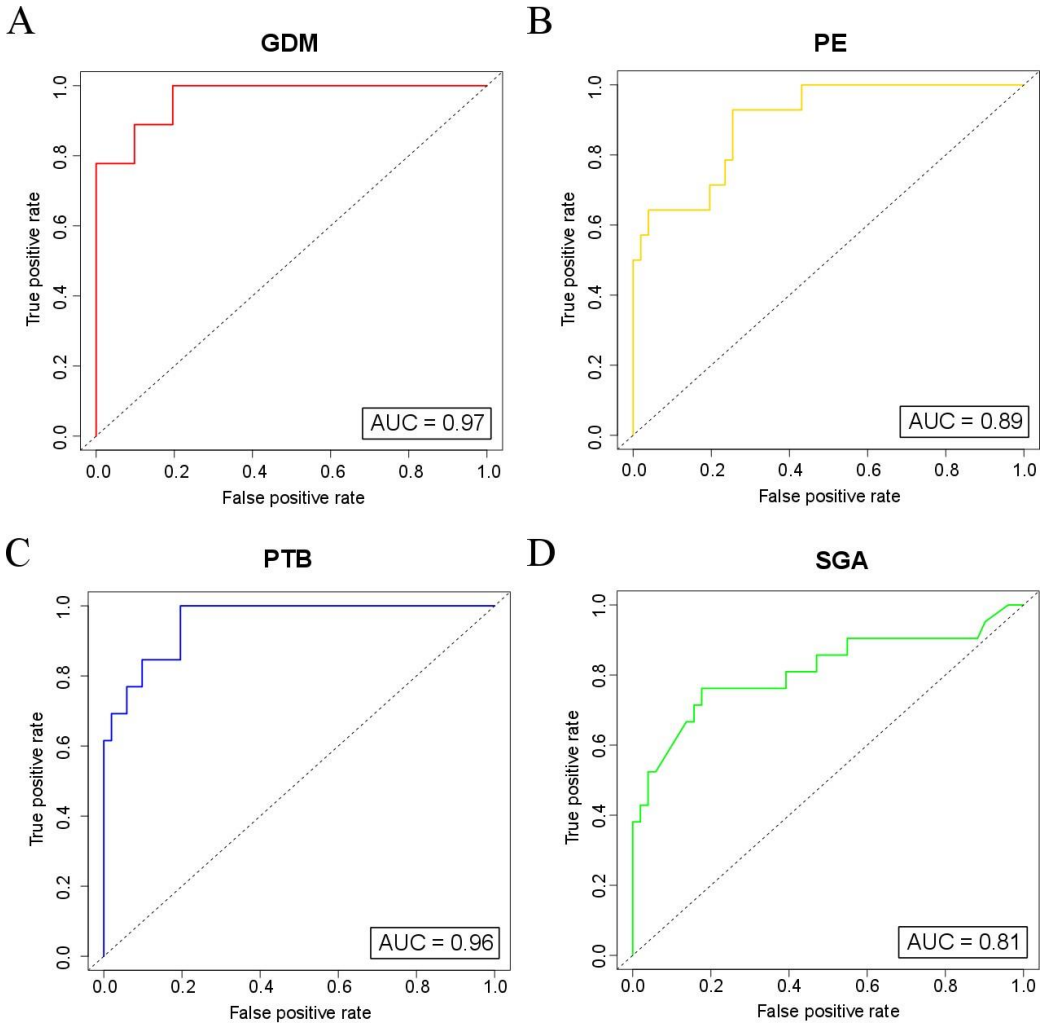


Figure 7-4: Receiver operating characteristic curves showing the false positive rate (100-Specificity) against the true positive rate (Sensitivity).

A. Gestational Diabetes Mellitus (GDM) **B.** Preeclampsia (PE) **C.** Preterm Birth (PTB) **D.** Small for Gestational Age (SGA).

Table 7-2: Accuracy measures and number of MspI cut sites required for each pregnancy complication model.

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) are represented as mean and standard deviation.

	GDM		**PE**		**PTB**		**SGA**	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Sensitivity	0.54 ± 0.13	0.49 ± 0.15	0.66 ± 0.05	0.66 ± 0.11	0.73 ± 0.06	0.61 ± 0.12	0.81 ± 0.05	0.55 ± 0.07
Specificity	0.98 ± 0.01	0.94 ± 0.05	0.95 ± 0.01	0.92 ± 0.03	0.96 ± 0.02	0.93 ± 0.05	0.95 ± 0.02	0.86 ± 0.04
PPV	0.84 ± 0.07	0.68 ± 0.23	0.80 ± 0.04	0.70 ± 0.10	0.83 ± 0.08	0.72 ± 0.17	0.88 ± 0.03	0.62 ± 0.08
NPV	0.92 ± 0.02	0.91 ± 0.02	0.91 ± 0.01	0.91 ± 0.03	0.93 ± 0.02	0.90 ± 0.03	0.92 ± 0.02	0.82 ± 0.02
AUC	0.98	0.97	0.94	0.89	0.95	0.96	0.92	0.81
No. of MspI cut sites	20		21		21		24	

AUC: area under the curve, NPV: negative predictive value, PPV: positive predictive value.

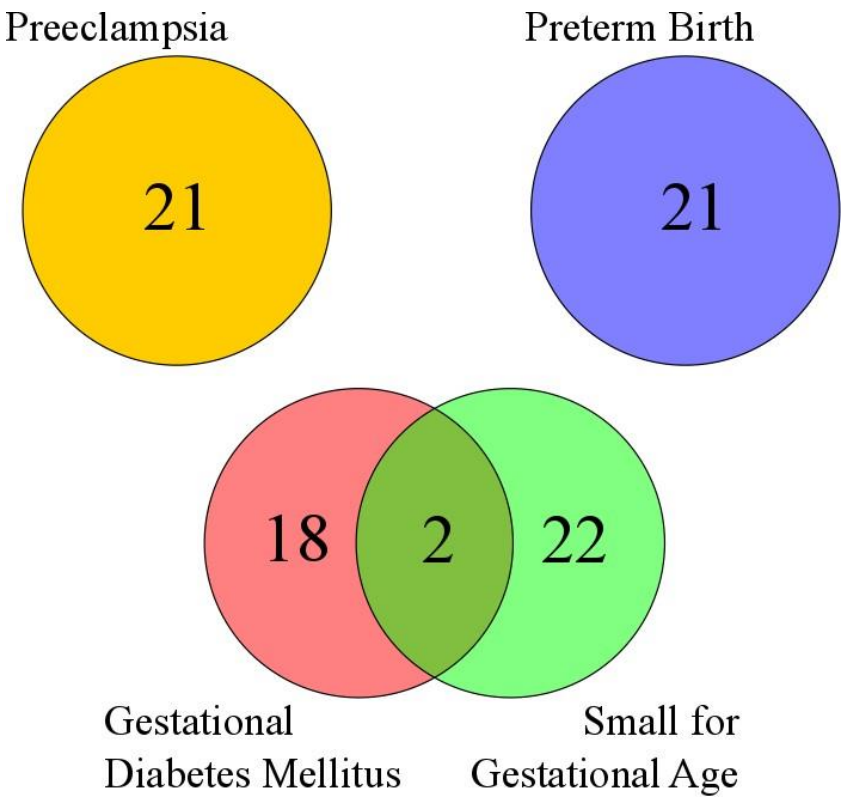


Figure 7-5: Euler diagram showing the separation and overlap of MspI sites in each pregnancy complication model.

7.3.4 What is known about the pregnancy outcome biomarkers?

In total, 84 MspI sites across all autosomal chromosomes were identified as biomarkers for the four pregnancy complications studied (Figure 7-6, Supplementary Table 7-3). 26 of the 84 sites were annotated to be within genomic locations of unknown biological function. The remaining sites were within (< 80kbp) known genomic features such as genes, exons, introns, 5´ untranslated regions (5´ UTR) and 3´ untranslated regions (3´ UTR). However, these sites can be a large distance from these genomic features (Supplementary Table 7-3). For example, one site was 79,030bp from the start of an exon (Supplementary Table 7-3). A GO enrichment for the sites to the closest genes returned no significant terms for any of the models. However, the aim of this study was to identify biomarkers and since there is no associated gene expression data we are unable to determine their effect on gene expression.

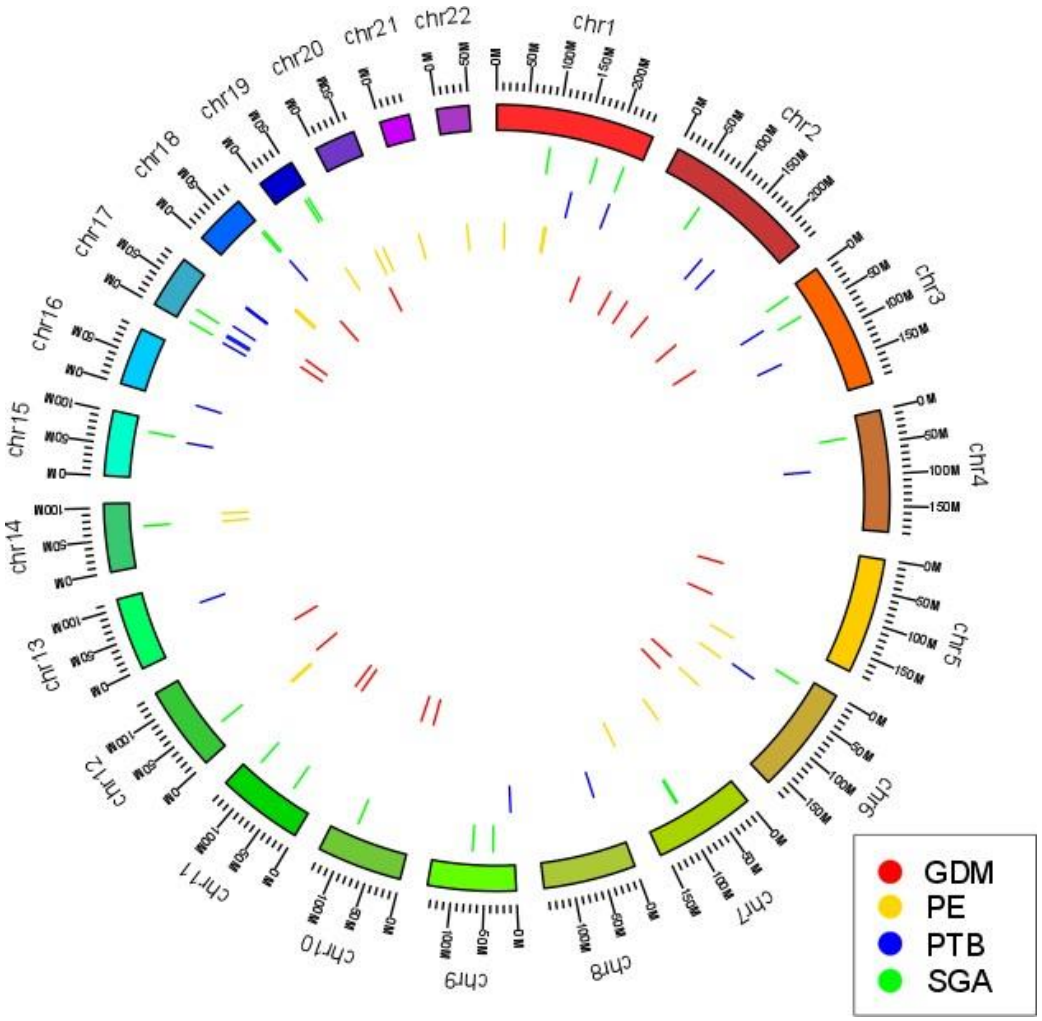


Figure 7-6: A circos plot showing the positions of the MspI cut sites identified as biomarkers for each pregnancy complication model.

7.3.5 Intrauterine growth restriction (IUGR)

In the binary classifier model for SGA we achieved AUC value of 0.81. Whereas the other models achieved much higher AUC values (> 0.89). The SGA model may potentially have not achieved as good an AUC value as the other models since some participants may have been classified as SGA, but are actually well nourished and are just genetically small. Therefore, some of the SGA participants may be genetically small but are not growth restricted. We then constructed another model using a more stringent cut off for SGA, using the participants who either had an uncomplicated pregnancy or had an infant that was < 5th centile, which was classified as IUGR. This model consisted of 37 IUGR participants (Supplementary Figure 7-2A) and achieved an improved AUC value of 0.85 (Supplementary Figure 7-2B, Supplementary Table 7-4). This also resulted in the IUGR group being distinctively different to the uncomplicated group. Interestingly, the IUGR and SGA models both contained 3 of the same chosen MspI cut sites (Supplementary Table 7-5). This suggests there may be similarities in methylation profiles in SGA and IUGR. However, IUGR is a subset of SGA and the overlapping sites may potentially be statistically stronger in distinguishing the separation between comparisons. As described above the classification for SGA and IUGR differs by birth weight centile threshold. Therefore, some of the individuals that were classified in the SGA group were also IUGR individuals. The IUGR model may have achieved better performance compared to the SGA model since these fetuses may have been under nourished creating a distinct methylation profile.

7.3.6 Comparing the DNA methylation models to existing pregnancy complication models

There have been previous attempts at building predictive models for pregnancy complications. Our sensitivity, specificity and AUC were either comparable or better than those reported in other studies for GDM [49] (AUC = 0.91), PE [50-54] (AUC = 0.89, 0.66, 76, 76, 0.97, respectively), PTB [55] (0.75) and SGA [56] (0.77) prediction. However, these studies did not use DNA methylation but focused on using serum concentrations of proteins. Serum levels of proteins are known to have high variability

[57, 58], which may have implications for biomarker design. Unlike protein levels, epigenetic marks including DNA methylation are relatively stable [59]. This makes methylated sites excellent candidates for biomarker prediction. Future studies should focus on incorporating DNA methylation into models for disease prediction. Due to epigenetic stability and low variability they can be used for more accurate prediction. Moreover, the declining cost of high throughput sequencing would increase the practicality of incorporating DNA methylation data into other disease prediction models.

7.3.7 Biomarkers of smoking

Cigarette smoking during pregnancy increases the risk of adverse outcomes for the developing fetus [60]. Many women who smoke quit once they find out they are pregnant. However, some continue to smoke throughout pregnancy. At 15 weeks' gestation we were able to identify distinct MspI sites in women who had quit and those who continued smoking compared to those who did not smoke. We built models on the basis of self-reported information about maternal smoking status. Women who were non-smokers were used as a reference group (n = 220) to build predictive models for women who had quit smoking (n = 59) or continued to smoke (n = 103). We built these models regardless of pregnancy outcome. Women who were either non-smokers, had quit smoking or continued smoking at 15 weeks' gestation were able to be distinguished by DNA methylation profiles (Supplementary Figure 7-3A). We achieved highly accurate models for both quit smoking and continued smoking (Supplementary Table 7-4, Supplementary Figure 7-3B). DNA methylation has been previously used as a biomarker of smoking [21] and our study, reinforces the notion that smoking can influence DNA methylation patterns. It also illustrates that despite quitting smoking during early pregnancy there can be an everlasting epigenetic effect in maternal blood and presumably other tissues that may impact pregnancy outcome and long term health. However, we have previously reported that women who quit smoking before 15 weeks' had similar outcomes with respect to SGA and PTB to those women who did not smoke during pregnancy [11].

7.3.8 Discrepancies in DNA methylation of women from different BMI categories

Obesity during pregnancy is associated with GDM and hypertension [61]. Furthermore, differential methylation is associated with differences in BMI [62, 63]. In this study, we also determined if models could be built using DNA methylation to predict BMI in pregnant women. These models were built using a normal BMI range (20-25) as a reference group and were trained to predict underweight (BMI <20), overweight (BMI 25-29.9) and obese (BMI ≥30) individuals. A PC-LDA revealed that these different BMI categories can be separated by DNA methylation (Supplementary Figure 7-4A). We achieved a high separation and accurate prediction (Supplementary Figure 7-4B, Supplementary Table 7-4) for underweight individuals (sensitivity = 0.96 ± 0.03, specificity = 0.58 ± 0.12, AUC = 0.95). We also achieved moderately accurate prediction for both overweight (sensitivity = 0.87 ± 0.05, specificity = 0.77 ± 0.07, AUC = 0.87) and obese individuals (sensitivity = 0.82 ± 0.08, specificity = 0.69 ± 0.06, AUC = 0.85). The BMI, along with the smoking and pregnancy outcome predictors are among the high performing models that were generated within this study. This analysis suggests that underweight individuals have more distinct methylation patterns than overweight and obese individuals compared to a normal BMI. Potentially, being underweight may also mean that these individuals are micronutrient deficient compared to the overweight and obese individuals. Therefore, the methylation profiles may also be capturing other factors such as micronutrient status in underweight individuals which may explain the distinct methylation profile. In addition, although BMI can be easily measured, this analysis shows that there are distinct methylation profiles. This suggests that BMI can exert profound effects on DNA methylation patterns. Therefore, this may uncover molecular mechanisms associated with BMI, which could assist in explaining the effects of BMI.

7.3.9 Circulating factors in maternal blood associated with DNA methylation

Folate is part of the one carbon metabolism pathway which is essential for the production of methyl groups for DNA methylation [64]. We set out to determine if DNA methylation

can be used to determine the folate status of an individual. We initially attempted, using an elastic net regression method, to predict serum folate however this had poor performance, potentially due to a low spread of the data (±10nM). We therefore found it advantageous to categorise the data. Serum folate status was categorised into two groups using the median (34.4nM) as a threshold. We were able to identify 21 sites (Supplementary Table 7-4, 7-8) that can moderately predict high serum folate from low levels (Supplementary Figure 7-5A, 7-5B). One of the disadvantages of categorising the data is that samples close to the median can be incorrectly predicted. There is evidence of this in the PC-LDA plot (Supplementary Figure 7-5A) where samples of high and low serum folate are overlapping. We also set out to determine if level of methylation between the two serum folate groups differed. We determined the mean level of DNA methylation using all the MspI sites (Supplementary Figure 7-6A) or the selected sites within the model (Supplementary Figure 7-6B). We found no difference between the groups when using all the MspI sites, however, we found a significantly higher level of DNA methylation in the low folate serum group when using the selected sites within the model. This may provide the rationale for why these sites were distinguished between the two groups.

Despite folate being well known for its role in the production of methyl groups a lower level of methylation was found in the higher folate status group. This is interesting as it may seem contradictory that having a high folate status is associated with low methylation levels. However, it is important not to forget that this assay only measures methylation at MspI sites and therefore the true extent of methylation has not been accurately quantified. Other sequencing methods such as whole genome bisulphite sequencing would capture the true extent of methylation. Since the true extent of the level of methylation in these samples has not be accurately quantified it is difficult to draw conclusion regarding the level of methylation and folate status.

7.3.10 Characteristics that are not predicted by DNA methylation

Previous studies have used DNA methylation to determine an individual's age [15, 65]. We too set out to determine if maternal age can be determined using our DNA methylation data set using an elastic net penalised regression method as previously described [15, 39] (see methods). Unfortunately, we were unable to build an accurate predictive model of maternal age. The model failed to predict the maternal age in the testing data set (r =0.0), which may have been in part due to the narrow age range in these pregnant women (Table 7-1). This was surprising since models using DNA methylation have been previously developed to determine age [15, 65]. However, these studies have used microarray technology to build predictive models. Although, this study captured 2,610,160 methylated sites, the sites required to predict age in other studies may not be present in our methylation data set. We therefore downloaded the positions of all sites in the Illumina Infinium HumanMethylation450 BeadChip from Illumina which have been used in other studies to predict age [15, 65]. We compared these sites to those generated in our study using msGBS. We found that none of the sties used in the microarray technology were present in our DNA methylation data set. This could potentially be a reason why we were unable to generate a model that can predict maternal age. In addition, this also makes it difficult to compare epigenome wide association studies (EWAS) since most studies use Illumina array technology. This suggests that determining age is site specific and that the sites captured in the current study may only be useful for prediction of certain characteristics.

Although this study has been successful in identifying sites that can be used to predict pregnancy complications, smoking status and BMI categories, we were unable to build predictive models for many other clinical outcomes and characteristics. For example, we were unable to determine maternal characteristics such as blood pressure (diastolic and systolic), height, micronutrient status and supplement use. These factors may not have been able to be predicted due to low sample sizes for these characteristics. We were also unable to determine other fetal characteristics such as sex, birth length, birth weight and final gestation. These fetal characteristics may potentially not have any influence on maternal blood factors such as DNA methylation. Factors within the cord blood are likely to be better indicators of fetal characteristics than maternal blood. However, it is estimated that there are approximately 28 million CpG sites in the human genome [66].

This study was able to capture 2.7 million methylated sites and therefore there may potentially be other sites more suitable for prediction of these characteristics.

7.3.11 Strengths and limitations

Conducting msGBS has its strengths and limitations. In comparison to other DNA methylation assays such as microarrays, it provides an unbiased approach to selection of methylated cytosine sites. Also compared to DNA methylation sequencing based techniques it is cost effective [67] and allows up 96 samples to be run at once. Although in this study we sequenced each sample three times, therefore a total of 32 samples could be run at once. A limitation of msGBS is that it is restricted to selected recognition sites. In this experiment, the restriction enzyme MspI was used, which has a recognition sequence of CCGG. DNA methylation was analysed at CCGG sites and other sites were not analysed. This potentially would have left other methylated sites out of the analysis that would be more suitable predictors. However, msGBS can be performed again, but with other methylation sensitive restriction enzymes, thereby covering more sites of the genome.

One of the challenges in building predictive models using DNA methylation data is working with the high complexity of potentially millions of variables. Due to the high number of variables of DNA methylation data sets, computational time can become overwhelming [68]. Since DNA methylation data sets suffer from high dimensionality it can be advantageous to perform feature selection [68]. This is one of the advantages of our method as by performing feature selection prior to building the models using the RFE method resulted in reduced computational time. This allowed us to build other predictive models in other clinical characteristics such as smoking and BMI. However, performing feature selection does have its pitfalls. A limitation is that it can result in overfitting which is a common problem where a model is trained on few data points. This results in poor performance of the testing data set. However, as shown in this study our predictive models perform very well on the testing data set. To ensure the robustness of our models it is also essential to validate our models on an independent data set to determine if any

potential overfitting of the data has occurred. Future studies are required to test our DNA methylation prognostic models to determine their true performance.

In this study, we have analysed the methylation profiles in the maternal blood from five distinct pregnancy outcomes (GDM, PE, PTB, SGA and Uncomplicated pregnancies). Although this study has shown that DNA methylation can be used to identify women having different pregnancy outcomes, it still needs to be developed more before it can be used clinically. As mentioned this study analysed five distinct pregnancy outcomes, however there are more different types of pregnancy outcomes. For example, PE can be divided into early onset (< 34 weeks' gestation) or late onset (≥ 34 weeks' gestation). This study only analysed pregnancies from late onset PE. It has been shown that there are differences in placental DNA methylation profiles between early and late onset PE pregnancies [69-71]. Therefore, DNA methylation profiles may be distinctively different in the maternal blood. Ideally, for a clinical application, models would have be constructed against a broader range of pregnancy outcomes. However, the aim of this study was to determine if DNA methylation can be used to identify different pregnancy outcomes.

Currently, DNA methylation as a biomarker has mostly dominated in cancer related applications [59]. However, this study shows that DNA methylation can be used in other fields of research such as a reproductive setting. Here in this study, we have developed prognostic models using the level of methylation at selected sites. These DNA methylation prediction models could potentially have a role in a clinical setting. They could identify high risk individuals such that early prevention can be administered thereby reducing or eliminating any risk of a pregnancy complication. We were also able to develop accurate prediction models for other clinical characteristics such as BMI and smoking. DNA methylation can be a valuable tool to determine such characteristics where clinical data may be unavailable.

7.4 Conclusion

In this study, we have identified DNA methylation signatures of pregnancy complications. Using DNA methylation we were able to build predictive models of pregnancy outcomes and other clinical characteristics of pregnant women at 15 weeks' gestation, a long time before the diseases become symptomatic. Future work is required to replicate our results and test our models in an independent cohort. This analysis highlights the possibility of using DNA methylation as a prognostic tool for pregnancy complications.

7.5 Supporting Information

For Supplementary Table 7-1, 7-2, 7-3, 7-4, 7-5, 7-6, 7-7 and 7-8 please refer to the electronic supporting information.

Supplementary Table 7-1: Barcode adapter sequences used for msGBS.

Supplementary Table 7-2: Total number of methylated MspI sites captured on each chromosome.

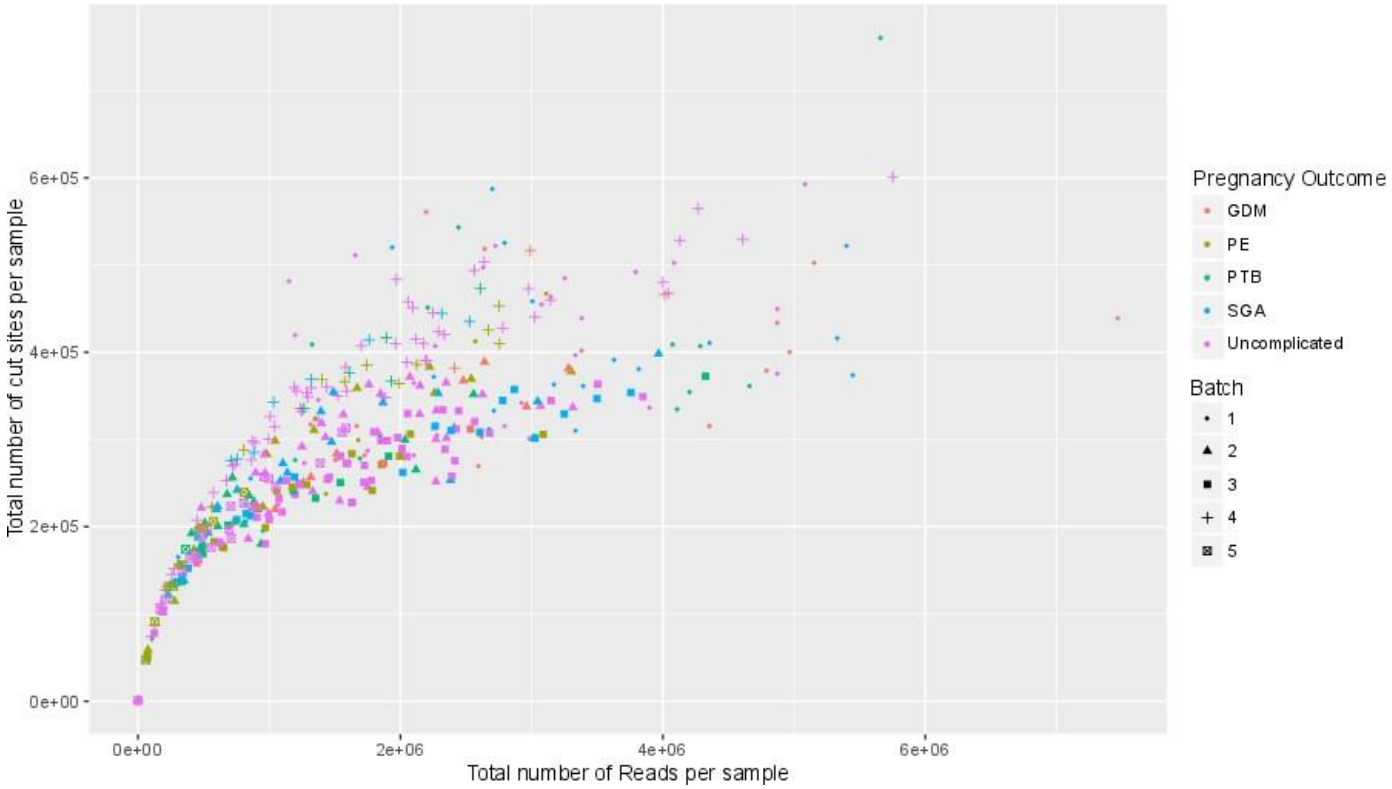
Supplementary Table 7-3: Locations of pregnancy outcome biomarkers and the closest genomic feature.

Supplementary Table 7-4: Accuracy measures and number of MspI cut sites required predictive models.
Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) are represented as mean and standard deviation.
Supplementary Table 7-5: Locations of biomarkers for the intrauterine growth restriction (IUGR) model.

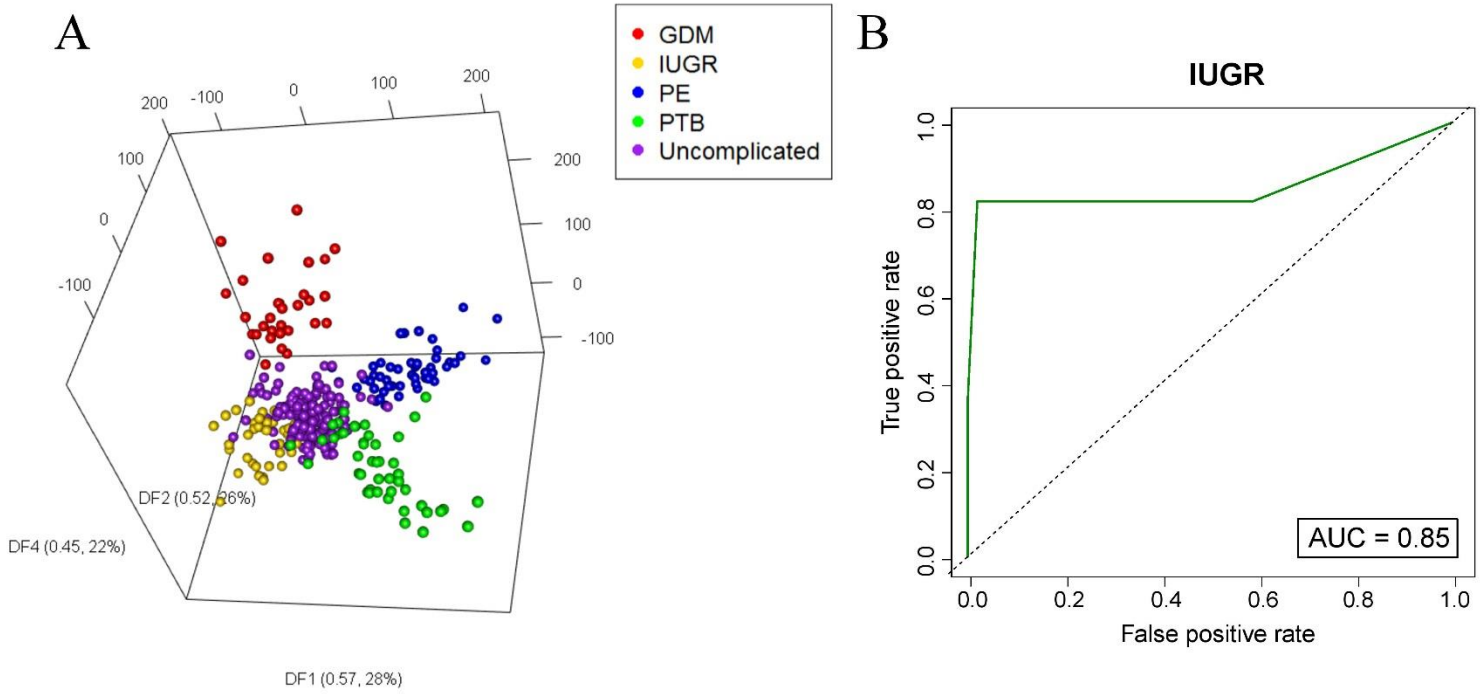
Supplementary Table 7-6: Locations of biomarkers for women who had quit smoking or continued smoking at 15 weeks' gestation.

Supplementary Table 7-7: Locations of biomarkers for women from different BMI categories.

Supplementary Table 7-8: Locations of biomarkers for women with high serum folate.

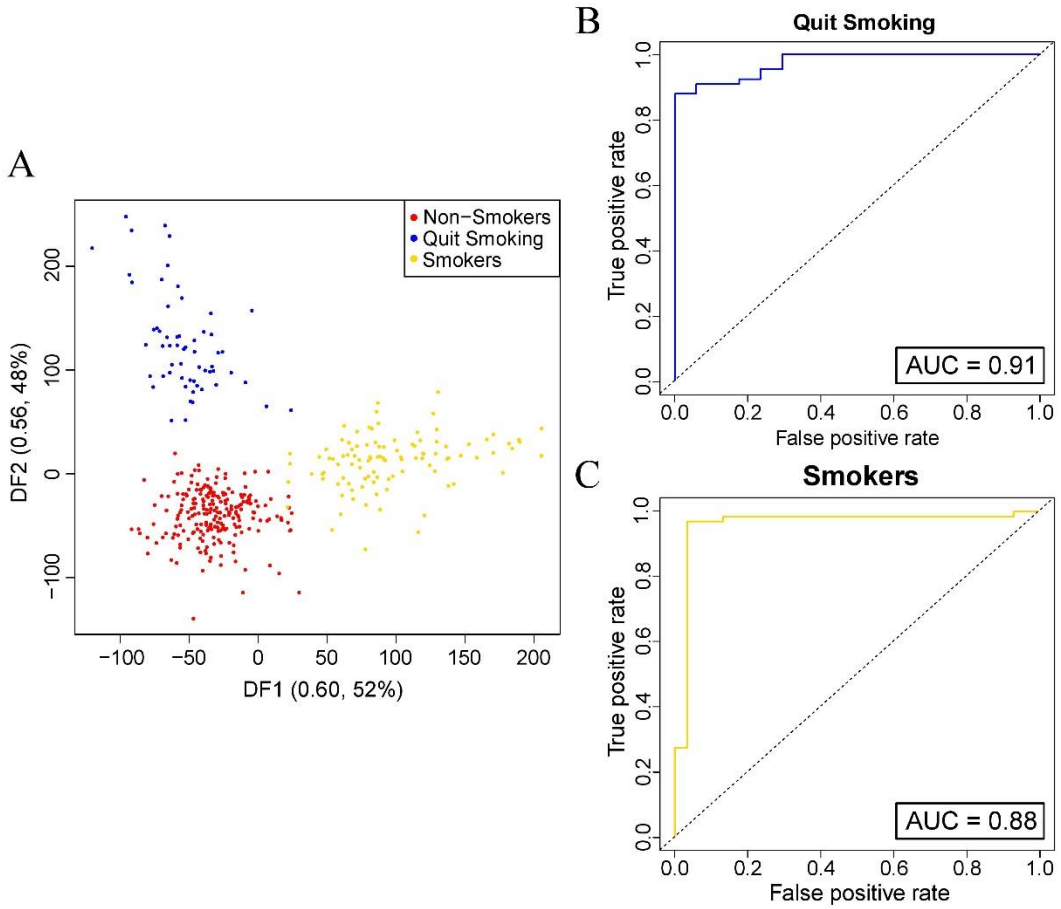


Supplementary Figure 7-1: Total number of reads per sample compared to the total number of cut sites generated per sample. Each colour and shape represents a pregnancy outcome and the plate that the sample was sequenced on, respectively.



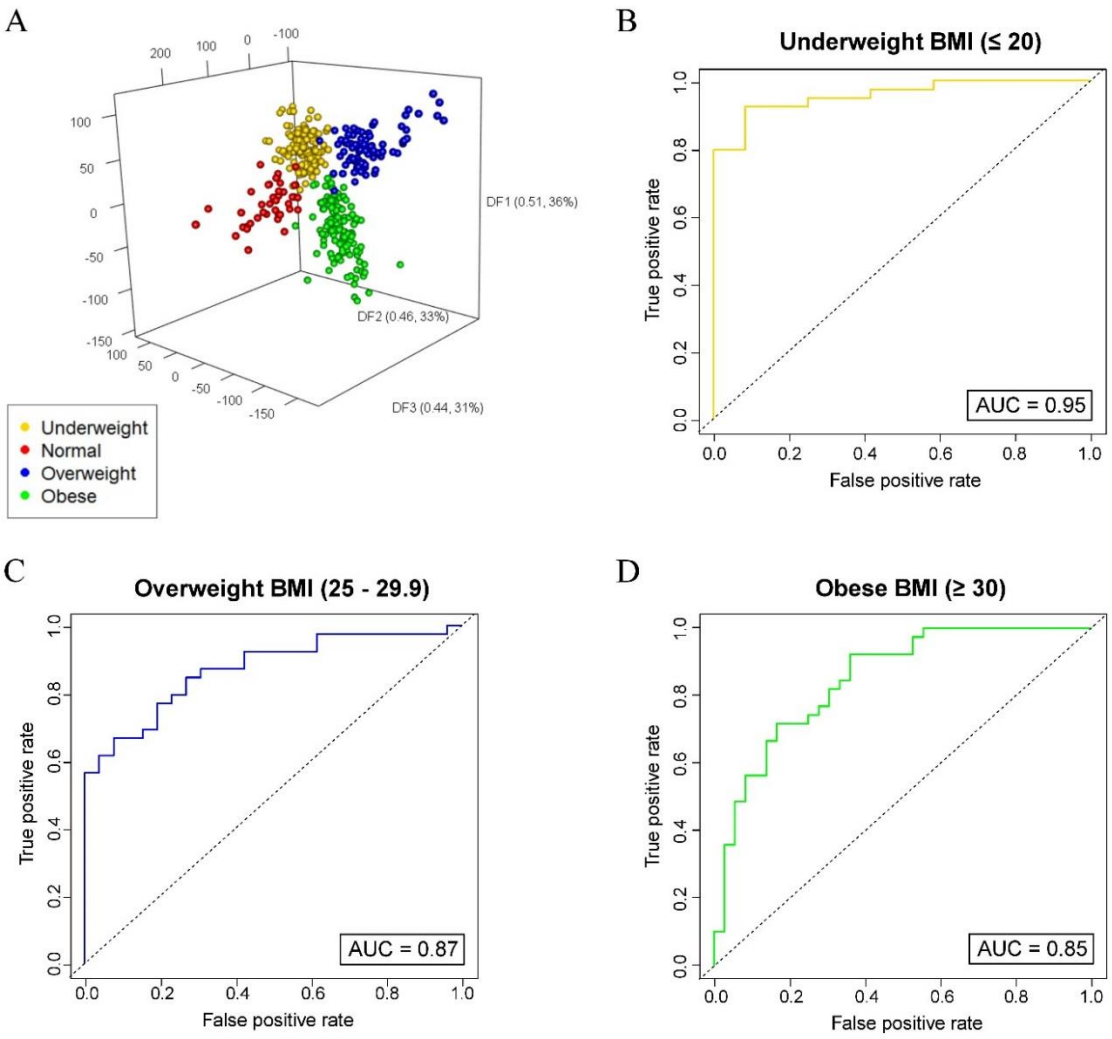
Supplementary Figure 7-2: Identifying biomarkers of intrauterine growth restriction (IUGR).

A. Principal component linear discriminant analysis (PC-LDA) of pregnancy outcome. **B.** ROC curve of intrauterine growth restriction (IUGR) model.



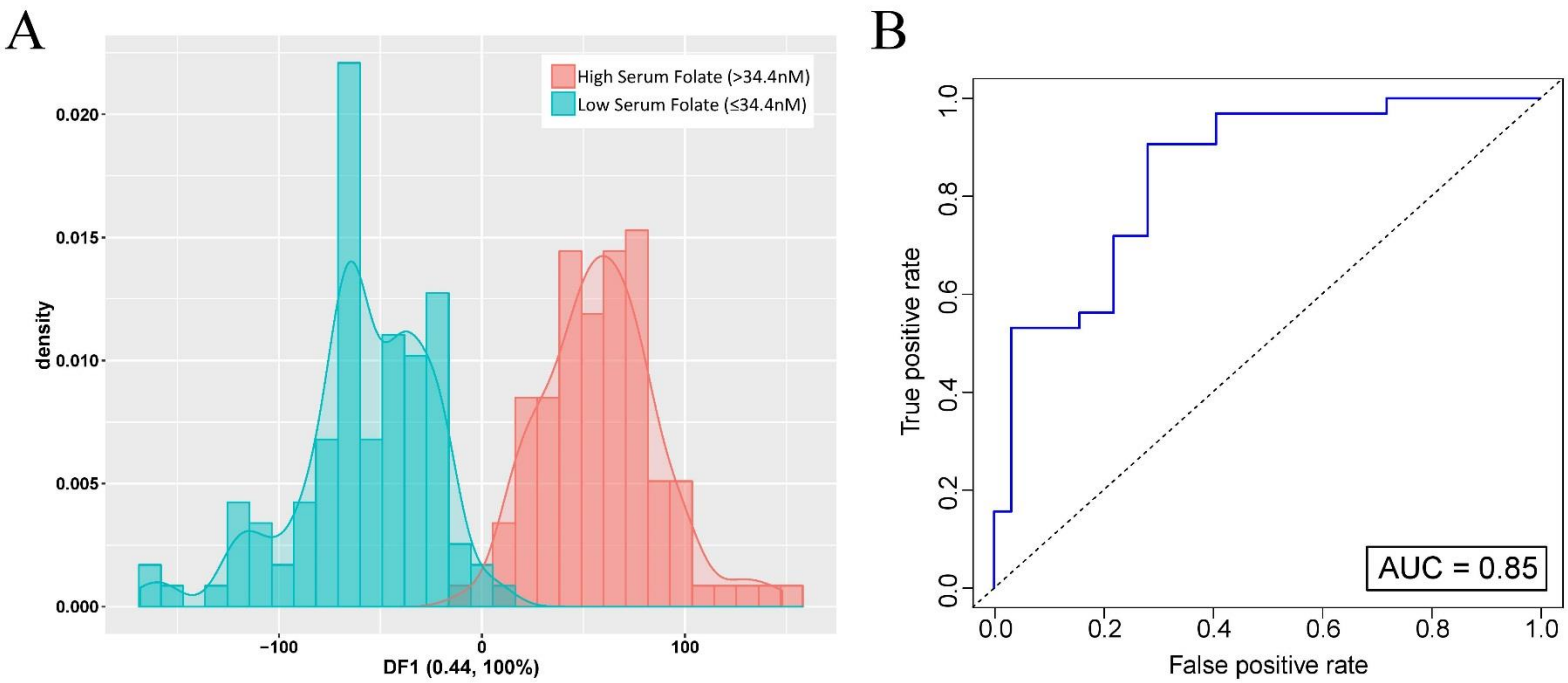
Supplementary Figure 7-3: Identifying biomarkers for women who had quit smoking and continued smoking at 15 weeks' gestation.

A. Principal component linear discriminant analysis (PC-LDA) of non-smokers, women who had quit smoking and women who continued smoking. ROC curves of the models to identify women **B.** who had quit smoking and **C.** women who continued smoking.



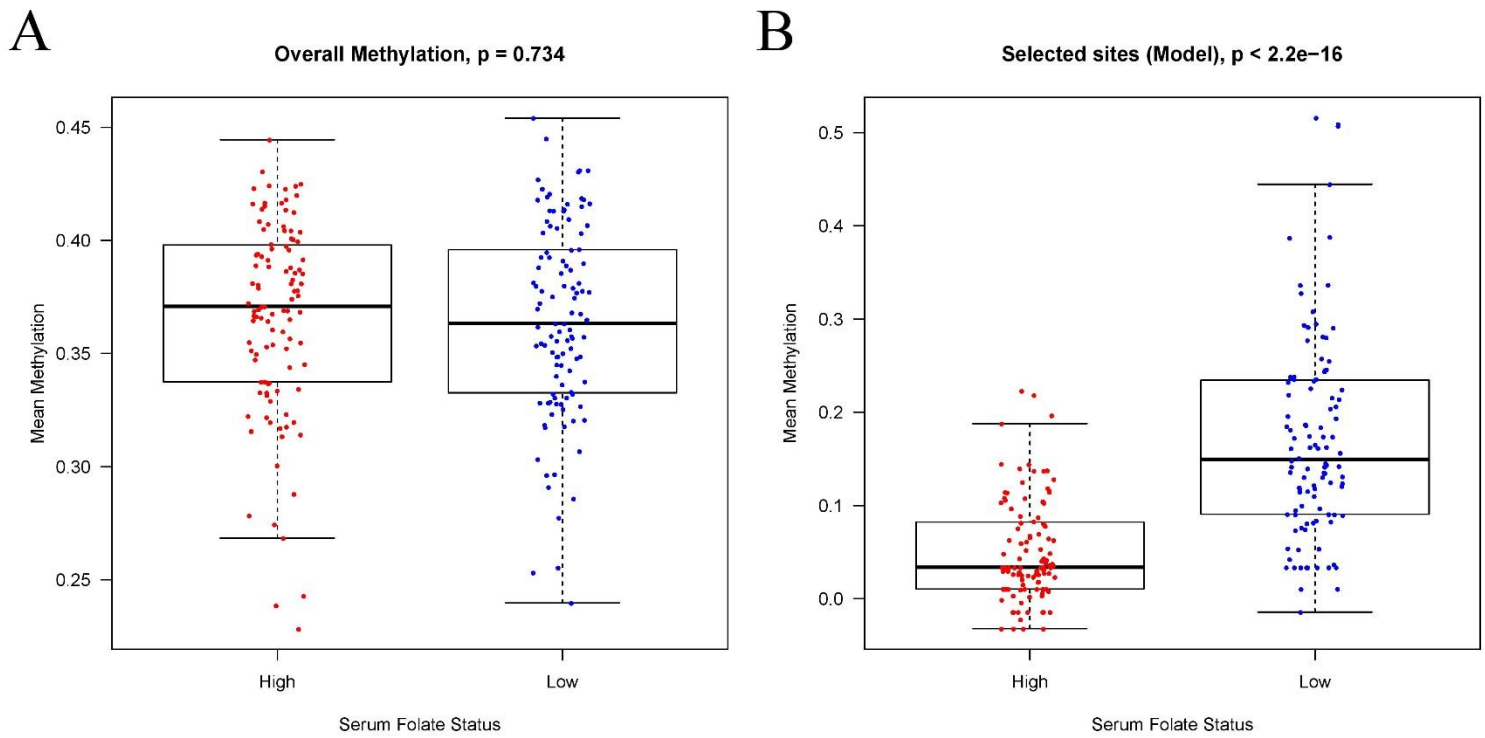
Supplementary Figure 7-4: Identifying biomarkers of different body mass index (BMI) categories.

A. Principal component linear discriminant analysis (PC-LDA) of individuals of different BMI categories. ROC curves of models to identify **B.** Underweight, **C.** Overweight and **D.** Obese.



Supplementary Figure 7-5: Biomarkers of serum folate status.

A. Principal component linear discriminant analysis (PC-LDA) of individuals with high or low serum folate levels. **B.** ROC curve of serum folate model.



Supplementary Figure 7-6: Average level of DNA methylation between the low and high serum folate groups.

A. all MspI sites or **B.** selected sites within the folate model.

References

1. Alfadhli EM: **Gestational diabetes mellitus.** *Saudi Med J* 2015, **36:**399-406.
2. Osungbade KO, Ige OK: **Public Health Perspectives of Preeclampsia in Developing Countries: Implication for Health System Strengthening.** *Journal of Pregnancy* 2011, **2011:**481095.
3. Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller AB, Narwal R, Adler A, Vera Garcia C, Rohde S, Say L, Lawn JE: **National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications.** *Lancet* 2012, **379:**2162-2172.
4. Romo A, Carceller R, Tobajas J: **Intrauterine growth retardation (IUGR): epidemiology and etiology.** *Pediatr Endocrinol Rev* 2009, **6 Suppl 3:**332-336.
5. Raty R, Koskinen P, Alanen A, Irjala K, Matinlauri I, Ekblad U: **Prediction of pre-eclampsia with maternal mid-trimester total renin, inhibin A, AFP and free beta-hCG levels.** *Prenat Diagn* 1999, **19:**122-127.
6. Tul N, Pusenjak S, Osredkar J, Spencer K, Novak-Antolic Z: **Predicting complications of pregnancy with first-trimester maternal serum free-betahCG, PAPP-A and inhibin-A.** *Prenat Diagn* 2003, **23:**990-996.
7. Brameld KJ, Dickinson JE, O'Leary P, Bower C, Goldblatt J, Hewitt B, Murch A, Stock R: **First trimester predictors of adverse pregnancy outcomes.** *Aust N Z J Obstet Gynaecol* 2008, **48:**529-535.
8. Poon LC, Nicolaides KH: **Early Prediction of Preeclampsia.** *Obstet Gynecol Int* 2014, **2014:**297397.
9. Kenny LC, Black MA, Poston L, Taylor R, Myers JE, Baker PN, McCowan LM, Simpson NA, Dekker GA, Roberts CT, et al: **Early pregnancy prediction of preeclampsia in nulliparous women, combining clinical risk and biomarkers: the Screening for Pregnancy Endpoints (SCOPE) international cohort study.** *Hypertension* 2014, **64:**644-652.
10. Dekker GA, Lee SY, North RA, McCowan LM, Simpson NAB, Roberts CT: **Risk Factors for Preterm Birth in an International Prospective Cohort of Nulliparous Women.** *PLoS One* 2012, **7:**e39154.
11. McCowan LME, Thompson JMD, Taylor RS, Baker PN, North RA, Poston L, Roberts CT, Simpson NAB, Walker JJ, Myers J, et al: **Prediction of Small for Gestational Age Infants in Healthy Nulliparous Women Using Clinical and Ultrasound Risk Factors Combined with Early Pregnancy Biomarkers.** *PLoS One* 2017, **12:**e0169311.
12. Thornburg KL, Shannon J, Thuillier P, Turker MS: **In utero life and epigenetic predisposition for disease.** *Adv Genet* 2010, **71:**57-78.
13. Bell JT, Spector TD: **DNA methylation studies using twins: what are they telling us?** *Genome Biol* 2012, **13:**172.
14. Mathers JC, Strathdee G, Relton CL: **Induction of epigenetic alterations by dietary and other environmental factors.** *Adv Genet* 2010, **71:**3-39.

15. Horvath S: **DNA methylation age of human tissues and cell types.** *Genome Biol* 2013, **14:**R115.
16. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S: **Genome-wide methylation profiles reveal quantitative views of human aging rates.** *Mol Cell* 2013, **49**.
17. Perna L, Zhang Y, Mons U, Holleczek B, Saum K-U, Brenner H: **Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort.** *Clin Epigenetics* 2016, **8:**64.
18. Ulirsch J, Fan C, Knafl G, Wu MJ, Coleman B, Perou CM, Swift-Scanlan T: **Vimentin DNA methylation predicts survival in breast cancer.** *Breast Cancer Res Treat* 2013, **137:**383-396.
19. Warton K, Samimi G: **Methylation of cell-free circulating DNA in the diagnosis of cancer.** *Frontiers in Molecular Biosciences* 2015, **2:**13.
20. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE: **DNA methylation age of blood predicts all-cause mortality in later life.** *Genome Biol* 2015, **16**.
21. Mikeska T, Craig JM: **DNA Methylation Biomarkers: Cancer and Beyond.** *Genes* 2014, **5:**821-864.
22. Knight AK, Craig JM, Theda C, Bækvad-Hansen M, Bybjerg-Grauholm J, Hansen CS, Hollegaard MV, Hougaard DM, Mortensen PB, Weinsheimer SM, et al: **An epigenetic clock for gestational age at birth based on blood methylation data.** *Genome Biol* 2016, **17:**206.
23. Mayne BT, Leemaqz SY, Smith AK, Breen J, Roberts CT, Bianco-Miotto T: **Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation.** *Epigenomics* 2016.
24. Chappell LC, Seed PT, Myers J, Taylor RS, Kenny LC, Dekker GA, Walker JJ, McCowan LM, North RA, Poston L: **Exploration and confirmation of factors associated with uncomplicated pregnancy in nulliparous women: prospective cohort study.** *Br J Sports Med* 2015, **49:**136.
25. Reyna-Lopez GE, Simpson J, Ruiz-Herrera J: **Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi by amplification of restriction polymorphisms.** *Mol Gen Genet* 1997, **253:**703-710.
26. Rodríguez López CM, Morán P, Lago F, Espiñeira M, Beckmann M, Consuegra S: **Detection and quantification of tissue of origin in salmon and veal products using methylation sensitive AFLPs.** *Food Chem* 2012, **131:**1493-1498.
27. Poland JA, Brown PJ, Sorrells ME, Jannink JL: **Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach.** *PLoS One* 2012, **7:**e32253.
28. Xia Z, Zou M, Zhang S, Feng B, Wang W: **AFSM sequencing approach: a simple and rapid method for genome-wide SNP and methylation site discovery and genetic mapping.** *Sci Rep* 2014, **4:**7300.
29. Herten K, Hestand MS, Vermeesch JR, Van Houdt JK: **GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments.** *BMC Bioinformatics* 2015, **16:**1-6.

30. Bushnell B: **BBMap. Available at: http://sourceforge.net/projects/bbmap/.** 2015.
31. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.
32. B M: **msgbsR: msgbsR: methylation sensitive genotyping by sequencing (MS-GBS) R functions. R package version 1.0.0.** 2017.
33. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al: **Ensembl 2016.** *Nucleic Acids Res* 2016, **44:**D710-D716.
34. Kitts A PL, Ward M, et al.: **The Database of Short Genetic Variation (dbSNP) 2013 Jun 30 [Updated 2014 Apr 3]. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK174586/.**
35. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges.** *PLoS Comput Biol* 2013, **9:**e1003118.
36. Enot DP, Lin W, Beckmann M, Parker D, Overy DP, Draper J: **Preprocessing, classification modeling and feature selection using flow injection electrospray mass spectrometry metabolite fingerprint data.** *Nat Protocols* 2008, **3:**446-470.
37. Kuhn M: **Building predictive models in R using the caret package.** *Journal of Statistical Software* 2008, **28:**1-26.
38. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R.** *Bioinformatics* 2005, **21:**3940-3941.
39. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *J Stat Softw* 2010, **33:**1-22.
40. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al: **Ensembl 2013.** *Nucleic Acids Res* 2013, **41:**D48-55.
41. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44-57.
42. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J: **g:Profiler-a web server for functional interpretation of gene lists (2016 update).** *Nucleic Acids Res* 2016, **44:**W83-89.
43. Carrel L, Willard HF: **X-inactivation profile reveals extensive variability in X-linked gene expression in females.** *Nature* 2005, **434:**400-404.
44. Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, Antonarakis SE: **DNA methylation profiles of human active and inactive X chromosomes.** *Genome Res* 2011, **21:**1592-1600.
45. Peeters SB, Cotton AM, Brown CJ: **Variable escape from X-chromosome inactivation: Identifying factors that tip the scales towards expression.** *Bioessays* 2014, **36:**746-756.
46. Wickström R: **Effects of Nicotine During Pregnancy: Human and Experimental Evidence.** *Curr Neuropharmacol* 2007, **5:**213-222.

47. Reeves S, Bernstein I: **Effects of maternal tobacco-smoke exposure on fetal growth and neonatal size.** *Expert Rev Obstet Gynecol* 2008, **3:**719-730.
48. Zaren B, Lindmark G, Bakketeig L: **Maternal smoking affects fetal growth more in the male fetus.** *Paediatr Perinat Epidemiol* 2000, **14:**118-126.
49. Rasanen JP, Snyder CK, Rao PV, Mihalache R, Heinonen S, Gravett MG, Roberts CT, Jr., Nagalla SR: **Glycosylated fibronectin as a first-trimester biomarker for prediction of gestational diabetes.** *Obstet Gynecol* 2013, **122:**586-594.
50. Bahado-Singh RO, Akolekar R, Mandal R, Dong E, Xia J, Kruger M, Wishart DS, Nicolaides K: **First-trimester metabolomic detection of late-onset preeclampsia.** *Am J Obstet Gynecol* 2013, **208:**58 e51-57.
51. Kusanovic JP, Romero R, Chaiworapongsa T, Erez O, Mittal P, Vaisbuch E, Mazaki-Tovi S, Gotsch F, Edwin SS, Gomez R, et al: **A prospective cohort study of the value of maternal plasma concentrations of angiogenic and anti-angiogenic factors in early pregnancy and midtrimester in the identification of patients destined to develop preeclampsia.** *J Matern Fetal Neonatal Med* 2009, **22:**1021-1038.
52. Myers JE, Kenny LC, McCowan LM, Chan EH, Dekker GA, Poston L, Simpson NA, North RA: **Angiogenic factors combined with clinical risk factors to predict preterm pre-eclampsia in nulliparous women: a predictive test accuracy study.** *BJOG* 2013, **120:**1215-1223.
53. Payne B, Hodgson S, Hutcheon JA, Joseph KS, Li J, Lee T, Magee LA, Qu Z, von Dadelszen P: **Performance of the fullPIERS model in predicting adverse maternal outcomes in pre-eclampsia using patient data from the PIERS (Pre-eclampsia Integrated Estimate of RiSk) cohort, collected on admission.** *BJOG* 2013, **120:**113-118.
54. Verlohren S, Galindo A, Schlembach D, Zeisler H, Herraiz I, Moertl MG, Pape J, Dudenhausen JW, Denk B, Stepan H: **An automated method for the determination of the sFlt-1/PIGF ratio in the assessment of preeclampsia.** *Am J Obstet Gynecol* 2010, **202:**161 e161-161 e111.
55. Paternoster D, Riboni F, Vitulo A, Plebani M, Dell'Avanzo M, Battagliarin G, Surico N, Nicolini U: **Phosphorylated insulin-like growth factor binding protein-1 in cervical secretions and sonographic cervical length in the prediction of spontaneous preterm delivery.** *Ultrasound Obstet Gynecol* 2009, **34:**437-440.
56. Reboul Q, Delabaere A, Luo ZC, Nuyt AM, Wu Y, Chauleur C, Fraser W, Audibert F: **Prediction of small for gestational age neonates by third trimester fetal biometry and impact of ultrasound-delivery interval.** *Ultrasound Obstet Gynecol* 2016.
57. Crosley LK, Duthie SJ, Polley AC, Bouwman FG, Heim C, Mulholland F, Horgan G, Johnson IT, Mariman EC, Elliott RM, et al: **Variation in protein levels obtained from human blood cells and biofluids for platelet, peripheral blood mononuclear cell, plasma, urine and saliva proteomics.** *Genes Nutr* 2009, **4:**95-102.
58. Liu Y, Buil A, Collins BC, Gillet LCJ, Blum LC, Cheng L-Y, Vitek O, Mouritsen J, Lachance G, Spector TD, et al: **Quantitative variability of 342 plasma proteins in a human twin population.** *Mol Syst Biol* 2015, **11:**786.

59. Levenson VV: **DNA methylation as a universal biomarker.** *Expert Rev Mol Diagn* 2010, **10:**481-488.
60. Shea AK, Steiner M: **Cigarette smoking during pregnancy.** *Nicotine Tob Res* 2008, **10:**267-278.
61. Heude B, Thiébaugeorges O, Goua V, Forhan A, Kaminski M, Foliguet B, Schweitzer M, Magnin G, Charles M-A, the EM-CCSg: **Pre-pregnancy body mass index and weight gain during pregnancy: relations with gestational diabetes and hypertension, and birth outcomes.** *Maternal and Child Health Journal* 2012, **16:**355-363.
62. Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aissi D, Wahl S, Meduri E, Morange PE, Gagnon F, Grallert H, et al: **DNA methylation and body-mass index: a genome-wide analysis.** *Lancet* 2014, **383:**1990-1998.
63. Ribel-Madsen R, Fraga MF, Jacobsen S, Bork-Jensen J, Lara E, Calvanese V, Fernandez AF, Friedrichsen M, Vind BF, Hojlund K, et al: **Genome-wide analysis of DNA methylation differences in muscle and fat from monozygotic twins discordant for type 2 diabetes.** *PLoS One* 2012, **7:**e51302.
64. Anderson OS, Sant KE, Dolinoy DC: **Nutrition and epigenetics: an interplay of dietary methyl donors, one-carbon metabolism and DNA methylation.** *J Nutr Biochem* 2012, **23:**853-859.
65. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y: **Genome-wide methylation profiles reveal quantitative views of human aging rates.** *Mol Cell* 2013, **49**.
66. Lövkvist C, Dodd IB, Sneppen K, Haerter JO: **DNA methylation in human epigenomes depends on local topology of CpG sites.** *Nucleic Acids Res* 2016, **44:**5123-5132.
67. Ziller MJ, Hansen KD, Meissner A, Aryee MJ: **Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing.** *Nat Methods* 2015, **12:**230-232, 231 p following 232.
68. Libbrecht MW, Noble WS: **Machine learning applications in genetics and genomics.** *Nat Rev Genet* 2015, **16:**321-332.
69. van den Berg CB, Chaves I, Herzog EM, Willemsen SP, van der Horst GTJ, Steegers-Theunissen RPM: **Early- and late-onset preeclampsia and the DNA methylation of circadian clock and clock-controlled genes in placental and newborn tissues.** *Chronobiol Int* 2017**:**1-12.
70. Mayne BT, Leemaqz SY, Smith AK, Breen J, Roberts CT, Bianco-Miotto T: **Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation.** *Epigenomics* 2017, **9:**279-289.
71. Hogg K, Blair JD, McFadden DE, von Dadelszen P, Robinson WP: **Early onset pre-eclampsia is associated with altered DNA methylation of cortisol-signalling and steroidogenic genes in the placenta.** *PLoS One* 2013, **8:**e62969.

8 Discussion

8.1 General Discussion

The work presented in this thesis has focused on identifying molecular markers for the identification of phenotypic traits. These molecular markers have primarily been the expression of genes and DNA methylation that were assessed by either large scale meta-analyses of genomic data or using next generation sequencing technology. This thesis has identified a significant number of sexually dimorphic candidate genes and cytosine-phosphate-guanine (CpG) sites that can predict the gestational age of a placenta. These candidate sex-biased genes and gestational age predictor sites were identified by taking full advantage of publicly available genomic data, a significant resource for health-specific data analyses. Additionally, this thesis has also used RNA sequencing technology to identify novel large intergenic non-coding RNAs (lincRNAs) in placental tissue. Next generation sequencing was also used to identify methylated cytosine sites in DNA from maternal peripheral leukocytes sampled early in gestation that predict later pregnancy complications. Overall, this thesis demonstrates the potential for using large genomic data sets and next generation sequencing applications to identify gene targets and biomarkers related to placental function and pregnancy outcome.

8.2 Overall Significance

8.2.1 Tissue specific sex-biased gene expression

The sex biased gene expression research presented in Chapter 3 uncovers new candidates for sexual dimorphism. Classically, sex biased gene expression was thought to only occur at sex chromosomes [1]. However, these results indicate that autosomal genes can also demonstrate sex differences in expression. This was the first study to utilise large genomic data to assess sex differences in gene expression across multiple human tissues. This study also shows that sex-biased gene expression is tissue specific. Furthermore, it has demonstrated sex-biased gene expression occurs in specific brain tissues. Therefore our data may provide novel sex specific targets for disease treatments and better understanding

of sex differences in disease prevalence and prognosis. Furthermore, this may have implications for clinical trials which have predominantly used male participants. In addition, when it comes to treatments for diseases our data suggest that different treatments for males and females should be explored. This is warranted because the large sex differences in gene expression may impact how drugs act and are metabolised. Thus, the research presented here in chapter 3 has implications in a range of biomedical clinical fields.

8.2.2 Novel placental specific transcripts

Previous work has uncovered approximately 50,000 previously unannotated long non-coding RNAs (lncRNAs) across a range of human tissues [2]. However, there has been no previous assessment of placental tissue for unannotated lncRNAs. The research presented in chapter 4 in this thesis fills this gap. Here, RNA-seq was used on placental tissue from first trimester, uncomplicated term and preeclamptic pregnancies. The data were used to perform a de novo transcript analysis which identified 23 large intergenic non-coding RNAs (lincRNAs). These, once previously unannotated transcripts, may potentially have roles in placental development as some were gestation specific and differentially expressed in group comparisons. Furthermore, these lincRNAs have now provided new targets for investigation of placental development and pregnancy complications. They may potentially have roles in regulating specific placental gene expression which is required for a successful uncomplicated pregnancy. However, more research is required to determine their functional roles in placental development and pregnancy complications. Nonetheless, this research highlights how little is known about gene expression in the placenta. Tools such as RNA-seq, have the potential to identify new transcripts in the placenta which may be important in placental development and identify new mechanisms in the development of pregnancy complications.

8.2.3 DNA methylation as a biomarker

DNA methylation has been proven to be an effective biomarker in a range of biomedical applications. However, this thesis has, for the first time, demonstrated the potential of

DNA methylation as a biomarker in pregnancy. The placenta undergoes dynamic changes across gestation. Although the placenta has low methylation levels compared to other human tissues, there are observable changes in DNA methylation across gestation. As detailed in Chapter 5, we developed a computational method which uses levels of DNA methylation at 62 specific sites to predict the gestational age of a placenta. The predicted gestational age can be interpreted as the biological age of a placenta. When used in placentas from early onset preeclampsia (EOPE), the predicted gestational age was greater than the chronological gestational age, which is the age at sampling or delivery. This suggests accelerated molecular aging indicating aberrant placental development and likely function. This tool can be used in other pregnancy complications to determine if accelerated placental aging is specific to EOPE. Furthermore, it also provides novel insight into the molecular mechanisms involved in PE.

DNA methylation, due to its high stability, has a strong potential as a biomarker. Besides gestational age prediction in the placenta, DNA methylation can also be used for predicting potential disease. As detailed in Chapter 7, we developed a computational tool that uses DNA methylation to predict pregnancy complications. We assessed DNA methylation from peripheral leukocytes isolated from maternal blood at 15 weeks' gestation. This has potential as a non-invasive tool in a clinical setting to identify women who will later develop a pregnancy complication, well before symptoms develop. Clinically, it may identify women at risk and enable early interventions to either prevent or reduce the severity of pregnancy complications. Models were built for each specific pregnancy complication using uncomplicated pregnancies as a reference group. The majority of sites chosen were specific to each pregnancy complication. This suggests that the onset of each pregnancy complication has a unique mechanism. Furthermore, predictive models were also built that distinguish clinical characteristics including maternal smoking and BMI.

8.3 Contributions to the field

8.3.1 Sex-biased gene expression data base

Many fields of biological research have historically neglected to account for sex differences [3, 4]. Therefore, there is a lack of sex-biased gene expression data available in

human tissues. In general, research is conducted mostly on male subjects, due to the assumption of female hormonal cycles being a confounding factor [3, 4]. Failing to account for sex differences can potentially have long lasting effects to the health and well-being of women. Some sex differences in disease prevalence and prognosis may originate from dimorphic gene expression, many of which are autosomal and not related to sex hormone profiles. The sex-biased gene expression meta-analysis has resulted in a wealth of data and therefore is potentially important for future research into sex differences in health and disease. The sex-biased genes identified in this thesis can be used as a basis to determine potential mechanisms by which sexual dimorphism occurs in disease susceptibility. In addition, it has also identified targets for sex specific treatments.

8.3.2 Placental gene expression

The work presented here comprises a comprehensive analysis of placental gene expression during first trimester and in either uncomplicated term or preeclamptic pregnancies. We have identified 7240 differentially expressed genes between placentas from first trimester and uncomplicated term pregnancies. These are two important time points during gestation and the gene lists contain genes that are essential for placental development. Furthermore, 8780 differentially expressed genes were found between placentas from first trimester and preeclamptic pregnancies. These are important analyses as they show differences that occur in PE compared to uncomplicated pregnancies across gestation.

The identification of previously unannotated lncRNAs has uncovered potential targets for disease treatment [2]. Previously, the placenta has been left out of published analyses. However, one focus of this thesis is the identification of novel lincRNAs in placenta and placenta specific transcripts. Future research using very deep sequencing is required to identify other novel transcripts that may be specific to the placenta. Furthermore, our analyses suggest there is still more to learn about placental gene expression. Additional research is needed to determine the role of the identified novel lincRNAs in this thesis. Since some were specific to either first trimester or term placental tissue, this suggests they

may have a critical role in placental development. Therefore, these lincRNAs may become furture targets to prevent or treat specific pregnancy complications.

8.3.3 Bioinformatics software

Next generation sequencing creates a wealth of complex data that is challenging to analyse. In this study, data from a methylation sensitive-genotyping by sequencing (msGBS) experiment was analysed. Although bioinformatic tools such as Tassel [5] and Stacks [6] exist which can analyse GBS data, these tools cannot be used to analyse the methylation part of the experiment. The new R package presented in this thesis enables analysis of DNA methylation from a msGBS experiment. The R package automates the generation of a table of read counts after alignment. It can also verify the recognition sites generated by the restriction enzyme. This R package fills in the missing link from data generation to downstream analyses. It has also enabled downstream analyses such as differential methylation.

8.3.4 Novel computational prediction tools for DNA methylation

DNA methylation has been proven to be an effective biomarker in many fields of biological research such as age [7, 8] and cancer [9, 10]. In this thesis, I have developed several computational tools to investigate the level and distribution of DNA methylation in human samples. The first is a computational method to predict the gestational age of a placenta. This method used DNA methylation data generated from the Illumina BeadChip Arrays, a commonly used tool to quantify DNA methylation. The predicted gestational age or DNA methylation gestational age (DNAm GA) can be seen as the biological age of a placenta. This gives the ability to assess the biological age of a placenta in pregnancy complications. As discussed in this thesis, placentas from early onset preeclampsia pregnancies aged at a faster rate than those from uncomplicated pregnancies. Therefore, this tool provides a way of assessing accelerated aging in the placenta. It can also be used to provide insight into placental aging and its relationship with pregnancy complications.

The work presented in Chapter 7 provides a possible prognostic tool for the prediction of pregnancy complications early in gestation. Many studies have focused on developing such

predictive tools. However, the work presented here is the first to use solely DNA methylation. Furthermore, this study has shown other clinical characteristics such as smoking and BMI can be identified by DNA methylation. The work presented here highlights the potential of using DNA methylation in a clinical setting. In addition, it has the potential application of one day identifying women at high risk for a pregnancy complication thereby, allowing early prevention methods to be implemented.

8.4 Strengths and Limitations

8.4.1 Phenotypic data

One of the disadvantages of using publicly available genomic data is the common lack of phenotypic data. This is because most studies either do not collect relevant information or the authors choose not to supply it when submitting their data to a publicly available data base. This has implications for future research which relies on publicly available data. Some research requires large sample sizes which can only be made possible with publicly available data. Unfortunately, since this data often does not contain all the relevant clinical information it may not all be very useful. Factors such as sample sex can influence gene expression such as shown in chapter 3. Knowing such information can assist in performing better and more detailed analyses.

The work presented in Chapter 3 focusing on sex-biased gene expression firstly had to overcome data sets lacking sex information. However, this was overcome using a Bioconductor package (massiR) which is able to determine sample sex. Although, it was not possible to determine other factors which may alter gene expression levels such as smoking, BMI, ethnicity and age as these were not supplied. Therefore, one of the limitations in this thesis, regarding the meta-analyses is not being able to block for phenotypic differences. Future publicly available data that is submitted online should supply more detailed phenotypic data which would enable more detailed comparisons and analyses, and enable researchers to test more complex scientific hypotheses which require large sample sizes.

8.4.2 Sequencing data

One of the challenges of working with sequencing data is the computational time. This thesis has had a focus on DNA methylation, a type of data where there can be millions of sites. DNA methylation data, along with many other types of genomic data, suffer from high dimensionality [11]. In other words there can be too much data for models to effectively select the best markers. Furthermore, building models using genomic data can be even more challenging since selecting the best predictor can take a long time with scarce computing resources. To overcome these challenges the research presented in this thesis has used the latest in machine learning applications to determine the best predictors for each model. For example, in Chapter 7, a feature selection method was used to reduce possible sites from approximately 2 million to within a few hundred, depending on the conditions of the model. This enabled the analysis to be performed in a relatively short time frame. This is also why so many different clinical characteristics were also analysed in this study. However, one of the pitfalls of performing such techniques as feature selection is overfitting. This occurs when a model is designed with too few data variables [12, 13]. However, mindful of this limitation, the statistics such as the sensitivity and specificity of each model were tested in both training and testing datasets with 10fold cross validation. Since we did not see a large drop in performance from training to testing data, it suggests overfitting was kept to a minimum.

8.5 Future directions

Recently, more studies have focused on using genomics to determine phenotypic traits. This field of research offers new opportunities to develop new models that may be able to predict or diagnose a wide range of diseases. Generally, it was thought the bulk of sex-biased gene expression occurred on the sex chromosomes. However, as shown in this thesis, this is not the case and sex-biased gene expression occurs on every autosome. Furthermore, as detailed in Chapter 3, many of the genes found to be sex-biased are also associated with many diseases. Future work should investigate if these genes have a role in the sexual dimorphism that occurs in many diseases. This could potentially lead the way to develop sex-specific treatments.

Although the transcriptomes in many human organs are well characterised, this thesis has shown that there is still relatively little known about placental gene expression. This tissue is important for the health of the developing feuts in utero, and therefore tissue is difficult to obtain until after delivery. RNA-seq has uncovered previously unannotated transcripts, furthering our knowledge of placental gene expression. However, the work presented here in this thesis has not identified all transcripts in the placenta. For example, there are potentially hundreds of small non-coding RNAs that have important roles in the placenta. Future research should focus on identifying and characterising these unannotated transcripts. Experiments such as sequencing to a greater depth may identify hundreds of other non-coding RNAs. In addition, future laboratory work such as cell and animal knock-out experiments are required to characterise the functions of these non-coding RNAs in vitro and in vivo. These transcripts may have important roles in placental development and disruption of their expression may lead to the development of pregnancy complications.

As shown in Chapter 5, the gestational age of a placenta can be determined by DNA methylation levels at 62 CpG sites. This work can be used in pregnancy complications to determine if placental aging occurs which may provide insight into how the disease occurs. However, predicting a disease such as pregnancy complication early on prior to disease onset has clinical application. This would enable early prevention measures to be taken and possibly prevention of the disease. A non-invasive measurement is ideal to reduce morbidity and potentially mortality. Here in this thesis, DNA methylation from leukocytes of the maternal blood was used to develop a prognostic model to predict pregnancy complications. As detailed in Chapter 7, great success was achieved using DNA methylation to predict pregnancy complications. This work has a lot to offer as a non-invasive tool in a clinical setting. However, much more research is required prior to being used a prognostic tool. For example, the models developed here in this thesis will need to be validated in other cohorts. In addition, further work is required to determine why the biomarkers identified in this study are associated with these pregnancy complications.

Overall, the work here in this thesis has shown that genomics has a lot to offer in understanding phenotypic differences in pregnancy. Much more work is required to further validate the models developed here in independent cohorts. Furthermore, sequencing

technology has allowed us to identify regions of the genome which were previously unannotated by microarray technology.

8.6 Conclusion

In this thesis the work presented has shown several aspects of gene regulation differences and the placental transcriptome. In addition, it has also demonstrated the use of DNA methylation as a biomarker of placental health and pregnancy outcome. Overall, this thesis has shown the use of genomics in the field of reproductive research. We have also demonstrated the importance of accounting for sex differences in biomedical research. The work presented here has provided a foundation for further placental research and biomarker discovery in prediction pf pregnancy complications. Finally the work presented here has the potential for applications in a real world clinical setting.

References

1. Ellegren H, Parsch J: **The evolution of sex-biased genes and sex-biased gene expression.** *Nat Rev Genet* 2007, **8:**689-698.
2. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y: **The landscape of long noncoding RNAs in the human transcriptome.** *Nat Genet* 2015, **47**.
3. Beery AK, Zucker I: **Sex bias in neuroscience and biomedical research.** *Neurosci Biobehav Rev* 2011, **35:**565-572.
4. Mogil JS, Chanda ML: **The case for the inclusion of female subjects in basic science studies of pain.** *Pain* 2005, **117:**1-5.
5. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES: **TASSEL: software for association mapping of complex traits in diverse samples.** *Bioinformatics* 2007, **23:**2633-2635.
6. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA: **Stacks: an analysis tool set for population genomics.** *Mol Ecol* 2013, **22**.
7. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S: **Genome-wide methylation profiles reveal quantitatives views of human aging rates.** *Mol Cell* 2013, **49**.
8. Horvath S: **DNA methylation age of human tissues and cell types.** *Genome Biol* 2013, **14:**R115.
9. Ulirsch J, Fan C, Knafl G, Wu MJ, Coleman B, Perou CM, Swift-Scanlan T: **Vimentin DNA methylation predicts survival in breast cancer.** *Breast Cancer Res Treat* 2013, **137:**383-396.
10. Warton K, Samimi G: **Methylation of cell-free circulating DNA in the diagnosis of cancer.** *Frontiers in Molecular Biosciences* 2015, **2:**13.
11. Libbrecht MW, Noble WS: **Machine learning applications in genetics and genomics.** *Nat Rev Genet* 2015, **16:**321-332.
12. Bilger M, Manning WG: **Measuring overfitting in nonlinear models: a new method and an application to health expenditures.** *Health Econ* 2015, **24:**75-85.
13. Waljee AK, Higgins PDR, Singal AG: **A Primer on Predictive Models.** *Clinical and Translational Gastroenterology* 2014, **5:**e44.

Publication Format: Large scale gene expression meta-analysis reveals tissue-specific, sex-biased gene expression in humans

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 October 2016
doi: 10.3389/fgene.2016.00183

Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans

Benjamin T. Mayne[1,2], *Tina Bianco-Miotto*[1,3], *Sam Buckberry*[4,5], *James Breen*[1,6], *Vicki Clifton*[7], *Cheryl Shoubridge*[1,2] *and Claire T. Roberts*[1,2]*

[1] Robinson Research Institute, University of Adelaide, Adelaide, SA, Australia, [2] Adelaide Medical School, University of Adelaide, Adelaide, SA, Australia, [3] School of Agriculture, Food and Wine, Waite Research Institute, University of Adelaide, Adelaide, SA, Australia, [4] Harry Perkins Institute of Medical Research, The University of Western Australia, Perth, WA, Australia, [5] Plant Energy Biology, Australian Research Council Centre of Excellence, The University of Western Australia, Perth, WA, Australia, [6] Bioinformatics Hub, School of Biological Sciences, University of Adelaide, Adelaide, SA, Australia, [7] Mater Research Institute, University of Queensland, Brisbane, QLD, Australia

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Francesco Russo,
University of Copenhagen, Denmark
Matteo Benelli,
University of Trento, Italy

***Correspondence:**
Claire T. Roberts
claire.roberts@adelaide.edu.au

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *12 August 2016*
Accepted: *27 September 2016*
Published: *13 October 2016*

Citation:
Mayne BT, Bianco-Miotto T, Buckberry S, Breen J, Clifton V, Shoubridge C and Roberts CT (2016) Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans. Front. Genet. 7:183. doi: 10.3389/fgene.2016.00183

The severity and prevalence of many diseases are known to differ between the sexes. Organ specific sex-biased gene expression may underpin these and other sexually dimorphic traits. To further our understanding of sex differences in transcriptional regulation, we performed meta-analyses of sex biased gene expression in multiple human tissues. We analyzed 22 publicly available human gene expression microarray data sets including over 2500 samples from 15 different tissues and 9 different organs. Briefly, by using an inverse-variance method we determined the effect size difference of gene expression between males and females. We found the greatest sex differences in gene expression in the brain, specifically in the anterior cingulate cortex, (1818 genes), followed by the heart (375 genes), kidney (224 genes), colon (218 genes), and thyroid (163 genes). More interestingly, we found different parts of the brain with varying numbers and identity of sex-biased genes, indicating that specific cortical regions may influence sexually dimorphic traits. The majority of sex-biased genes in other tissues such as the bladder, liver, lungs, and pancreas were on the sex chromosomes or involved in sex hormone production. On average in each tissue, 32% of autosomal genes that were expressed in a sex-biased fashion contained androgen or estrogen hormone response elements. Interestingly, across all tissues, we found approximately two-thirds of autosomal genes that were sex-biased were not under direct influence of sex hormones. To our knowledge this is the largest analysis of sex-biased gene expression in human tissues to date. We identified many sex-biased genes that were not under the direct influence of sex chromosome genes or sex hormones. These may provide targets for future development of sex-specific treatments for diseases.

Keywords: sex-biased gene expression, meta-analysis, microarray, human, organs

INTRODUCTION

Differences in both disease severity, prevalence, symptoms, and age of onset vary greatly between males and females (Morrow, 2015). For example, cardiovascular disease is one of the leading causes of death, affecting up to 55% of females but only 44% of males in Europe (Möller-Leimkühler, 2007). Sex differences are also evident in the risk factors for cardiovascular disease, such as diabetes which increases the risk for cardiovascular disease 2–3 fold in males but 3–7 fold in females (Eastwood and Doering, 2005). Sex differences have also been identified in the age of onset of brain diseases such as schizophrenia, where males develop symptoms between 18 and 25 years of age whereas females develop symptoms between 25 and 35 years (Ochoa et al., 2012). Moreover, reported atonic seizures in epilepsy are more frequent in males compared to females (6.5 vs. 1.7%; Carlson et al., 2014). These sex differences in diseases may be the result of tissue-specific differential gene expression between males and females. In schizophrenia, genes relating to energy metabolism have been found to have altered expression in the prefrontal cortex of only males (Qin et al., 2016). Therefore, gene expression may have a role in orchestrating sex differences in the prevalence of diseases.

Many studies neglect to account for sample sex in the design and analysis of their experiments (Mogil and Chanda, 2005; Beery and Zucker, 2011). Historically, females have been excluded from biomedical studies, due to the assumption that their hormonal cycles are a confounding factor in experimental manipulations (Zucker and Beery, 2010; Beery and Zucker, 2011). Despite females and males sharing highly similar genomes, there are numerous sex-specific traits in phenotype, physiology, and pathology. Sexually dimorphic traits can be influenced by sex chromosome genes or sex hormones, but may extend beyond these influences. Sex differences may arise through alterations in autosomal gene regulation but the true extent of sex specific differential gene regulation is not fully known. Understanding these differences may dictate that future research should consider sex as a biological confounder (Zucker and Beery, 2010). Sex differences in many traits are often small and require large sample sizes for studies to be sufficiently powered. The substantial increase in the number of large publicly available genomic data sets could assist in determining the true extent of sex-biased gene expression but to date there are no large-scale meta-analyses investigating this in adult human tissues.

Previous studies have reported sex-biased gene expression in the human brain (Vawter et al., 2004; Reinius and Jazin, 2009; Weickert et al., 2009; Kang et al., 2011; Trabzuni et al., 2013), pancreas (Hall et al., 2014), heart (Fermin et al., 2008), and liver (Zhang et al., 2011). Most studies identify sex-biased genes as those located on the sex chromosomes and it is well-known that these are a source of differentially expressed genes between the sexes (Carrel and Willard, 2005). In mammalian, female, somatic cells, one X chromosome is randomly inactivated by a process referred to as X chromosome inactivation (XCI; Carrel and Willard, 2005; Yang et al., 2010). In normal human XX females, up to 15% of genes on the X chromosome escape XCI, unlike the case in mice where very few escape inactivation (Carrel and Willard, 2005; Yang et al., 2010). Escape from XCI results in a number of genes that are expressed more highly in females compared to males. In addition, autosomal genes have also been shown to be sex-biased in human tissues including the brain (Trabzuni et al., 2013), heart (Fermin et al., 2008) and placenta (Buckberry et al., 2014b). Furthermore, sex differences in the brain in diseases such as multiple sclerosis (MS) are related to autosomal genes and are not regulated by sex chromosome genes (Voskuhl and Palaszynski, 2001; Ebers et al., 2004). These studies highlight the importance of investigating sex differences outside the context of reproductive and sex chromosome factors. In order to characterize the true extent of sex-biased gene expression in humans, we performed a large meta-analysis of publicly available microarray data. We limited our analysis to tissue samples from healthy individuals, reducing the possible effect that diseases may have on gene expression. Our analysis revealed consistencies in sex differences that are widespread in a range of human tissues. Furthermore, we have identified sex-biased genes that are disease-related, suggesting possible mechanisms for the associations of sex with an increased risk of certain diseases.

MATERIALS AND METHODS

Data Collection

Data sets were from different microarray platforms and therefore pre-processing was tailored to each platform. Briefly, data from Illumina platforms were pre-processed using Beadarray prior to quantile normalization (Dunning et al., 2007). Data from Affymetrix platforms were pre-processed and quantile normalized using the robust multiarray average (RMA) or GeneChip-RMA (GC-RMA) where appropriate that is implemented in Simpleaffy (Wilson and Miller, 2005). Batch effects in data sets were corrected for using the "combat" function in the SVA package (Leek et al., 2012). Outliers were identified and removed using ArrayQualityMetrics by analysing MA plots (Kauffmann et al., 2009).

Sample Sex Identification

To identify sample sex in each data set we used the massiR Bioconductor package (Buckberry et al., 2014a). This R package uses unsupervised clustering of probes that target Y chromosome genes to identify sample sex. In data sets where sample sex was supplied, we found an agreement in all predicted and supplied sample sex identification.

Differential Gene Expression Analysis

Probes were re-annotated to Ensembl gene identifiers using biomaRt (Durinck et al., 2009). In tissues where only one data set was found to be useable, sex-biased gene expression was determined using the Empirical Bayes methods within limma (Ritchie et al., 2015). For tissues that were present in >1 data set, differential gene expression analysis was performed using the *meta*GEM package (https://spiral.imperial.ac.uk/handle/10044/1/4217) and using the inverse–variance method as previously described (Ramasamy et al., 2008). For each probe, study specific effect sizes were calculated, by determining the mean and standard deviation for each probe which was corrected using Hedges' g (accounts for the number of samples in each dataset). *Z*

statistics were calculated for each gene identifier which was used to calculate a nominal p-value to give a corrected p-value (false discovery rate, FDR).

Androgen and Estrogen Response Elements

To determine which genes contained androgen response elements (AREs), we firstly downloaded the coordinates of AREs from JASPAR (Hu et al., 2010; Mathelier et al., 2014) and determined the positions within the genome in relation to genes and genomic locations. This was performed using the matchGenes function in the bumphunter Bioconductor package (Jaffe et al., 2012) and UCSC hg19 annotation package (BP)[1]. For estrogen response elements (EREs) we used a previous study that lists genes that are targets of ERα (Jin et al., 2004).

Identifying Enriched Transcription Factors

Transcription factor (TF) binding sites within 10 kb upstream/downstream of sex-biased genes were analyzed using oPOSSUM-3 and the JASPAR vertebrate core profiles (Kwon et al., 2012; Mathelier et al., 2014). We chose 10 kb upstream/downstream of genes as this was the largest range the oPOSSUM-3 would allow. Thus, we sought to identify all possible TF binding sites enriched within sex-biased genes. For each sex-biased gene in each tissue, the TF binding site motifs were searched with a conservation cut-off of 0.4, an 85% threshold for the matrix score and minimum specificity of 8 bits. The resulting TF analysis was limited to the most enriched TFs which were defined as those with the highest Fisher's exact test and z-score rankings.

Gene Ontology

Gene ontology (GO) analysis was performed using all human genes in the Database for Annotation, Visualization, and Integrated Discovery (DAVID) v6.7 (Huang da et al., 2009) and g:Profiler (Reimand et al., 2016). GO terms were considered significant if the corrected p-value (FDR) < 0.05.

A more detailed account of the methodology is provided in **File S1**.

RESULTS AND DISCUSSION

Overview of Publicly Available Microarray Data

Using the Gene Expression Omnibus (GEO; Barrett et al., 2013) and ArrayExpress (Brazma et al., 2003) we identified 22 microarray data sets containing a total of 2502 samples, in 15 different human tissues (**Table 1**). We excluded pooled samples and limited our analyses to data sets with >10 samples to allow better determination of sample sex. To increase the number of useable data sets we used massiR (Buckberry et al., 2014a) to identify and to verify the sample sex in all data sets. From the 22 chosen studies, 10 had sample sex metadata and within these we found concordance with all the predicted and supplied sample sex information. Female samples ($N = 803$) made up 32% of all samples across all data sets (**Table 1**).

Sex differences in autosomal gene expression are typically small so in order to increase statistical robustness, we performed multiple testing corrections in three different analyses for each tissue. We determined the adjusted p-value implemented by Benjamini and Hochberg (1995) for each autosomal gene, where (1) all the chromosomes were included, (2) the Y chromosome was excluded, and (3) both the X and Y chromosomes were excluded in the analysis (**Table 2**). In general, we observed a reduction in the number of autosomal genes that were significantly sex-biased when we removed sex chromosomes from the analysis. Since most genes located on the sex chromosomes had the smallest adjusted p-value, their removal from the analysis slightly increased the adjusted p-value for all other genes. Here we supply the adjusted p-values for all three analyses (**Tables S1**– **S3**) but discuss only autosomal genes that were significantly different in all three cases. Furthermore, the sample size in each tissue was not reflective of the total number of genes differentially expressed between males and females (**Figure 1**). For example, despite the frontal lobe of the cerebral cortex or frontal cortex (FC) and cerebellum (CB) data sets containing the greatest number of samples, with 455 and 553 samples, respectively, we detected only a small number of sex-biased genes compared to other tissues such as the anterior cingulate cortex (AnCg) and the heart which contained the greatest number of sex-biased genes with average sample sizes (**Figure 1**, **Table 2**).

Sex-Biased Gene Expression in the Human Brain

Previous studies have found sex-biased gene expression in the human brain (Vawter et al., 2004; Reinius and Jazin, 2009; Weickert et al., 2009; Kang et al., 2011). We identified five data sets for seven brain regions and our analyses showed that each region had different numbers of differentially expressed genes (**Tables 1, 2**). Our findings were consistent with previous studies (Reinius and Jazin, 2009; Weickert et al., 2009; Kang et al., 2011), whereby the most striking differences in gene expression between the sexes were sex chromosome genes. These comprised most of the sex-biased genes in the amygdala (65%; AMY) and FC (78%). However, a large proportion of sex-biased genes were autosomal in the nucleus accumbens (91%; NC), AnCg (95%), dorsolateral prefrontal cortex (91%; DLPFC), CB (60%) and the hippocampus (89%; HC). Of the 1690 autosomal sex-biased genes in AnCg, 65% were expressed more highly in males (**Figure 2A**, **Tables S1**–**S3**). Conversely, we observed a greater proportion of autosomal genes expressed more highly in females in the NC (75%), DLPFC (68%), and the HC (62%). We also found that each brain region was unique in its proportion of sex-biased genes, with as many sex-biased genes in one brain region that were not sex-biased in another (**Figure 2B**).

An increase in the expression of heat shock proteins (HSPs) has been shown to have protective roles in pro-inflammatory responses (Grundtman et al., 2011). Consistent with a previous study (Lin et al., 2011), we found genes that encode for HSPs

[1]BP., C. M. a. M., TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s). R package version 3.2.2.

TABLE 1 | Gene expression data involving 15 healthy tissues.

Organ/tissue	GEO accession	Microarray manufacturer	Samples in data set	Control samples	Sample after pre-processing	Males	Females
Bladder	GSE13507	Affymetrix	256	68	68	48	20
Brain	GSE45642	Affymetrix	670	670	659	493	166
Brain	GSE11512	Affymetrix	80	44	44	29	15
Brain	GSE54572	Affymetrix	24	12	12	5	7
Brain	GSE36192	Illumina	911	911	911	622	289
Brain	GSE44456	Affymetrix	39	39	39	28	11
Colon	GSE8671	Affymetrix	62	25	23	15	8
Colon	GSE41328	Affymetrix	20	10	10	8	2
Heart	GSE55231	Illumina	129	129	118	69	49
Heart	GSE26887	Affymetrix	24	24	23	19	4
Heart	GSE57338	Affymetrix	313	136	136	97	39
Kidney	GSE43974	Illumina	554	118	118	73	45
Kidney	GSE50892	Affymetrix	17	17	15	9	6
Liver	GSE61276	Illumina	106	50	48	22	26
Liver	GSE23649	Illumina	69	69	68	42	26
Liver	GSE38941	Affymetrix	27	10	10	4	6
Lung	GSE10072	Affymetrix	107	49	46	32	14
Lung	GSE18995	Affymetrix	35	35	34	15	19
Lung	GSE51024	Affymetrix	96	41	39	34	5
Pancreas	GSE15471	Affymetrix	78	36	35	19	16
Thyroid	GSE33630	Affymetrix	105	45	35	10	25
Thyroid	GSE65144	Affymetrix	25	13	12	7	5
		Total	3747	2551	2502	1699	803

Each row corresponds to a data set where only healthy tissue was used within this analysis. The columns report the Microarray manufacturer, total number of samples, and which data sets supplied sample sex.

TABLE 2 | Total number of sex-biased genes in each tissue.

Organ/tissue	No. of sex-biased genes (All chromosomes)	No. of autosomal sex-biased genes (Sex chromosomes included in analysis)	No. of autosomal sex-biased genes (Sex chromosomes removed)	No. of autosomal sex-biased genes (Y chromosome removed)
Bladder	16	0	0	0
Brain (Nucleus Accumbens)	264	239	216	244
Brain (Amygdala)	17	6	0	0
Brain (Cerebellum)	98	59	45	52
Brain (Anterior Cingulate Cortex)	1818	1726	1690	1728
Brain (Dorsolateral Prefrontal Cortex)	198	180	165	169
Brain (Frontal Cortex)	45	10	27	7
Brain (Hippocampus)	205	183	174	180
Colon	218	199	162	190
Heart	375	348	334	346
Kidney	224	196	194	194
Liver	32	21	16	28
Lung	36	14	2	12
Pancreas	22	0	0	0
Thyroid	163	151	133	135

Each column corresponds to the total number of genes that were differentially expressed between males and females in each analysis.

to have sex-biased expression in the human brain. Our analyses also identified genes that are involved in pro-inflammatory responses, such as those encoding interleukins, that are more highly expressed in females in NC, AnCg, DLPFC, and HC tissues (**Tables S1**–**S3**). By contrast, genes expressed more highly in males within the brain were related to energy production

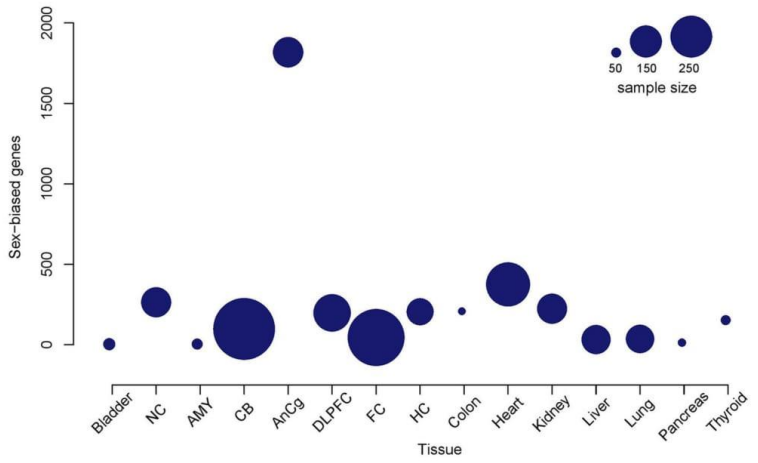


FIGURE 1 | Total number of detectable sex-biased genes relative to the sample size in each tissue. A bubble plot of each tissue where the size of the bubble is proportional to the sample size of the tissue. Bubbles that are higher on the y-axis are tissues that demonstrate a higher number of detectable sex-biased genes. Nucleus accumbens (NC); amygdala (AMY); cerebellum (CB); anterior cingulate cortex (AnCg); dorsolateral frontal cortex (DLPFC); frontal cortex (FC); hippocampus (HC).

and growth, including ATPase's and insulin-like growth factors in the HC and NC, respectively, and *GAPDH* in the AnCg. We found sex-biased genes in the NC, AnCg, and HC to be enriched for GO as defined by DAVID v6.7 for terms relating to cellular functions, the immune response and energy production (**Figures 2C–E**, **Table S4**). We also used g:Profiler (Reimand et al., 2016) for a comparison of GO terms and found similar results to what was found by DAVID v6.7. For example, in the NC, AnCg, and HC we found that the gene upregulated in females were enriched for those involved in the immune response (GO:0006955). Whereas, genes upregulated in males were found to be enriched for GO terms such as generation of precursor metabolites and energy (GO:0006091). Overall, varying proportions and types of sex-biased genes were identified within different locations of the brain, suggesting that specific cortical regions may influence sexually dimorphic traits. As mentioned above, the AnCg contained the largest number of genes differentially expressed between males and females. The AnCg is one of the most recently evolved parts of the mammalian brain (Allman et al., 2001) and also has been shown to regulate behavior and act in a sex-specific manner (Liu et al., 2012). Furthermore, previous studies have identified sex differences in mood disorders and the AnCg is known to have a role in regulating mood (Seney and Sibille, 2014; Yang et al., 2015). In mice, the AnCg has also been shown to have a critical role in sexual interest of males for females (Wu et al., 2009) and hence the large number of genes that were differentially expressed between sexes in the AnCg may assist in the explanation for sexual dimorphism in behavior.

Sex biased gene expression in the brain may potentially contribute to differences in certain neurological diseases between sexes, such as the previously mentioned epilepsy. Sex differences in gene expression may mediate these differences in susceptibility or comprise part of the mechanistic pathways involved in their pathology. Previously, sex biased gene expression in the brain has been proposed to underlie the sex differences in schizophrenia (Trabzuni et al., 2013) which has an incidence of 1.4:1 between males and females (Abel et al., 2010). We found several genes that have been associated with brain disorders to be sex-biased within specific locations of the brain. For example in the AnCg, *NOTCH3*, a gene associated with hereditary stroke disorder (Joutel et al., 1996), and *ALDH3B1*, a gene associated with schizophrenia (Wang et al., 2009), were more highly expressed in females than males. On the other hand, *KCNH3*, a gene associated with epilepsy (Zhang et al., 2010), *GABRB3*, a gene associated with schizophrenia (Huang et al., 2014), epilepsy (Gurba et al., 2012), and autism (Buxbaum et al., 2002), *SNCA*, a gene associated with Parkinson's disease (Wang et al., 2015), and *RGS4*, a gene associated with schizophrenia (Jönsson et al., 2012), were all expressed more highly in males. Recently, sex-biased gene expression has also been identified during developmental stages of the human brain (Shi et al., 2016). Furthermore, genes associated with schizophrenia have been found to be upregulated in male brains as opposed to females across different developmental stages (Shi et al., 2016). This demonstrates consistency in sex-biased genes within the human brain across different studies. Taken together, these findings suggest possible mechanisms by which sex-specific prevalence of brain disorders may occur.

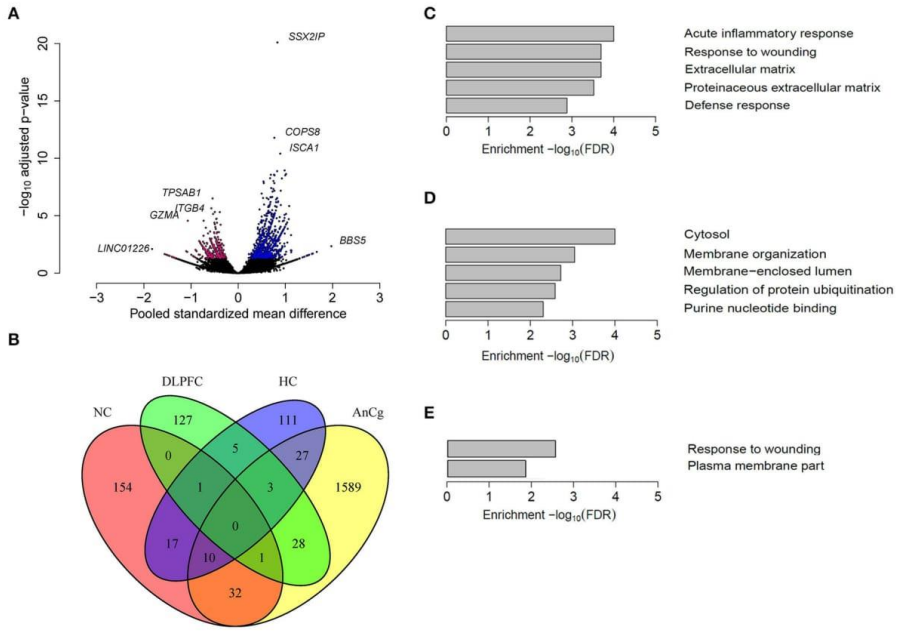


FIGURE 2 | Sex differences in autosomal gene expression in the human brain. (A) A volcano plot representing the autosomal genes that were sex-biased in the AnCg. Pink colored dots represent genes that were significantly expressed more highly in females and blue colored dots represent genes that were expressed more highly in males. **(B)** A four-way venn diagram showing the overlap of sex-biased autosomal gene expression in different regions of the human brain. Most genes that were found to be sex-biased in one region were not sex-biased in another region. The top GO terms that were enriched for sex-biased genes in **(C)** Nucleus accumbens (NC), **(D)** anterior cingulate cortex (AnCg) and **(E)** hippocampus (HC).

The Heart and Kidney Show Opposite Trends in Sex Differences in Gene Expression

Most of the heart gene expression data used in this study are from individuals with an average age of 47 years and we observed many sex differences in expression of genes associated with heart disease. It has been reported in elderly individuals (>75 years), isolated systolic hypertension can be up to 14% more prevalent in females than males (Maas and Appelman, 2010). We found *SCN10A*, a gene associated with hypertrophic cardiomyopathy (Iio et al., 2015), and *KCNE1,* a gene associated with long-QT syndrome (Splawski et al., 2000), to be expressed more highly in hearts from females. Interestingly, 62% of the 334 autosomal sex-biased genes in the heart were expressed more highly in females. The distribution of sex-biased genes across all chromosomes in the heart was similar to that in a previous study (Fermin et al., 2008). However, we report a much smaller number of sex-biased genes in the heart [375 genes in 277 samples (**Table 2**, **Table S1**) compared to 1800 genes in 102 samples in that study (Fermin et al., 2008)].

Conversely, compared to the heart, we found an opposite trend in the kidneys, with 72% of a total of 194 autosomal genes being expressed more highly in males. We also identified six genes located on chromosome 1 that were expressed more highly in females in the heart that were more abundantly expressed in males in the kidney (**Figure 3**). These genes are from the RNA U1 family (*RNU1-1*, *RNU1-2*, *RNU1-3*, *RNU1-4*, *RNVU1-7*, and *RNU1-18*) that includes genes that regulate transcription, elongation and pre-mRNA splicing events (O'Reilly et al., 2013; Guiro and O'Reilly, 2015). It has been suggested that the expression of these genes is different between tissues to regulate organ specific alternative splicing events (Guiro and O'Reilly, 2015). Sex differences in alternative splicing have also previously been detected in the brain, where it has been found to affect 2.5% of expressed genes (Trabzuni et al., 2013). Apart from RNA U1 family all other sex-biased genes were only found to be expressed more highly in one sex.

Sex Hormones and Gene Expression

Many of the sex-biased genes we identified encode enzymes that are known to regulate the production of sex hormones. In the AnCg, three genes from the sulfotransferase family that regulates sulfate conjugation in estrogen precursors (Adjei et al., 2003; *SULT2A1*, *SULT1B1*, and *SULT1C1*) were expressed more highly

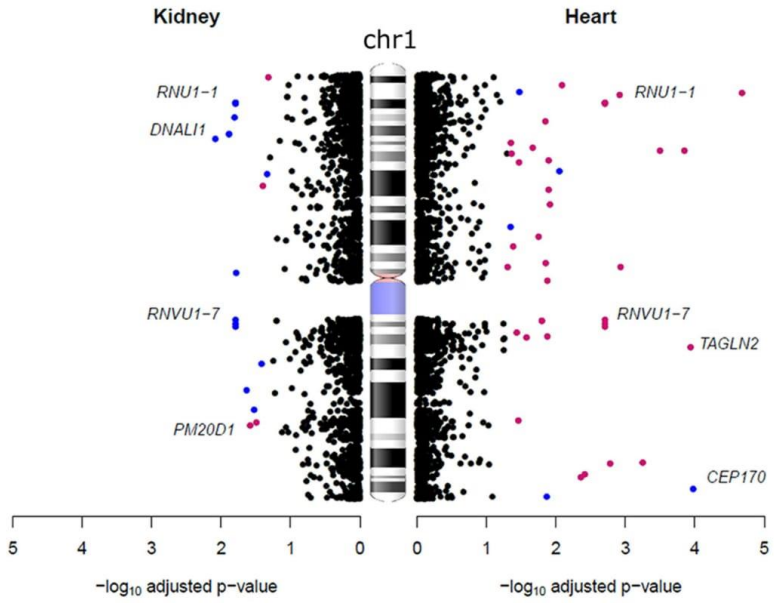


FIGURE 3 | Sex-biased gene expression differences on chromosome 1 in the heart and kidney. Each dot represents a gene, blue dots are genes that were expressed more highly in males and pink dots are those expressed more highly in females. The ideogram of chromosome 1 was obtained from the (National Centre for Biotechnology Information, 2016) (NCBI).

in females. In addition, we also found *STS* [a gene involved in the production of estrogen precursors (Miki et al., 2002)] to be expressed more highly in females in the FC and CB, as well as in the heart and lung. We did not find any major sex differences in gene expression in the bladder, liver, lung, or pancreas, apart from genes located on the sex chromosomes and those that are involved in sex hormone production. This can be contradictory to that which has been found in mouse studies where thousands of genes have been found to be sex-biased (Yang et al., 2006; van Nas et al., 2009). This may reflect an evolutionary difference between the species. Apart from the brain, we found the largest number of sex-biased gene expression differences in the heart, kidney, colon, and thyroid (**Table 2**). Thyroid hormones are known to regulate sex hormone-binding globulin (SHBG) production, which transports androgens and estrogens through the bloodstream (Selva and Hammond, 2009). In the thyroid, 133 autosomal genes were sex-biased, 75% of which were expressed more highly in males. Genes that encode for growth factors and signaling molecules were highly expressed in the thyroid of males, such as *CCL28*, a growth factor in hematopoietic stem cells (Karlsson et al., 2013), *CMTM4*, a chemokine that regulates the cell cycle (Plate et al., 2010), and *GH1*, a gene that encodes for growth hormone (Vakili et al., 2014). These findings suggest a functional role for the thyroid in influencing sexually dimorphic traits such as metabolism, as well as sex differences in thyroid hormone secretion (Ehrenkranz et al., 2015). There is also evidence to suggest that thyroid hormones significantly influence testosterone levels (Meikle, 2004).

To determine if the differentially expressed genes between sexes were regulated by sex hormones, we quantified the number of genes that contained either AREs or EREs. For AREs we downloaded the coordinates of AR binding sites from the JASPAR database (Hu et al., 2010; Mathelier et al., 2014) and for EREs we used a list of previously reported ERα targets (Jin et al., 2004). In total, we identified 3014 different genes that were expressed more highly in either sex in at least one tissue. Of the 3014 genes, 875 contained AREs, 239 contained EREs and 86 contained both. On average 32% of autosomal genes that were sex biased in tissues contained AREs or EREs. Therefore, across all tissues analyzed approximately two-thirds of autosomal genes did not contain either AREs or EREs. Four hundred and eighty-nine genes contained AREs within gene bodies such as introns and exons, 216 genes contained AREs upstream and within the promoters, and 170 genes contained AREs located downstream of the gene. The precise locations of EREs were unknown as we were using a list of previously defined ERα targets. GO enrichment for genes that contained both AREs and EREs in each individual tissue did not produce any significant enrichment, most likely due to the lists of genes being too small. We therefore found it advantageous to combine the list of genes

across different tissues since the list of genes in each tissue were too small to produce any significant results. The genes that contained either or both AREs or EREs and were expressed more highly in females were enriched for GO terms relating to response to wounding and inflammatory response. For example, we found genes related to interleukin signaling and inflammatory processes to be expressed more highly in females such as *TNFAIP6*, *IL10RB*, and *IFNA2* in the DLPFC, HC, and AnCg, respectively. On the other hand genes containing either or both AREs or EREs that were expressed more highly in males were enriched for GO terms relating to mitochondrion and generation of precursor metabolites and energy. As already mentioned, we found a variety of ATPase's to be expressed more highly in males in the AnCg, NC, DLPFC, CB, thyroid, colon, and kidney such as *ATP5G1*, *ATP6V1B2*, *ATP6V0B*, *ATP6V1C1*, and *ATP6V1A*. These results indicate that sex chromosome genes and sex hormones are key regulators of sex-biased gene expression across a range of tissues. However, our data also suggest a significant number of genes that have sex-biased expression may potentially be independent of direct influence by sex chromosomes or sex hormones.

Sex-Biased Epigenetic Modifications

Genes that are involved in the regulation of transcription and histone modifications also showed sex differences. In the colon, genes expressed more highly in males included those that encode for histones (*H3F3A*, *H3F3AP4*, *H3F3AP6*, and *H3F3BP1*) and ribosomal proteins (*RPS3A*, *RPS3AP26*, *RPS3AP6*, *RPL13A*, *RPL4*, *RPL4P4*, *RPL13AP5*, *RPS3AP5*, *RPS3AP47*, *RPL7A*, *RPL7AP6*, *RPL23AP74*, *RPL4P5*, *RPL3P4*, *RPL13AP20*, and *RPL13AP25*). These genes were also expressed more highly in males in other tissues such as the brain, heart, and kidney. It is worth mentioning that we also found other members of the RPL gene family to be more highly expressed in females in other tissues (**Tables S1–S3**). We also found sex bias in some genes that encode for enzymes that regulate histone modifications. For example, *SET*, a gene that inhibits nucleosome and histone H4 acetylation (Krajewski and Vassiliev, 2011) was expressed more highly in males in the DLPFC, *SMYD3*, a histone methyltransferase (Hamamoto et al., 2004), *PRMT2*, *PRMT5*, and *PRMT8* [histone arginine methyltransferases (Di Lorenzo and Bedford, 2011)] were more highly expressed in males in the AnCg and DLPFC (**Tables S1–S3**). Together these findings suggest that sex differences in tissue-specific gene expression extend from sex hormones and into genes that regulate gene expression and translation. Furthermore, our findings of sex bias in genes that encode for histones and histone modifying enzymes in most tissues suggest the possibility that sex-specific epigenetic modifications act on transcription that may result in phenotypic sex differences.

X-Linked Sex-Biased Gene Expression

As expected, a majority of X-linked, sex-biased genes were expressed more highly in females (**Figure 4**), with the exception of those in the AnCg in which 75% were more abundantly expressed in males. The mechanism by which genes on the single copy X chromosome in males could be expressed more highly than in females with two copies is obviously likely to be associated with XCI but another mechanism is likely to be active and requires investigation. Although we do report Y chromosome genes in our analysis (**Table 2**, **Table S1**), we do not consider these genes as differentially expressed between sexes, since females do not have a Y chromosome. We do, however, consider the reported Y chromosome genes as detectable in the analyzed tissues and act as a positive control and these genes may have potential roles in the male phenotype in these tissues. Many X-linked genes that were expressed more highly in females have been previously reported to escape XCI (Cotton et al., 2015). Not surprisingly, we consistently found *XIST* and *JPX* [genes that orchestrate XCI (Augui et al., 2011; Lee, 2011)] to be expressed more highly in females and interestingly, many sex-biased X-linked genes that are known to regulate gene expression have been defined previously (Bellott et al., 2014). For example, we found *KDM6A* (**Figure 5**), a gene that regulates chromatin modifications, to be expressed more highly in females in the liver, lung, DLPFC, NC, AMY, FC, bladder, and CB. In addition, forest plots (**Figure 5**) demonstrate consistency between individual data sets of *KDM6A* expression showing higher expression in females across different tissues. Furthermore, we also found *KDM5C* to be expressed more highly in females in the lung, FC, bladder, and CB. Genes that are involved in post-transcriptional processes and more highly expressed in females in the liver, thyroid, FC, and CB, include *ZRSR2*, *DDX3X* which are involved in alternative splicing. In addition, we also found translation regulators *EIF1AX* and *RPS4X*, to be expressed more highly in females in the lung, pancreas, HC, and colon.

Across all tissues, we found a total of 86 different genes on the X chromosome to be more highly expressed in males in at least one tissue. Twenty-two of the 86 X chromosome genes more highly expressed in males have homologous counterparts on the Y chromosome and are located within pseudoautosomal region 1 (PAR1; Ross et al., 2005), which may explain the differences in expression. However, not all X chromosome genes that were expressed more highly in males were within PAR1 or had homologous Y chromosome counterparts, such as *SMARCA2*, an ATPase and chromatin re-modeler (Takeshima et al., 2015). These findings suggest that X-linked sex-biased genes may potentially regulate autosomal gene expression such as the possible case of *SMARCA2*, through epigenetic modifications and post-transcriptional processes.

Enriched Transcription Factors

We next investigated which TFs were enriched in the sex-biased genes by running a TF binding site (TFBS) enrichment analysis using oPOSSUM-3 and the JASPAR core motifs (Kwon et al., 2012; Mathelier et al., 2014). Both the Sry-related HMG box (SOX) and the Forkhead-box (FOX) family of TFs were enriched within 10 kb of the transcription start site (TSS) of sex-biased genes across all tissues (**Table S5**). The SOX TFs are vital for sex determination (Huang et al., 2015) and the FOX TFs are essential for embryonic development and also have roles in regulating the immune system (Coffer and Burgering, 2004; Jackson et al., 2010; Lam et al., 2013). Sex chromosome derived TFs such as *ZFX* and *SRY* were also enriched within 10 kb of the TSS. We also found the androgen receptor (*AR*) as an enriched TF within the AMY,

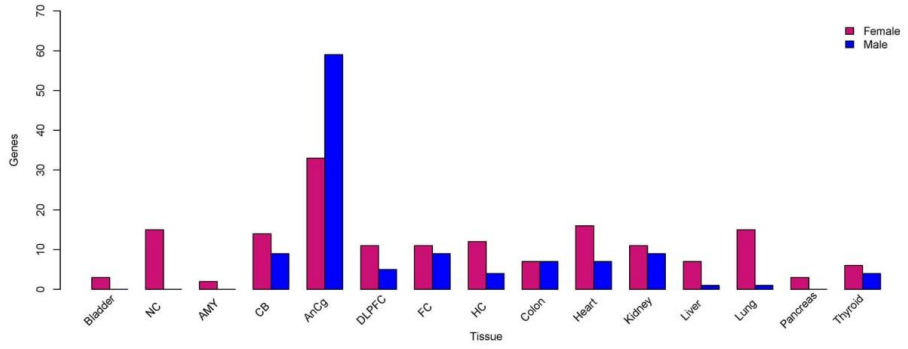


FIGURE 4 | X-linked sex-biased gene expression. The total number of genes located on the X chromosome that were expressed more highly in females (pink) and males (blue) compared to the opposite sex, respectively.

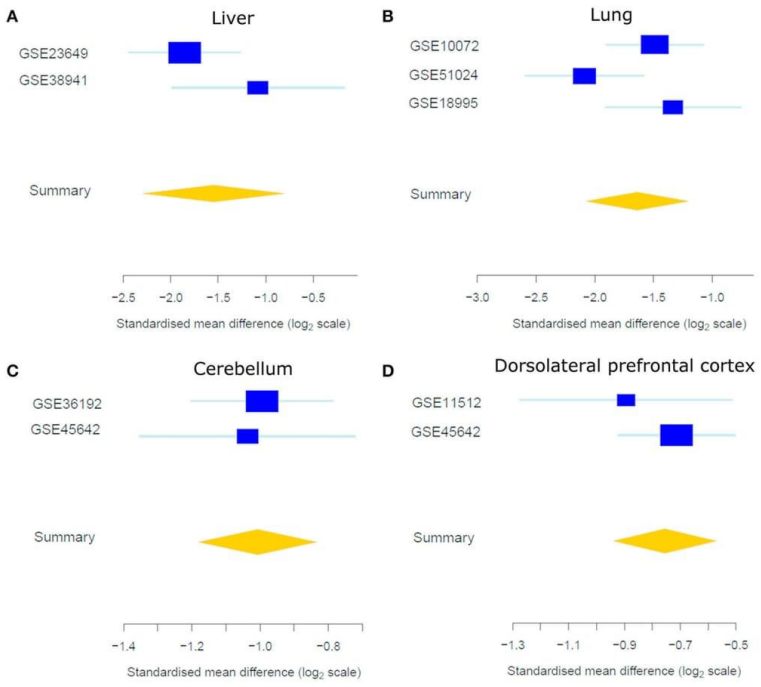


FIGURE 5 | Forest plots of the standardized mean difference of *KDM6A* expression showing higher expression in females in the liver (A), lung (B), CB (C), and DLPFC (D). Each blue box is representative of the study size in each data set and horizontal lines are standard error. The yellow diamond represents the overall gene summary for *KDM6A* in each tissue.

CB, FC, bladder, and lung. In addition, *HNF1A* and *HNF1B* were enriched in genes upregulated in both males and females within all tissues apart from the NC and DLPFC. *HNF1A* and *HNF1B* are homeobox TFs that are required for expression of specific liver genes (Shih et al., 2001). These findings reveal TFs that may have important roles in regulating sexually dimorphic gene expression such as *HNF1A* and *HNF1B*, which as mentioned earlier have only previously been shown to be required for expression of

specific liver genes (Shih et al., 2001). However, the genes that encode for the majority of the TFs that were enriched within sex-biased genes were not themselves differentially expressed between the sexes. Although in this study, we focus on gene expression, TFs undergo more processing post-transcription and therefore their protein abundance within tissues may differ between sexes.

Sex Differences in Other Tissues

In this study, we have analyzed sex-biased gene expression in 15 human tissues. However, we must acknowledge other studies that have also analyzed sex-biased gene expression. One of the largest studies that has analyzed sex-biased gene expression is the Genotype-Tissue Expression (GTEx) project (Melé et al., 2015). The GTEx project has used RNA-seq to analyse gene expression in a variety of different human tissues which would give a broader comparison of gene expression differences between tissues. In comparison to GTEx (Melé et al., 2015) we have analyzed sex-biased gene expression in five of the same human tissues which is represented as a Venn diagram (**Figure S1**). We found an overlap of sex chromosome genes as being sex-biased between this study and GTEx. However, there were many genes that we found to be sex-biased that were not in GTEx (Melé et al., 2015). A possible explanation for the difference between studies is that in GTEx only samples from175 individuals were used (Melé et al., 2015) as opposed to over 2500 in this study which provides much greater statistical power compared to GTEx (Melé et al., 2015). In addition, GTEx also used RNA-seq and were therefore able to quantify the expression of genes for which no probes were available in the microarrays used in this study.

Bias of Male Samples

To prevent any biases in our analyses we have performed differential gene expression in tissues from all publicly available data to our knowledge. However, since most studies neglect to account for samples sex (Mogil and Chanda, 2005; Beery and Zucker, 2011), we unfortunately had a ratio of 2.1:1 males to females on average across all tissues analyzed. Therefore, this in itself may create some biases in our analyses. Across all data sets (**Table 1**) the ratio of males to females was skewed toward males apart from one data set containing thyroid samples (GSE33630), where the ratio was 2.5 females for every male.

To determine if the ratio of males to females affects the differential expression analyses we conducted a 10-fold cross validation of the differential gene expression analyses in the tissue where the ratio of males to females was the greatest. The AMY gene expression data had a ratio of 4.5 males to every female. In this analysis we randomly removed male samples from the analysis to make the number of each sex the same and then assessed which genes were differentially expressed between males and females. We performed this analysis 10 times and then compared which genes were consistently identified as sex-biased to our original analysis where we did not sub-set any male samples. In the analysis with the sex chromosomes included we found the sex chromosome genes (*XIST*, *RPS4Y1*, *DDX3Y*, *KDM5D*, *USP9Y*, *EIF1AY*, and *TTTY15*) consistently classified as sex-biased in the 10-fold cross validation. However, in the original analysis we identified four autosomal genes to be sex-biased and upregulated in females (**Table S1**). However, these four autosomal genes were not found to be sex-biased in the 10-fold cross validation. By performing the 10-fold cross validation, we removed samples which would have decreased our statistical power and therefore increased the magnitude of the adjusted *p*-value which is what occurred. Therefore, caution should be taken when interpreting the results of genes that were found to be sex-biased with an adjusted *p*-value close to 0.05 and in tissues where there is a large ratio of males to females. However, this analysis does provide reassurance that the sex chromosome genes that were found to be sex-biased in the original analysis were not greatly affected by the bias in male samples.

STRENGTHS AND LIMITATIONS

While our analyses reveal many sex differences in gene expression within a variety of tissues, there are several limitations to this study. Firstly, most tissues (where age was provided) were from individuals who were post-reproductive age (average age = 47 years) which may not have captured the true extent of sex-biased gene expression that would otherwise be evident during early adulthood when sex hormones are at their peak production. Thus, using data from older individuals limited our ability to assess sex-biased gene expression in individuals of reproductive age. We also report a number of genes previously associated with diseases and disorders that were differentially expressed between sexes. RNA expression differences do not necessarily cause phenotypic variation, as there are multiple levels of gene and protein regulation that can occur post-transcription. Next-generation sequencing, as opposed to microarrays used in this study, would allow a more complete assessment of sex-dependent gene expression differences but there is currently more samples that have been analyzed using microarrays and therefore more statistical power can be achieved. Furthermore, on average, 64% of genes differentially expressed between sexes in each tissue had a magnitude $\log_2 FC < 1$. Most genes that were found to be sex-biased do not have large $\log_2 FC$ apart from genes located on the sex chromosomes. In addition, most genes that were found to be sex-biased across all tissues had a magnitude $\log_2 FC < 1.5$ (**Table S7**). Therefore, future studies would need to be adequately powered to replicate our findings. Despite these limitations, to our knowledge this is the largest analysis of sex differences in gene expression across a range of human tissues.

Despite the large amount of genomic data that was available for this study it was unfortunate not to consider clinical and lifestyle factors such as age, smoking status, sample heterogeneity and body mass index (BMI) which may potentially have an effect on gene expression. We were unable to correct for these potential confounding factors because, as detailed in **Table S6**, most studies provide little or no clinical information about the samples. Furthermore, only 32% of all the samples analyzed in this study were from females which may potentially create a bias for genes to be more highly expressed in males. However, by acknowledging this limitation we draw attention to the bias toward using only males in biomedical research. We therefore

urge future research in all fields of biomedical science to use an equal sex ratio in study design.

CONCLUSIONS

Our analyses have revealed substantial differences in the transcriptional landscape between sexes across a range of human organs and tissues and highlight possible mechanisms by which gene expression may contribute to sexually dimorphic traits. Improved understanding of these is fundamental to understanding diseases with different prevalence between the sexes. Our data show that sex differences in gene expression vary widely across different tissues. We identified a consistent trend for genes known to regulate the immune system to be more highly expressed in females and those involved in energy production and growth were more highly expressed in males. These may be the result of different evolutionary pressures between the sexes. The brain demonstrates the largest differences in sex-biased gene expression with several sex-biased genes associated with specific brain disorders, providing insight into possible mechanisms for the association of sex-specific prevalence of certain brain disorders.

Our findings also indicate that many sex biased genes within tissues are independent of sex chromosome genes or sex hormones. Approximately 32% of autosomal genes in each tissue contained an ARE or ERE, which suggests there are other mechanisms that underpin sex differences in gene expression. One potential mechanism is through epigenetic factors, such as chromatin modeling which has been suggested to have sex specific functional roles (Silkaitis and Lemos, 2014).

Finally, our data demonstrate why it is important to consider sex as a biological confounder in biomedical studies. Future studies should incorporate sex differences in their analyses which will help to provide new insights in health and disease. The sex-biased genes identified in this study provide a basis for determining the mechanism by which sexual dimorphism occurs and potential causal pathways for sexually biased disease susceptibility. More importantly however, they provide potential targets for novel sex specific treatments.

AUTHORS CONTRIBUTIONS

BM designed, conducted the study, analyzed and interpreted the data, and wrote the manuscript. SB conceived the initial part of the study and provided intellectual input into the manuscript. JB, TB, and CR were all involved in the study design, provided critical discussion and intellectual input into the manuscript. CS and VC provided critical discussion and intellectual input into the manuscript. All authors read and approved the final manuscript.

FUNDING

This project was funded in part by a National Health and Medical Research Council of Australia (NHMRC) Project Grant (GNT1059120) awarded to CR, CS, VC, and TB. CR is supported by a NHMRC Senior Research Fellowship GNT1020749. CS is supported by an Australian Research Council Future Fellowship (FT120100086). VC is supported by a NHMRC Senior Research Fellowship GNT1041918. SB is supported by an NHMRC-ARC Dementia Research Development Fellowship Grant (APP1111206). BM is supported by an Australian Post-graduate Award.

ACKNOWLEDGMENTS

The authors would like to thank the generosity of all individuals who were involved in the data creation of all data sets that were available for public analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00183

Table S1 | Differential gene expression analysis between males and females in each tissue, including the sex chromosomes. A list of differentially expressed genes between sexes in each tissue with all the chromosomes included in the analysis. A fold change > 0 indicates the gene is expressed more highly in males and a fold change < 0 indicates the gene is expressed more highly in females.

Table S2 | Differential gene expression analysis results with genes on the Y chromosome removed from the analysis. A list of sex-biased genes with the Y chromosome genes removed from the analysis. A fold change > 0 indicates the gene is expressed more highly in males and a fold change < 0 indicates the gene is expressed more highly in females.

Table S3 | Differential gene expression analysis results with genes on the X and Y chromosomes removed from the analysis. A list of sex-biased genes with the sex chromosome genes removed from the analysis. A fold change > 0 indicates the gene is expressed more highly in males and a fold change < 0 indicates the gene is expressed more highly in females.

Table S4 | Gene ontology results. This table lists all the GO terms that were found to be enriched within each tissue. Only significant GO terms were found for the NC, AnCg, and HC.

Table S5 | Transcription factors that were found to contain enriched motifs with 10 kb of the transcription start site of sex-biased genes in each tissue. A list of enriched transcription factors of the sex-biased genes in each tissue.

Table S6 | Clinical and lifestyle factors supplied by each data set. A table representing which data set supplied sample information such as age, ethnicity, sex, smoking status, and disease status.

Table S7 | Total number of sex-biased genes at different log_2FC cut-offs. A table listing the total number of genes that were found to be sex-biased in each tissue at different log_2FC cut-offs. This analysis was performed with the sex chromosomes included.

Figure S1 | Venn diagrams representing the overlap of defined sex-biased genes between this study and a previous study (Melé et al., 2015). Each Venn diagram represents an individual tissue and the overlap of genes that were found to be sex-biased between studies.

File S1 | Detailed methodology. A description of the precise methods used involved in data collection, data processing, normalization, batch correction, and differential expression.

REFERENCES

Abel, K. M., Drake, R., and Goldstein, J. M. (2010). Sex differences in schizophrenia. *Int. Rev. Psychiatry* 22, 417–428. doi: 10.3109/09540261.2010.515205

Adjei, A. A., Thomae, B. A., Prondzinski, J. L., Eckloff, B. W., Wieben, E. D., Weinshilboum, R. M., et al. (2003). Human estrogen sulfotransferase (SULT1E1) pharmacogenomics: gene resequencing and functional genomics. *Br. J. Pharmacol.* 139, 1373–1382. doi: 10.1038/sj.bjp.0705369

Allman, J. M., Hakeem, A., Erwin, J. M., Nimchinsky, E., and Hof, P. (2001). The anterior cingulate cortex. The evolution of an interface between emotion and cognition. *Ann. N.Y Acad. Sci.* 935, 107–117. doi: 10.1111/j.1749-6632.2001.tb03476.x

Augui, S., Nora, E. P., and Heard, E. (2011). Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat. Rev. Genet.* 12, 429–442. doi: 10.1038/nrg2987

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Beery, A. K., and Zucker, I. (2011). Sex bias in neuroscience and biomedical research. *Neurosci. Biobehav. Rev.* 35, 565–572. doi: 10.1016/j.neubiorev.2010.07.002

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T. J., et al. (2014). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* 508, 494–499. doi: 10.1038/nature13206

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B* 57, 289–300. doi: 10.2307/2346101

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J. Abeygunawardena, N., et al. (2003). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31, 68–71. doi: 10.1093/nar/gkg091

Buckberry, S., Bent, S. J., Bianco-Miotto, T., and Roberts, C. T. (2014a). massiR: a method for predicting the sex of samples in gene expression microarray datasets. *Bioinformatics* 30, 2084–2085. doi: 10.1093/bioinformatics/btu161

Buckberry, S., Bianco-Miotto, T., Bent, S. J., Dekker, G. A., and Roberts, C. T. (2014b). Integrative transcriptome meta-analysis reveals widespread sex-biased gene expression at the human fetal-maternal interface. *Mol. Hum. Reprod.* 20, 810–819. doi: 10.1093/molehr/gau035

Buxbaum, J. D., Silverman, J. M., Smith, C. J., Greenberg, D. A., and Kilifarski, M., Reichert, J., et al. (2002). Association between a GABRB3 polymorphism and autism. *Mol. Psychiatry* 7, 311–316. doi: 10.1038/sj.mp.4001011

Carlson, C., Dugan, P., Kirsch, H. E., and Friedman, D. (2014). Sex differences in seizure types and symptoms. *Epilepsy Behav.* 41, 103–108. doi: 10.1016/j.yebeh.2014.09.051

Carrel, L., and Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434, 400–404. doi: 10.1038/nature03479

Coffer, P. J., and Burgering, B. M. (2004). Forkhead-box transcription factors and their role in the immune system. *Nat. Rev. Immunol.* 4, 889–899. doi: 10.1038/nri1488

Cotton, A. M., Price, E. M., Jones, M. J., Balaton, B. P., Kobor, M. S., Brown, C. J., et al. (2015). Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* 24, 1528–1539. doi: 10.1093/hmg/ddu564

Di Lorenzo, A., and Bedford, M. T. (2011). Histone arginine methylation. *FEBS Lett.* 585, 2024–2031. doi: 10.1016/j.febslet.2010.11.010

Dunning, M. J., Smith, M. L., Ritchie, M. E., and Tavaré, S. (2007). beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* 23, 2183–2184. doi: 10.1093/bioinformatics/btm311

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97

Eastwood, J. A., and Doering, L. V. (2005). Gender differences in coronary artery disease. *J. Cardiovasc. Nurs.* 20, 340–351. *quiz*: 352–343. doi: 10.1097/00005082-200509000-00008

Ebers, G. C., Sadovnick, A. D., Dyment, D. A., Yee, I. M., Willer, C. J., and Risch, N. (2004). Parent-of-origin effect in multiple sclerosis: observations in half-siblings. *Lancet* 363, 1773–1774. doi: 10.1016/S0140-6736(04)16304-6

Ehrenkranz, J., Bach, P. R., Snow, G. L., Schneider, A., Lee, J. L., Ilstrup, S., et al. (2015). Circadian and circannual rhythms in thyroid hormones: determining the TSH and Free T4 reference intervals based upon time of day, age, and sex. *Thyroid* 25, 954–961. doi: 10.1089/thy.2014.0589

Fermin, D. R., Barac, A., Lee, S., Polster, S. P., Hannenhalli, S., Bergemann, T. L., et al. (2008). Sex and age dimorphism of myocardial gene expression in nonischemic human heart failure. *Circ. Cardiovasc. Genet.* 1, 117–125. doi: 10.1161/CIRCGENETICS.108.802652

Grundtman, C., Kreutmayer, S. B., Almanzar, G., Wick, M. C., and Wick, G. (2011). Heat shock protein 60 and immune inflammatory responses in atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* 31, 960–968. doi: 10.1161/ATVBAHA.110.217877

Guiro, J., and O'Reilly, D. (2015). Insights into the U1 small nuclear ribonucleoprotein complex superfamily. *Wiley Interdiscip. Rev. RNA* 6, 79–92. doi: 10.1002/wrna.1257

Gurba, K. N., Hernandez, C. C., Hu, N., and Macdonald, R. L. (2012). GABRB3 mutation, G32R, associated with childhood absence epilepsy alters alpha1beta3gamma2L gamma-aminobutyric acid type A (GABAA) receptor expression and channel gating. *J. Biol. Chem.* 287, 12083–12097. doi: 10.1074/jbc.M111.332528

Hall, E., Volkov, P., Dayeh, T., Esguerra, J. L., Salö, S., Taneera, J., et al. (2014). Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets. *Genome Biol.* 15:522. doi: 10.1186/s13059-014-0522-z

Hamamoto, R., Furukawa, Y., Morita, M., Iimura, Y., Silva, F. P., Li, M., et al. (2004). SMYD3 encodes a histone methyltransferase involved in the proliferation of cancer cells. *Nat. Cell. Biol.* 6, 731–740. doi: 10.1038/ncb1151

Hu, S., Yao, G., Guan, X., Ni, Z., Ma, W., Wilson, E. M., et al. (2010). Research resource: genome-wide mapping of *in vivo* androgen receptor binding sites in mouse epididymis. *Mol. Endocrinol.* 24, 2392–2405. doi: 10.1210/me.2010-0226

Huang, C. C., Cheng, M. C., Tsai, H. M., Lai, C. H., and Chen, C. H. (2014). Genetic analysis of GABRB3 at 15q12 as a candidate gene of schizophrenia. *Psychiatr. Genet.* 24, 151–157. doi: 10.1097/YPG.0000000000000032

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, Y. H., Jankowski, A., Cheah, K. S., Prabhakar, S., and Jauch, R. (2015). SOXE transcription factors form selective dimers on non-compact DNA motifs through multifaceted interactions between dimerization and high-mobility group domains. *Sci. Rep.* 5:10398. doi: 10.1038/srep10398

Iio, C., Ogimoto, A., Nagai, T., Suzuki, J., Inoue, K., Nishimura, K., et al. (2015). Association between genetic variation in the scn10a gene and cardiac conduction abnormalities in patients with hypertrophic cardiomyopathy. *Int. Heart J.* 56, 421–427. doi: 10.1536/ihj.14-411

Jackson, B. C., Carpenter, C., Nebert, D. W., and Vasiliou, V. (2010). Update of human and mouse forkhead box (FOX) gene families. *Hum. Genomics* 4, 345–352. doi: 10.1186/1479-7364-4-5-345

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., et al. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* 41, 200–209. doi: 10.1093/ije/dyr238

Jin, V. X., Leu, Y.-W., Liyanarachchi, S., Sun, H., Fan, M., Andreassen, O. A., et al. (2004). Identifying estrogen receptor α target genes using integrated computational genomics and chromatin immunoprecipitation microarray. *Nucleic Acids Res.* 32, 6627–6635. doi: 10.1093/nar/gkh1005

Jönsson, E. G., Saetre, P., Nyholm, H., Djurovic, S., Melle, I., Andreassen, O. A., et al. (2012). Lack of association between the regulator of G-protein signaling 4 (RGS4) rs951436 polymorphism and schizophrenia. *Psychiatr. Genet.* 22, 263–264. doi: 10.1097/YPG.0b013e32834f3558

Joutel, A., Corpechot, C., Ducros, A., Vahedi, K., Chabriat, H., Mouton, P., et al. (1996). Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature* 383, 707–710. doi: 10.1038/383707a0

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M.,et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature* 478, 483–489. doi: 10.1038/nature10523

Karlsson, C., Baudet, A., Miharada, N., Soneji, S., Gupta, R., Magnusson, M., et al. (2013). Identification of the chemokine CCL28 as a growth and survival factor

for human hematopoietic stem and progenitor cells. *Blood* 121, 3838–3842, S3831–S3815. doi: 10.1182/blood-2013-02-481192

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics* 25, 415–416. doi: 10.1093/bioinformatics/btn647

Krajewski, W. A., and Vassiliev, O. L. (2011). Interaction of SET domains with histones and nucleic acid structures in active chromatin. *Clin. Epigenetics* 2, 17–25. doi: 10.1007/s13148-010-0015-1

Kwon, A. T., Arenillas, D. J., Worsley Hunt, R., and Wasserman, W. W. (2012). oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 (Bethesda)* 2, 987–1002. doi: 10.1534/g3.112.003202

Lam, E. W. F., Brosens, J. J., Gomes, A. R., and Koo, C. Y. (2013). Forkhead box proteins: tuning forks for transcriptional harmony. *Nat. Rev. Cancer* 13, 482–495. doi: 10.1038/nrc3539

Lee, J. T. (2011). Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nat. Rev. Mol. Cell. Biol.* 12, 815–826. doi: 10.1038/nrm3231

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi: 10.1093/bioinformatics/bts034

Lin, L. C., Lewis, D. A., and Sibille, E. (2011). A human-mouse conserved sex bias in amygdala gene expression related to circadian clock and energy metabolism. *Mol. Brain* 4:18. doi: 10.1186/1756-6606-4-18

Liu, J., Zubieta, J. K., and Heitzeg, M. (2012). Sex differences in anterior cingulate cortex activation during impulse inhibition and behavioral correlates. *Psychiatry Res.* 201, 54–62. doi: 10.1016/j.pscychresns.2011.05.008

Maas, A. H. E. M., and Appelman, Y. E. A. (2010). Gender differences in coronary heart disease. *Neth. Heart J.* 18, 598–602. doi: 10.1007/s12471-010-0841-y

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., et al. (2014). JASPAR 2014, an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42, D142–D147. doi: 10.1093/nar/gkt997

Meikle, A. W. (2004). The interrelationships between thyroid dysfunction and hypogonadism in men and boys. *Thyroid* 14(Suppl 1.), S17–S25. doi: 10.1089/105072504323024552

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). The human transcriptome across tissues and individuals. *Science* 348, 660–665. doi: 10.1126/science.aaa0355

Miki, Y., Nakata, T., Suzuki, T., Darnel, A. D., Moriya, T., Kaneko, C., et al. (2002). Systemic distribution of steroid sulfatase and estrogen sulfotransferase in human adult and fetal tissues. *J. Clin. Endocrinol. Metab.* 87, 5760–5768. doi: 10.1210/jc.2002-020670

Mogil, J. S., and Chanda, M. L. (2005). The case for the inclusion of female subjects in basic science studies of pain. *Pain* 117, 1–5. doi: 10.1016/j.pain.2005.06.020

Möller-Leimkühler, A. M. (2007). Gender differences in cardiovascular disease and comorbid depression. *Dialogues Clin. Neurosci.* 9, 71–83.

Morrow, E. H. (2015). The evolution of sex differences in disease. *Biol. Sex Dif.* 6:5. doi: 10.1186/s13293-015-0023-0

National Centre for Biotechnology Information (2016). U. S. N. L. o. M. w. n. n. n. g. g. t. g. A. D. Available online at: https://www.ncbi.nlm.nih.gov/genome/tools/gdp

O'Reilly, D., Dienstbier, M., Cowley, S. A., Vazquez, P., Drozdz, M., Taylor, S., et al. (2013). Differentially expressed, variant U1 snRNAs regulate gene expression in human cells. *Genome Res.* 23, 281–291. doi: 10.1101/gr.142968.112

Ochoa, S., Usall, J., Cobo, J., Labad, X., and Kulkarni, J. (2012). Gender differences in schizophrenia and first-episode psychosis: a comprehensive literature review. *Schizophr. Res. Treat.* 2012:916198. doi: 10.1155/2012/916198

Plate, M., Li, T., Wang, Y., Mo, X., Zhang, Y., Ma, D., et al. (2010). Identification and characterization of CMTM4, a novel gene with inhibitory effects on HeLa cell growth through Inducing G2/M phase accumulation. *Mol. Cells* 29, 355–361. doi: 10.1007/s10059-010-0038-7

Qin, W., Liu, C., Sodhi, M., and Lu, H. (2016). Meta-analysis of sex differences in gene expression in schizophrenia. *BMC Syst. Biol.* 10(Suppl. 1), 9. doi: 10.1186/s12918-015-0250-3

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* 5:e184. doi: 10.1371/journal.pmed.0050184

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., et al. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* 44, W83–W89. doi: 10.1093/nar/gkw199

Reinius, B., and Jazin, E. (2009). Prenatal sex differences in the human brain. *Mol. Psychiatry* 14, 988–989. doi: 10.1038/mp.2009.79

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., et al. (2005). The DNA sequence of the human X chromosome. *Nature* 434, 325–337. doi: 10.1038/nature03440

Selva, D. M., and Hammond, G. L. (2009). Thyroid hormones act indirectly to increase sex hormone-binding globulin production by liver via hepatocyte nuclear factor-4alpha. *J. Mol. Endocrinol.* 43, 19–27. doi: 10.1677/JME-09-0025

Seney, M. L., and Sibille, E. (2014). Sex differences in mood disorders: perspectives from humans and rodent models. *Biol. Sex Dif.* 5:17. doi: 10.1186/s13293-014-0017-3

Shi, L., Zhang, Z., and Su, B. (2016). Sex biased gene expression profiling of human brains at major developmental stages. *Sci. Rep.* 6:21181. doi: 10.1038/srep21181

Shih, D. Q., Bussen, M., Sehayek, E., Ananthanarayanan, M., Shneider, B. L., Suchy, F. J., et al. (2001). Hepatocyte nuclear factor-1alpha is an essential regulator of bile acid and plasma cholesterol metabolism. *Nat. Genet.* 27, 375–382. doi: 10.1038/86871

Silkaitis, K., and Lemos, B. (2014). Sex-biased chromatin and regulatory cross-talk between sex chromosomes, autosomes, and mitochondria. *Biol. Sex Dif.* 5:2. doi: 10.1186/2042-6410-5-2

Splawski, I., Shen, J., Timothy, K. W., Lehmann, M. H., Priori, S., Robinson, J. L., et al. (2000). Spectrum of mutations in long-QT syndrome genes. KVLQT1, HERG, SCN5A, KCNE1, and KCNE2. *Circulation* 102, 1178–1185. doi: 10.1161/01.CIR.102.10.1178

Takeshima, H., Niwa, T., Takahashi, T., Wakabayashi, M., Yamashita, S., Ando, T., et al. (2015). Frequent involvement of chromatin remodeler alterations in gastric field cancerization. *Cancer Lett.* 357, 328–338. doi: 10.1016/j.canlet.2014.11.038

Trabzuni, D., Ramasamy, A., Imran, S., Walker, R., Smith, C., Weale, M. E, et al. (2013). Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* 4:2771. doi: 10.1038/ncomms3771

Vakili, H., Jin, Y., and Cattini, P. A. (2014). Energy homeostasis targets chromosomal reconfiguration of the human GH1 locus. *J. Clin. Invest.* 124, 5002–5012. doi: 10.1172/JCI77126

van Nas, A., Guhathakurta, D., Wang, S. S., Yehya, N., Horvath, S., Zhang, B., et al. (2009). Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. *Endocrinology* 150, 1235–1249. doi: 10.1210/en.2008-0563

Vawter, M. P., Evans, S., Choudary, P., Tomita, H., Meador-Woodruff, J., Molnar, M., et al. (2004). Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology* 29, 373–384. doi: 10.1038/sj.npp.1300337

Voskuhl, R. R., and Palaszynski, K. (2001). Sex hormones in experimental autoimmune encephalomyelitis: implications for multiple sclerosis. *Neuroscientist* 7, 258–270. doi: 10.1177/107385840100700310

Wang, X., Yu, S., Li, F., and Feng, T. (2015). Detection of alpha-synuclein oligomers in red blood cells as a potential biomarker of Parkinson's disease. *Neurosci. Lett.* 599, 115–119. doi: 10.1016/j.neulet.2015.05.030

Wang, Y., Hu, Y., Fang, Y., Zhang, K., Yang, H., Ma, J., et al. (2009). Evidence of epistasis between the catechol-O-methyltransferase and aldehyde dehydrogenase 3B1 genes in paranoid schizophrenia. *Biol. Psychiatry* 65, 1048–1054. doi: 10.1016/j.biopsych.2008.11.027

Weickert, C. S., Elashoff, M., Richards, A. B., Sinclair, D., Bahn, S., Paabo, S., et al. (2009). Transcriptome analysis of male-female differences in prefrontal cortical development. *Mol. Psychiatry* 14, 558–561. doi: 10.1038/mp.2009.5

Wilson, C. L., and Miller, C. J. (2005). Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* 21, 3683–3685. doi: 10.1093/bioinformatics/bti605

Wu, L. J., Kim, S. S., Li, X., Zhang, F., and Zhuo, M. (2009). Sexual attraction enhances glutamate transmission in mammalian anterior cingulate cortex. *Mol. Brain* 2:9. doi: 10.1186/1756-6606-2-9

Yang, F., Babak, T., Shendure, J., and Disteche, C. M. (2010). Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.* 20, 614–622. doi: 10.1101/gr.103200.109

Yang, X., Schadt, E. E., Wang, S., Wang, H., Arnold, A. P., Ingram-Drake, L., et al. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* 16, 995–1004. doi: 10.1101/gr.5217506

Yang, X., Wang, S., Kendrick, K. M., Wu, X., Yao, L., Lei, D., et al. (2015). Sex differences in intrinsic brain functional connectivity underlying human shyness. *Soc. Cogn. Affect. Neurosci.* 10, 1634–1643. doi: 10.1093/scan/nsv052

Zhang, X., Bertaso, F., Yoo, J. W., Baumgärtel, K., Clancy, S. M., Lee, V., et al. (2010). Deletion of the potassium channel Kv12.2 causes hippocampal hyperexcitability and epilepsy. *Nat. Neurosci.* 13, 1056–1058. doi: 10.1038/nn.2610

Zhang, Y., Klein, K., Sugathan, A., Nassery, N., Dombkowski, A., Zanger, U. M., et al. (2011). Transcriptional profiling of human liver identifies sex-biased genes associated with polygenic dyslipidemia and coronary artery disease. *PLoS ONE* 6:e23506. doi: 10.1371/journal.pone.0023506

Zucker, I., and Beery, A. K. (2010). Males still dominate animal studies. *Nature* 465:690. doi: 10.1038/465690a

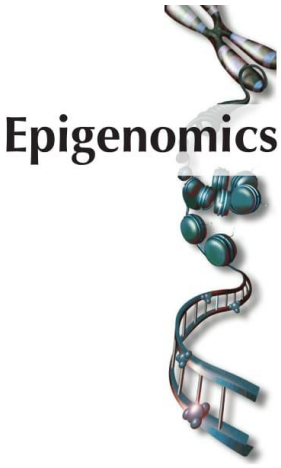
Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publication Format: Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation

SPECIAL FOCUS I Effects of the *in utero* environment on the epigenome

Research Article

For reprint orders, please contact: reprints@futuremedicine.com



Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation

Benjamin T Mayne[1,2], Shalem Y Leemaqz[1,2], Alicia K Smith[3], James Breen[1,4], Claire T Roberts[1,2] & Tina Bianco-Miotto*[1,5]

[1]Robinson Research Institute, University of Adelaide, SA, 5005, Australia
[2]Adelaide Medical School, University of Adelaide, SA, 5005, Australia
[3]Department of Gynecology and Obstetrics & Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA 30322, USA
[4]Bioinformatics Hub, School of Biological Sciences, University of Adelaide, SA, 5005, Australia
[5]Waite Research Institute, School of Agriculture, Food & Wine, University of Adelaide, SA, 5005, Australia
*Author for correspondence: tina.bianco@adelaide.edu.au

Aim: To determine whether dynamic DNA methylation changes in the human placenta can be used to predict gestational age. **Materials & methods:** Publicly available placental DNA methylation data from 12 studies, together with our own dataset, using Illumina Infinium Human Methylation BeadChip arrays. **Results & conclusion:** We developed an accurate tool for predicting gestational age of placentas using 62 CpG sites. There was a higher predicted gestational age for placentas from early onset preeclampsia cases, but not term preeclampsia, compared with their chronological age. Therefore, early onset preeclampsia is associated with placental aging. Gestational age acceleration prediction from DNA methylation array data may provide insight into the molecular mechanisms of pregnancy disorders.

First draft submitted: 10 August 2016; Accepted for publication: 1 November 2016; Published online: 29 November 2016

Keywords: DNA methylation • gestational age • placenta

DNA methylation is a heritable epigenetic process that can regulate important genetic mechanisms and processes such as gene expression, X chromosome inactivation (XCI) [1], cellular identity [2] and genomic imprinting [3]. DNA methylation is the covalent attachment of a methyl group to a cytosine ring by a DNA methyl transferase [4]. In this article we focus on cytosine-5 DNA methylation within CpG dinucleotides as opposed to other methylated cytosines such as CHG and CHH. Recently, DNA methylation levels for 353 CpG sites have been used to measure the epigenetic age, defined as the predicted age by DNA methylation, of a variety of human tissues, which has a high correlation (r = 0.92) with the actual age [5]. Moreover, another study found that accelerated aging [5], defined as the difference between the estimated epigenetic age and the actual known age, is associated with maternal characteristics in pregnancy such as smoking, weight, BMI, selenium and cholesterol in peripheral blood [6]. The placenta is unique compared with other tissues, with the exception of cancer tissues [7,8] and a human fetal fibroblast cell line (IMR90) [9], in that it has been shown that it has low levels of genome-wide CpG methylation [10,11]. Despite this, overall placental genome CpG methylation has been observed to increase during gestation [12]. However, precisely what DNA methylation changes occur during gestation and how these changes relate to pregnancy success is unknown.

Poor placental function, due to impaired placentation has been proposed to be a cause of preeclampsia (PE) [13], which is characterized by high-maternal blood pressure and proteinuria [14]. Adversities during pregnancy may cause epigenetic changes and altered fetal development outcomes [15], which may be orchestrated by the placenta [16]. Differential DNA methylation in the placenta has been shown to occur in pregnancy complications [17] including PE [18–21], gestational diabetes mellitus [18,22,23] and intrauterine growth restriction [24]. DNA methylation is

Future Medicine part of fsg

Table 1. A description of the DNA methylation datasets containing placental tissue used in this study.

GEO accession	Dataset summary	Platform	Number of placental tissue samples	Gestational age range (weeks)
GSE31781	Placental tissue samples from three gestational age time points	27k	18 first trimester, 10 second trimester, 14 term	8–42
GSE36642	Cord blood and placentas from MZ and DZ twins	27k	28 third trimester	32–38
GSE36829	Placental tissue from term pregnancies	27k	48 term	37–42
GSE59274	Placental tissue samples from women with PE or uncomplicated pregnancies	27k	24 uncomplicated, 24 PE	28–41
GSE46573	Epigenome-wide association study	450k	2 term	NA
GSE52576	Genome-wide human imprinting analysis of different healthy human tissue	450k	4 term	NA
GSE54399	Whole cord blood and placental tissue from normal pregnancies	450k	24 uncomplicated	NA
GSE57767	Placental tissue samples from normal term, preterm PE and term PE women	450k	14 uncomplicated, 12 preterm PE, 19 term PE	NA
GSE44667	Placental tissue from women with EOPET and gestational age matched controls	450k	20 third trimester, 20 PE	25–37
GSE66210	First trimester chorionic villus samples from normal and trisomy pregnancies	450k	12 normal, 12 trisomy 21, 12 trisomy 18, 6 trisomy 13	NA
GSE73375	Placental tissue samples from preeclamptic and normotensive women	450k	17 uncomplicated, 19 PE	NA
GSE74738	Placental tissue from uncomplicated pregnancies	450k	28 term	36–42
This study	Placental tissue from uncomplicated elective cesarean pregnancies	27k	22 term	35–40

27k: Illumina Infinium Human Methylation27; 450k: Illumina Infinium HumanMethylation450; DZ: Dizygotic; EOPET: Early onset preeclampsia; GEO: Gene Expression Omnibus; MZ: Monozygotic; NA: Not available; PE: Preeclampsia.

critical for optimal placental and fetal development. For example, genomic imprinting regulates expression of *IGF2* and *H19*, which are both required for proper placental development. Hypomethylation within the imprinting control region of *IGF2* has been shown to be associated with reduced fetal growth [25,26] whereas loss of imprinting of *H19* has been reported to be associated with PE and with small for gestational age infants [27].

Previous studies have reported altered placental gene expression in human pregnancy complications including PE [28–32], gestational diabetes mellitus [31] and fetal growth restriction [28,33]. However, little is known whether changes in DNA methylation and gene expression overlap in the placenta. The relationship between global placental gene expression and DNA methylation is conflicting as one study has reported a general trend between the increase in DNA methylation and decrease in gene expression levels during gestation [12], whereas another study using matched samples has reported no overlap between gene expression and DNA methylation changes between placentas from PE and uncomplicated pregnancies [21]. In addition, genes within partially methylated domains in the term placenta have been reported to be repressed [10].

In this study, we assembled a large dataset of publicly available placental tissue DNA methylation data, which has been measured using the Illumina Infinium HumanMethylation BeadChip arrays (Illumina). We sought to determine the precise changes in DNA methylation that occur in the placenta across gestation and determined whether DNA methylation data can be used to predict the gestational age of a placenta. Finally, we investigated what happens to the predicted gestational age in placentas from preeclamptic pregnancies. We hypothesized that the gestational age of the placenta can be estimated by its DNA methylation levels and that pregnancy complications such as PE would be characterized by accelerated placental aging. Our computational analysis of DNA methylation data reveals accurate predictions of the gestational age of the placenta. Moreover,

gestational age prediction of the placenta may identify novel mechanisms in pregnancy complications, such as those associated with placental aging.

Materials & methods

Quantification of the DNA methylation levels

Quantification of the DNA methylation level of each CpG site that is annotated in either the Illumina Infinium HumanMethylation27 or 450 BeadChip (Illumina) was performed using standard techniques. The publicly available human placental datasets (Table 1) were obtained from the Gene Expression Omnibus using the GEOquery Bioconductor package. Briefly, DNA methylation levels of each CpG site from each dataset were quantified by the β-value. The β-value is calculated by taking the ratio of the two fluorescent signals (methylated and unmethylated signals) in studies that provided the two fluorescent signals.

Differential methylation analysis

Differential methylation analysis was performed using the 450k datasets for which we had data for 469,017 probes. Normalization was performed using the Beta MIxture Quantile dilation (BMIQ) method [34], which corrects for the probe design bias in the Illumina Infinium HumanMethylation450 BeadChip followed by quantile normalization. Since the data were from multiple datasets, batch effects were corrected using the ComBat function in the ChAMP Bioconductor package [35,36]. Multidimensional scaling plots of the 1000 most variable probes of the data were used to check for outliers. Sample sex was identified using the minfi package in which the median value of the β-values for probes that mapped uniquely for the X and Y chromosome, respectively, were determined [37]. Differentially methylated CpG sites were identified using empirical Bayesian variance method in limma [38]. The Bumphunter Bioconductor package was used to identify differentially methylated regions (DMRs) by running 1000 permutations of the data [39]. We selected datasets (GSE44667, GSE46573, GSE52576, GSE54399, GSE57767 and GSE73375) that had used the Illumina Infinium HumanMethylation450 BeadChip array to assess sex differences in placental DNA methylation and differential methylation between PE and uncomplicated pregnancies (Table 1). In total, we had placental DNA methylation data for 70 preeclamptic and 62 uncomplicated pregnancies.

Gestational age prediction

We selected datasets for placental tissue samples that were publicly available and which contained individual gestational age information and were from healthy singleton pregnancies (GSE31781, GSE36829, GSE59274, GSE44667 and GSE74738) (Table 1). We also combined these datasets with our own dataset for 22 term placentas from uncomplicated pregnancies. Probes that were present in both the 27k and 450k array were kept, and probes that were annotated to the sex chromosomes were removed. Every sample in the training dataset contained β-values for 18,437 probes. Normalization was performed as described by Horvath [5], using a modified version of the BMIQ [34] method. This modified version of BMIQ rescales the Infinium II probes to the mean β-value of each probe in the largest dataset (GSE36829) prior to quantile normalization. Using the R package glmnet [40], we regressed gestational age with the 18,437 probes. The elastic net regression or the α-parameter of glmnet was 0.5 and the minimum λ-value based on the training data was 0.6807. The elastic net model automatically selected 62 CpG sites, such that given the level of methylation of the 62 CpG sites, the gestational age of a placenta can be calculated.

Gestational age acceleration heritability

To determine the heritability of gestational age acceleration, which is defined as the difference between the chronological and predicted gestational age of a placenta we used Falconer's formula (H^2 = 2(cor(MZ)-cor(DZ))). Falconer's formula was used to determine the broad sense heritability, which is the proportion of variance of gestational age acceleration as a result of genetic factors. The broad sense heritability was determined by using a dataset (GSE36642) that contained monozygotic (MZ) and dizygotic (DZ) twins of the same sex. First, the predicted gestational age of the twin placentas were determined as described above. The twin dataset was split into either MZ or DZ twins, and for each twin pair the gestational age acceleration was calculated. For each twin pair, each sibling was randomly selected as either twin 1 or twin 2 and the pearson correlation of gestational age acceleration was determined in both MZ and DZ twins. The correlations of both MZ and DZ twins were inputted into Falconer's formula to determine the broad sense heritability.

Annotation of CpG sites

Annotation of all the CpG sites within the analysis of this study for the Illumina Infinium HumanMethylation27 and HumanMethylation450 BeadChip was performed using two annotation Bioconductor packages [41,42].

Results

Differential methylation in placentas from preeclamptic pregnancies

We searched Gene Expression Omnibus [43] for DNA methylation datasets containing placental tis-

sue that were measured on either the Illumina Infinium HumanMethylation27 BeadChip or Infinium HumanMethylation450 BeadChip. In total, we identified 387 placental tissue samples from 12 different datasets (Table 1). We selected only placentas assessed using the Illumina DNA methylation arrays, the most commonly used platform to quantify DNA methylation in the placenta. Here, we used publicly available placental DNA methylation data (Table 1) to assess differential methylation between placentas from PE and uncomplicated pregnancies. We selected only datasets that used the Infinium HumanMethylation450 BeadChip for differential methylation analysis since most studies involving placental samples from women with PE had used this platform (Table 1) and it contains the largest number of CpG sites available for analysis. When comparing 70 placentas from preeclamptic pregnancies and 62 placentas from term uncomplicated pregnancies we found that a total of 741 CpG sites (false discovery rate [FDR] <0.01) were differentially methylated (Supplementary Table 1). We then tested for DMRs and identified three DMRs in placentas from preeclamptic pregnancies, which overlapped the 5′ region of *MARC2*, *FAM3B* and *TP53TG1*.

Sex differences in DNA methylation

To identify whether sex differences also occur in the placental DNA methylome, we identified a total of 2898 differentially methylated CpG sites (FDR <0.01) between 35 male and 27 female placentas from uncomplicated singleton term pregnancies that had been analyzed using the Infinium HumanMethylation450 BeadChip (Supplementary Table 2). Of the 2898 CpG sites, 420 were located on autosomes, 2464 on the X chromosome and 14 on the Y chromosome. Although we are reporting Y chromosome CpG sites, we do not consider these CpG sites as differentially methylated between sexes. In addition, upon removing the Y chromosome from our analysis we did not observe a difference in the total number of differentially methylated X chromosome or autosomal CpG sites. We also identified a total of 396 DMRs between sexes (Supplementary Table 3). All of these were located on the X chromosome with 311 and 85 being hypermethylated in females and males, respectively. The 85 hypermethylated DMRs on the X chromosome in males did not overlap with any reported sex-biased genes [44], with the exception of *XIST*, which is upregulated in females and is well known for its role in XCI in female mammalian somatic cells [45]. The most statistically significant differentially methylated autosomal CpG site as defined by the empirical Bayesian variance method in limma between sexes was hypermethylated in females and was within exon 1 of *TLE1*, which is a marker of synovial sarcoma [46]. Although *TLE1* is expressed in the placenta, it has not been reported to be differentially expressed between fetal sexes in the placenta [44,47].

Gestational age calculator training dataset

Although a multitissue age predictor using DNA methylation data has been previously developed [5], we set out to determine if DNA methylation can be used to predict the gestational age of the placenta. To develop our placental gestational age calculator, only placentas from healthy singleton pregnancies with individual gestational age information were included. Placental tissue samples from PE pregnancies were excluded from the training dataset to reduce potential confounding factors caused by the disease. In total, we used 170 placental tissue samples that had individual gestational age information to generate the gestational age calculator. The 170 placental tissue samples were taken from five publicly available datasets, along with our own generated data analyzing 22 term placentas (Table 1). Four of the six datasets were obtained using the Infinium HumanMethylation27 BeadChip and the other two on the Infinium HumanMethylation450 BeadChip, and the included datasets contained placental tissue samples that spanned 8–42 weeks gestation. We selected probes that were present in all six datasets and removed probes that were found on sex chromosomes, leaving a total of 18,437 probes (no missing data). We randomly assigned half of the 170 placental tissue samples to a training dataset, leaving the other 85 samples for validation.

Identifying & validating the gestational age calculator

Briefly, we first normalized the training data using the modified version of the BMIQ [5,34]. The mean β-value of each probe in the largest dataset (GSE3829) was used as the gold standard of the probes, similarly as previously described [5], in the normalization step (Supplementary Table 4). A gold standard of the probes was used for normalization since the datasets were from two different microarray platforms and it rescaled the probes that were present in both microarray platforms. After normalization, we regressed the chronological gestational age against the 18,437 CpG sites using an elastic net penalized regression model [40]. The model automatically selected 62 CpG sites (Supplementary Table 5) to predict the gestational age of a placenta. In the training dataset we found an extremely high correlation ($r = 0.99$, $p < 2.2e\text{-}16$) between the chronological and the predicted gestation age (Figure 1A). In addition, the median absolute difference between the predicted and chronological gesta-

tional age in the training dataset was found to be 0.23 weeks. We then tested these 62 CpG sites in the validation dataset (Figure 1B) and also found a high correlation between the chronological and predicted gestational age ($r = 0.95$, $p < 2.2e\text{-}16$). The median absolute difference in the validation dataset was 1.47 weeks and the root mean square error was 2.3. The heatmap (Figure 1C) allows visualization of the CpG sites and shows changes in DNA methylation across gestation. Furthermore, the lack of vertical lines in the heatmap suggests that the CpG sites are robust against dataset effects. In order to further validate the 62 CpG sites, we predicted the gestational age of all remaining publicly available placental tissue samples from uncomplicated pregnancies that did not have individual gestational age information (Supplementary Table 6). Although these samples did not have individual gestational age information, we sought to determine if the predicted gestational age matched the labeled trimester of pregnancy for each sample. We found a concordance between the predicted gestational age and labeled trimester of pregnancy, which assured us that it is an accurate predictor of gestational age. Here in this study we refer to the predicted gestational age of each placenta as the DNA methylation gestational age (DNAm GA).

The 62 gestational clock CpG sites

The 62 gestational clock CpG sites can be characterized into two groups depending on the direction of their correlation with gestational age. Twenty-seven CpG sites were found to positively correlate and become hypermethylated with increasing gestational age whereas, the other 35 CpG sites negatively correlated and became hypomethylated with increasing gestational age.

Gestational age acceleration in placentas from preeclamptic pregnancies

Since differential DNA methylation occurs in placentas from PE compared with uncomplicated pregnancies [17], we investigated if accurate prediction of gestation age can also be achieved in placentas from PE pregnancies. Two datasets (GSE44667 and GSE59274) contained individual gestational age information for placental tissue samples from PE pregnancies (26 early-onset PE and 18 late-onset PE). We then compared the chronological gestational age with the DNAm GA (Figure 2) and found that placentas from early-onset PE pregnancies (<34 weeks gestation) had a higher DNAm GA compared with their chronological gestational age ($p = 3.44e\text{-}6$, two-tailed, paired t-test). However, late-onset placentas from PE pregnancies (≥34 weeks gestation) did not show any significant difference between their chronological and DNAm GA ($p = 0.38$) indicating that late-onset PE does not affect placental aging. From here on, we refer to the difference between chronological and DNAm GA as gestational age acceleration, similar to what has been defined previously [5]. One dataset (GSE36642) contained placentas from MZ and DZ twins with individual gestational age information, allowing the determination of gestational age acceleration heritability by calculating the broad-sense heritability using Falconer's formula ($H^2 = 2(\text{cor(MZ)-cor(DZ)})$). We conducted our analysis on gestational age acceleration heritability on twin samples of the same sex. The broad-sense heritability was used to determine the proportion of variance of gestational age acceleration as a result of genetic variation. Despite having a small sample size, we calculated the gestational age acceleration for each sample and determined the broad-sense heritability to be 57.2% in MZ and DZ twin pairs (Supplementary Figure 1).

Discussion

In this study, we investigated DNA methylation differences in the placenta across gestation and in different pregnancy outcomes. We found 34 genes that have been reported to be differentially expressed in placenta from PE compared with uncomplicated pregnancies [30] to contain at least one differentially methylated CpG site (Supplementary Table 1). For example, there was a differentially methylated CpG site in the promoter of *RAC1* (Supplementary Table 1), a member of the RAS superfamily [48], which has been found to be upregulated in placenta from preeclamptic pregnancies [30]. We also tested for DMRs in placentas from preeclamptic and uncomplicated pregnancies. We identified three DMRs, which overlapped the 5′ region of *MARC2*, *FAM3B* and *TP53TG1*. However, to our knowledge these genes have not been reported to be differentially expressed in placenta from women with PE.

The placenta has been implicated in a number of pregnancy complications, including preterm birth [49]. Sex differences in pregnancy outcomes have also been reported, for example, women bearing a male fetus are at a 20% higher risk of preterm birth [50–52]. The fetus and placenta are genetically identical [53]. Therefore, it is reasonable to suggest that sex differences in outcomes may potentially be orchestrated by the placenta. Sex differences in placental gene expression have been previously reported [44,47] and in comparison to the placental sex-biased gene expression meta-analysis [44], 20 genes (located on the X chromosome and upregulated in females) were found to contain at least one CpG site or DMR that was hypermethylated in females. However, this may be the result of XCI in females.

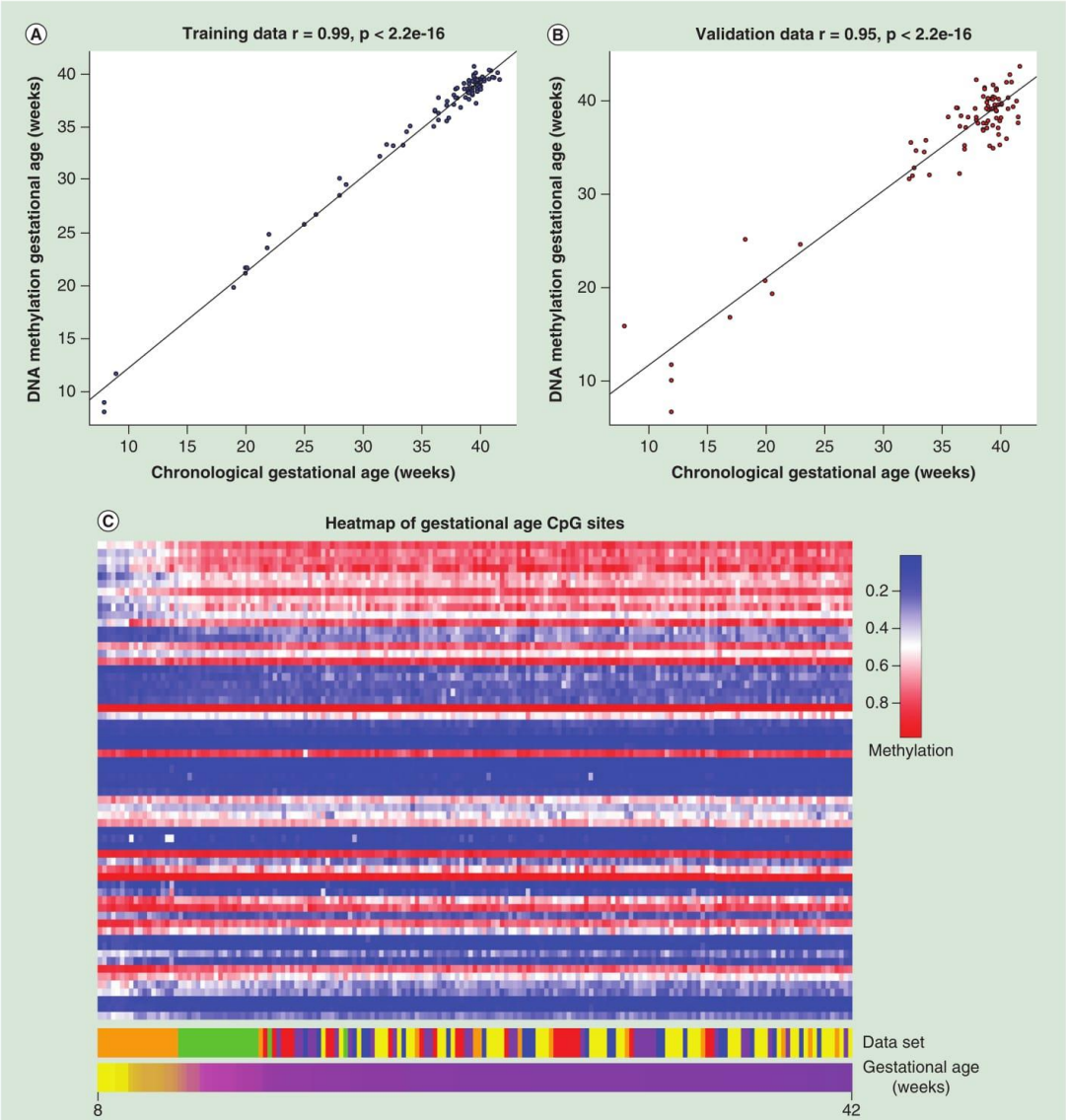


Figure 1. DNA methylation accurately predicts gestational age of the placenta. The correlation between the chronological and DNA methylation gestational age of each placenta in the **(A)** training dataset and the **(B)** validation dataset. **(C)** A heatmap visualizing the gradual changes in DNA methylation in each of the 62 CpG sites (rows) across gestation in all samples (columns). The samples have been ordered by increasing gestational age and the probes have been ordered by the increasing magnitude of correlation with gestational age. The dataset heatmap represents the origin of each sample.

It is unclear why 85 X chromosome DMRs would be hypermethylated in males considering XCI occurs in females. Despite being hypermethylated in males, there are no reports of sex differences in expression of the genes in which these DMRs occur. Potentially they may regulate other sex-specific differentially expressed

RNAs including ncRNAs or affect transcription factor binding. Further research is required to elucidate the precise mechanisms by which the 85 X chromosome hypermethylated DMRs in males act.

We therefore, found a lack of overlap with our differential DNA methylation analysis and two gene expression meta-analyses [30,44]. These findings were consistent with a previous study in which changes in gene expression and DNA methylation did not overlap from matched placental samples from PE and uncomplicated pregnancies [21]. Furthermore, changes in DNA methylation within the promoters of genes do not always alter gene expression as reported for colon cancer [54]. We also found little overlap between genes that have previously been reported to be sex biased in the placenta [44] and our DNA methylation results. Interestingly, we did find 420 differentially methylated autosomal CpG sites and 85 X chromosome hypermethylated DMRs in males. However, due to the poor overlap with gene expression and the effect of XCI it is difficult to draw any conclusions on what the effect of the differences in DNA methylation that were observed had on gene expression.

The lack of overlap between DNA methylation and reported gene expression changes has several limitations. First, in comparison to the two gene expression meta-analyses, we analyzed far fewer samples and therefore may not have been statistically powered to detect small differences. Second, the Illumina DNA Methylation BeadChip arrays only assess approximately 2% of the CpG sites in the human genome and do not assess other methylated cytosine sites such as CHG and CHH. Therefore, we may not have captured the true landscape of DNA methylation in the placenta. Finally, the comparison between DNA methylation and gene expression was not in the same samples and therefore there may have been too much biological variability to detect any overlap. Despite these limitations, our findings suggest that CpG methylation in the placenta may not be a key regulator of gene expression and therefore may be more dependent on other epigenetic factors such as histone modifications, small RNA regulation or ncRNA changes. It has been reported that the placenta with the exception of the brain, has high levels of non-CpG methylation compared with other human tissues [55], which may have a bigger influence on gene expression levels.

In this study, we identified 62 CpG sites, which together can be used to determine the gestational age of a placenta. Several limitations of our study for predicting gestational age do require discussion. First, the training dataset consisted of placentas from 8 to 42 weeks gestation with a bias of samples being from late third trimester. First and second trimester samples comprised only 11 and 9.5% of the training and validation data, respectively. This may have created some biases in the CpG sites chosen and may cause some inaccuracy in identifying the gestational age of placentas from first and second trimester. Second, in relation to gestational age acceleration heritability the twin dataset only contained 14 twin pairs. Therefore we may have not captured the true extent of gestational age heritability within this study. Future studies with large sample sizes of twin pairs are required to determine the true extent of the heritability of gestational age acceleration. In addition to gestational age prediction, we used first and second trimester placentas from terminated pregnancies. One limitation is that some of these placentas may have been from women destined to develop a pregnancy complication, which may have implications for our gestational age prediction. A possible approach to overcome this limitation is to use placental villi from chorionic villus sampling in ongoing pregnancies. Thereby, samples that were from complicated pregnancies could be excluded from the analysis. Unfortunately, to our knowledge there is no publicly available DNA methylation data on such samples that we could use to test our gestational age prediction.

Placentas from early-onset preeclamptic pregnancies were found to have a higher DNAm GA compared with their chronological gestational age. Using a twin dataset we were able to determine the broad-sense heritability of gestational age acceleration to be 57.2%. This finding suggests environmental factors also have an influence together with genetic factors on gestational age acceleration. Maternal lifestyle factors such as smoking [56,57] are known to alter DNA methylation levels in the placenta. Therefore, the maternal environment can affect the intrauterine environment, and thereby could influence gestational age acceleration in the placenta. Furthermore, other lifestyle factors such as BMI have been found to increase the epigenetic age in certain tissues such as the liver [58]. It would therefore be important to investigate the effect of maternal lifestyle factors on gestational age acceleration as they may have implications for pregnancy success. Unfortunately these data are often not recorded for publicly available datasets. This limitation also applies to the exact nature of the twin placentas included in the publicly available data. For example, we do not know if the twin pregnancies had fused placentas or not. Although each individual twin is listed as having a separate placenta, it may be possible that some of the twin pregnancies had fused placentas and therefore the exact sampling sites could confound the data on heritability. Therefore, we suggest some caution in interpreting the heritability of gestational age acceleration analysis as we do not have full clinical details for the publicly available twin dataset (GSE36642).

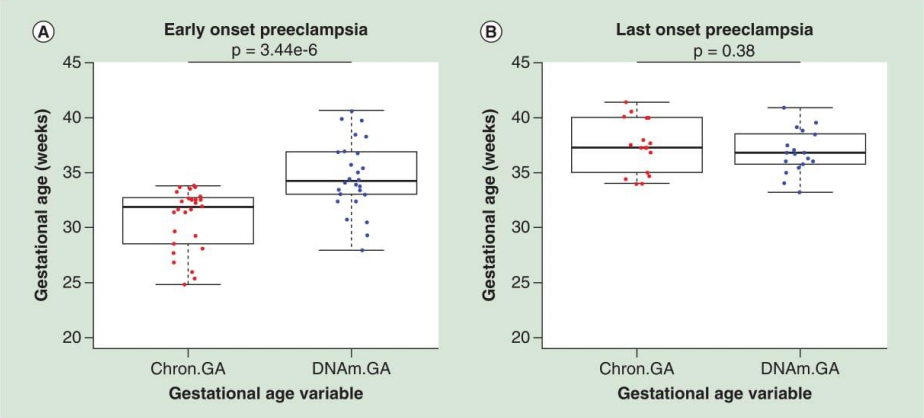


Figure 2. Gestational age acceleration in early-onset preeclampsia. Placentas from pregnancies complicated by early-onset PE **(A)** show a higher DNAm GA compared with their Chron GA. **(B)** Placentas from late-onset PE do not show any difference between the chronological and DNA methylation gestational age. Error bars are represented as standard deviations.
Chron GA: Chronological gestational age; DNAm GA: DNA methylation gestational age.

Using our gestational age tool, we also provide an estimation of the gestational age of all remaining publicly available placenta samples from uncomplicated pregnancies for which gestational age information is not recorded as a resource to the scientific community (Supplementary Table 6). It has also been reported that the placental transcriptome is clearly distinct in PE compared with other pregnancy complications [31]. It would therefore be of interest to determine if the placental gestational age acceleration observed in early-onset PE pregnancies also occurs in other pregnancy complications. Likewise, the gestational age acceleration observed in placentas from early- but not late-onset PE highlights potential differences in the etiology of the two diseases. Furthermore, gestational age acceleration may potentially reveal potential mechanisms in the development of early-onset PE and other pregnancy complications. First trimester chorionic villus samples could potentially be used to determine when accelerated placental aging first occurs. This may provide insight into the mechanisms and the association of placental aging and pregnancy complications such as early-onset PE. However, chorionic villus sampling is only used in some high-risk pregnancies but has a miscarriage risk, so tissue availability is limited.

Conclusion

In conclusion, we have identified 62 CpG sites that can be used to determine the DNAm GA of a placenta. Furthermore, we found evidence of placental aging in placentas from early-onset PE. Future studies are required to determine if gestational age acceleration is unique to early-onset PE or is common to other pregnancy complications. In addition, future research should also determine if gestational age acceleration or placental aging could be detected perhaps in maternal blood early in pregnancy in women who are destined to develop a pregnancy complication. Although, we found little overlap between DNA methylation and gene expression changes, further studies involving matched samples are required to confirm these findings.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/full/10.2217/epi-2016-0103

Author contributions

BT Mayne designed, conducted the study, analyzed and interpreted the data, and wrote the manuscript. SY Leemaqz analyzed the data, provided intellectual input into the manuscript. AK Smith was involved in data creation, provided critical discussion and intellectual input into the manuscript. J Breen, CT Roberts and T Bianco-Miotto were all involved in the study design, provided critical discussion and intellectual input into the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the generosity of all the individuals who were involved in the data creation of all datasets that were available for public analysis.

Financial & competing interests disclosure

This project was funded in part by a National Health and Medical Research Council of Australia Project Grant (GNT1059120) awarded to CT Roberts and T Bianco-Miotto, and support from the NIH R01MD009064 awarded to AK Smith. BT Mayne is supported by an Australian Postgraduate Award. CT Roberts is supported by an National Health and Medical Research Council of Australia Senior Research Fellowship GNT1020749. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Aim

- Preeclampsia is characterized by high-maternal blood pressure and proteinuria.
- Differential placental gene expression is known to occur between preeclamptic and uncomplicated pregnancies.
- Overall CpG methylation in the placenta has been shown to increase across gestation.
- We sought to investigate DNA methylation changes in the placenta across gestation and pregnancy complications by using publicly available data.

Results

- We identified a total of 741 differentially methylated CpG sites in the placenta between preeclampsia and uncomplicated pregnancies.
- We developed an accurate tool using 62 CpG sites to predict the gestational age of a placenta.
- Placentas from early-onset preeclampsia had a higher predicted gestational age compared with their chronological age.

Conclusion

- Placentas from early-onset preeclampsia pregnancies are associated with placental aging.
- Our gestational age prediction tool may offer important insight into the molecular mechanisms associated with pregnancy complications.

References

1 Sharp AJ, Stathaki E, Migliavacca E *et al.* DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.* 21(10), 1592–1600 (2011).

2 Novak P, Stampfer MR, Munoz-Rodriguez JL *et al.* Cell-type specific DNA methylation patterns define human breast cellular identity. *PLoS ONE* 7(12), e52299 (2012).

3 Paulsen M, Ferguson-Smith AC. DNA methylation in genomic imprinting, development, and disease. *J. Pathol.* 195(1), 97–110 (2001).

4 Robertson KD. DNA methylation and human disease. *Nat. Rev. Genet.* 6(8), 597–610 (2005).

5 Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 14(10), R115 (2013).

6 Simpkin AJ, Hemani G, Suderman M *et al.* Prenatal and early life influences on epigenetic age in children: a study of mother-offspring pairs from two cohort studies. *Hum. Mol. Genet.* 25(1), 191–201 (2016).

7 Hon GC, Hawkins RD, Caballero OL *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* 22(2), 246–258 (2012).

8 Berman BP, Weisenberger DJ, Aman JF *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* 44(1), 40–46 (2012).

9 Lister R, Pelizzola M, Dowen RH *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271), 315–322 (2009).

10 Schroeder DI, Blair JD, Lott P *et al.* The human placenta methylome. *Proc. Natl Acad. Sci. USA* 110(15), 6037–6042 (2013).

11 Schultz MD, He Y, Whitaker JW *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* 523(7559), 212–216 (2015).

12 Novakovic B, Yuen RK, Gordon L *et al.* Evidence for widespread changes in promoter methylation profile in human placenta in response to increasing gestational age and environmental/stochastic factors. *BMC Genomics* 12 529 (2011).

13 Redman CW, Sargent IL, Staff AC. IFPA Senior Award Lecture: making sense of pre-eclampsia – two placental causes of preeclampsia? *Placenta* 35(Suppl.) S20–S25 (2014).

14 Williams D, Craft N. Pre-eclampsia. *BMJ* 345, e4437 (2012).

15 Monk C, Spicer J, Champagne FA. Linking prenatal maternal adversity to developmental outcomes in infants:

the role of epigenetic pathways. *Dev. Psychopathol.* 24(4), 1361–1376 (2012).

16 Novakovic B, Saffery R. The ever growing complexity of placental epigenetics – role in adverse pregnancy outcomes and fetal programming. *Placenta* 33(12), 959–970 (2012).

17 Bianco-Miotto T, Mayne B, Buckberry S, Breen J, Rodriguez Lopez C, Roberts C. Recent progress towards understanding the role of DNA methylation in human placental development. *Reproduction* 152(1), R23–R30 (2016).

18 Liu L, Zhang X, Rong C *et al.* Distinct DNA methylomes of human placentas between pre-eclampsia and gestational diabetes mellitus. *Cell. Physiol. Biochem.* 34(6), 1877–1889 (2014).

19 Anton L, Brown AG, Bartolomei MS, Elovitz MA. Differential methylation of genes associated with cell adhesion in preeclamptic placentas. *PLoS ONE* 9(6), e100148 (2014).

20 Blair JD, Yuen RK, Lim BK, Mcfadden DE, Von Dadelszen P, Robinson WP. Widespread DNA hypomethylation at gene enhancer regions in placentas associated with early-onset pre-eclampsia. *Mol. Hum. Reprod.* 19(10), 697–708 (2013).

21 Chu T, Bunce K, Shaw P *et al.* Comprehensive analysis of preeclampsia-associated DNA methylation in the placenta. *PLoS ONE* 9(9), e107318 (2014).

22 Finer S, Mathews C, Lowe R *et al.* Maternal gestational diabetes is associated with genome-wide DNA methylation variation in placenta and cord blood of exposed offspring. *Hum. Mol. Genet.* 24(11), 3021–3029 (2015).

23 Ruchat SM, Houde AA, Voisin G *et al.* Gestational diabetes mellitus epigenetically affects genes predominantly involved in metabolic diseases. *Epigenetics* 8(9), 935–943 (2013).

24 Hillman SL, Finer S, Smart MC *et al.* Novel DNA methylation profiles associated with key gene regulation and transcription pathways in blood and placenta of growth-restricted neonates. *Epigenetics* 10(1), 50–61 (2015).

25 Jacob KJ, Robinson WP, Lefebvre L. Beckwith–Wiedemann and Silver–Russell syndromes: opposite developmental imbalances in imprinted regulators of placental function and embryonic growth. *Clin. Genet.* 84(4), 326–334 (2013).

26 Fowden AL, Sibley C, Reik W, Constancia M. Imprinted genes, placental development and fetal growth. *Horm. Res.* 65(Suppl. 3), 50–58 (2006).

27 Yu L, Chen M, Zhao D *et al.* The H19 gene imprinting in normal pregnancy and pre-eclampsia. *Placenta* 30(5), 443–447 (2009).

28 Guo L, Tsai SQ, Hardison NE *et al.* Differentially expressed microRNAs and affected biological pathways revealed by modulated modularity clustering (MMC) analysis of human preeclamptic and IUGR placentas. *Placenta* 34(7), 599–605 (2013).

29 Nishizawa H, Pryor-Koishi K, Kato T, Kowa H, Kurahashi H, Udagawa Y. Microarray analysis of differentially expressed fetal genes in placental tissue derived from early and late onset severe pre-eclampsia. *Placenta* 28(5–6), 487–497 (2007).

30 Van Uitert M, Moerland PD, Enquobahrie DA *et al.* Meta-analysis of placental transcriptome data identifies a novel molecular pathway related to preeclampsia. *PLoS ONE* 10(7), e0132468 (2015).

31 Sober S, Reiman M, Kikas T *et al.* Extensive shift in placental transcriptome profile in preeclampsia and placental origin of adverse pregnancy outcomes. *Sci. Rep.* 5, 13336 (2015).

32 Kaartokallio T, Cervera A, Kyllonen A, Laivuori K. Gene expression profiling of pre-eclamptic placentae by RNA sequencing. *Sci. Rep.* 5, 14107 (2015).

33 Nishizawa H, Ota S, Suzuki M *et al.* Comparative gene expression profiling of placentas from patients with severe pre-eclampsia and unexplained fetal growth restriction. *Reprod. Biol. Endocrinol.* 9, 107 (2011).

34 Teschendorff AE, Marabita F, Lechner M *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29(2), 189–196 (2013).

35 Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28(6), 882–883 (2012).

36 Morris TJ, Butcher LM, Feber A *et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* 30(3), 428–430 (2014).

37 Aryee MJ, Jaffe AE, Corrada-Bravo H *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30(10), 1363–1369 (2014).

38 Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43(7), e47 (2015).

39 Jaffe AE, Murakami P, Lee H *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* 41(1), 200–209 (2012).

40 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33(1), 1–22 (2010).

41 Davis S. IlluminaHumanMethylation27k.db: Illumina Illumina Human Methylation 27k annotation data (chip IlluminaHumanMethylation27k). R package version 1.4.8 (2014).

42 Triche T and Jr. IlluminaHumanMethylation450k.db: Illumina Human Methylation 450k annotation data. R package version 2.0.9 (2014).

43 Barrett T, Wilhite SE, Ledoux P *et al.* NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res.* 41, D991–D995 (2013).

44 Buckberry S, Bianco-Miotto T, Bent SJ, Dekker GA, Roberts CT. Integrative transcriptome meta-analysis reveals widespread sex-biased gene expression at the human fetal–maternal interface. *Mol. Hum. Reprod.* 20(8), 810–819 (2014).

45 Augui S, Nora EP, Heard E. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat. Rev. Genet.* 12(6), 429–442 (2011).

▶46 Terry J, Saito T, Subramanian S *et al.* TLE1 as a diagnostic immunohistochemical marker for synovial sarcoma emerging from gene expression profiling studies. *Am. J. Surg. Pathol.* 31(2), 240–246 (2007).

▶47 Sood R, Zehnder JL, Druzin ML, Brown PO. Gene expression patterns in human placenta. *Proc. Natl Acad. Sci. USA* 103(14), 5478–5483 (2006).

48 Bai Y, Xiang X, Liang C, Shi L. Regulating Rac in the nervous system: molecular function and disease implication of Rac GEFs and GAPs. *Biomed. Res. Int.* 2015, 632450 (2015).

▶49 Roberts CT. IFPA award in placentology lecture: complicated interactions between genes and the environment in placentation, pregnancy outcome and long term health. *Placenta* 31(Suppl.), S47–S53 (2010).

▶50 Clifton VL. Review: sex and the human placenta: mediating differential strategies of fetal growth and survival. *Placenta* 31(Suppl.), S33–S39 (2010).

▶51 Vatten LJ, Skjaerven R. Offspring sex and pregnancy outcome by length of gestation. *Early Hum. Dev.* 76(1), 47–54 (2004).

▶52 Verburg PE, Tucker G, Scheil W, Erwich JJ, Dekker GA, Roberts CT. Sexual dimorphism in adverse pregnancy outcomes – a Retrospective Australian Population Study 1981–2011. *PLoS ONE* 11(7), e0158807 (2016).

▶53 Gude NM, Roberts CT, Kalionis B, King RG. Growth and function of the normal human placenta. *Thromb. Res.* 114(5–6), 397–407 (2004).

▶54 Moarii M, Boeva V, Vert J-P, Reyal F. Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics* 16, 873 (2015).

▶55 Guo JU, Su Y, Shin JH *et al.* Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* 17(2), 215–222 (2014).

▶56 Suter M, Ma J, Harris A *et al.* Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. *Epigenetics* 6(11), 1284–1294 (2011).

▶57 Maccani JZ, Koestler DC, Houseman EA, Marsit CJ, Kelsey KT. Placental DNA methylation alterations associated with maternal tobacco smoking at the RUNX3 gene are also associated with gestational age. *Epigenomics* 5(6), 619–630 (2013).

▶58 Horvath S, Erhart W, Brosch M *et al.* Obesity accelerates epigenetic aging of human liver. *Proc. Natl Acad. Sci. USA* 111(43), 15538–15543 (2014).

Bioconductor software manual: msgbsR: an R package to analyse methylation sensitive genotyping by sequencing (MS-GBS) data

msgbsR: an R package to analyse methylation sensitive genotyping by sequencing (MS-GBS) data

Benjamin Mayne

April 24, 2017

Contents

1 Introduction

Current data analysis tools do not fulfil all experimental designs. For example, GBS experiments using methylation sensitive restriction enzymes (REs), which is also known as methylation sensitive genotyping by sequencing (MS-GBS), is an effective method to identify differentially methylated sites that may not be accessible in other technologies such as microarrays and methyl capture sequencing. However, current data analysis tools do not satisfy the requirements for these types of experimental designs.

Here we present msgbsR, an R package for data analysis of MS-GBS experiments. Read counts and cut sites from a MS-GBS experiment can be read directly into the R environment from a sorted and indexed BAM file(s).

2 Reading data into R

The analysis with the msgbsR pipeline begins with a directory which contains sorted and indexed BAM file(s). msgbsR contains an example data set containing 6 samples from a MS-GBS experiment using the restriction enzyme MspI. In this example the 6 samples are from the prostate of a rat and have been truncated for chromosome 20. 3 of the samples were fed a control diet and the other 3 were fed an experimental high fat diet.

To read in the data directly into the R environment can be done using the rawCounts() function, which requires the directory path to where the sorted and indexed files are located and the desired number of threads to be run (Default = 1).

```
> library(msgbsR)
> library(GenomicRanges)
> library(SummarizedExperiment)
> my_path <- system.file("extdata", package = "msgbsR")
> se <- rawCounts(bamFilepath = my_path)
> dim(assay(se))
```

```
[1] 16047     6
```

The result is an RangedSummarizedExperiment object containing the read counts. The columns are samples and the rows contain the location of each unique cut sites. Each cut site has been given a unique ID (chr:position-position:strand). The cut site IDs can be turned into a GRanges object. Information regarding the samples such as treatment or other groups can be added into the return object as shown below

```
> colData(se) <- DataFrame(Group = c(rep("Control", 3), rep("Experimental", 3)),
+                          row.names = colnames(assay(se)))
```

3 Confirmation of correct cut sites

After the data has been generated into the R environment, the next step is to confirm that the cut sites were the correctly generated sites. In this example, the methylated sensitive restriction enzyme that has been used is MspI which recognizes a 4bp sequence (C/CGG). MspI cuts between the two cytosines when the outside cytosine is methylated.

The first step is to extract the location of the cut sites from se and adjust the cut sites such that the region will cover the recognition sequence of MspI. It is important to note that in this example the user must adjust the region over the cut sites specifically for each strand. In other words although the enzyme cuts at C/CGG on the minus strand this would appear as CCG/G. The code below shows how to adjust the postioining of the cut sites to cover the recginition site on each strand.

```
> cutSites <- rowRanges(se)
> # # Adjust the cut sites to overlap recognition site on each strand
> start(cutSites) <- ifelse(test = strand(cutSites) == '+',
+                           yes = start(cutSites) - 1, no = start(cutSites) - 2)
> end(cutSites) <- ifelse(test = strand(cutSites) == '+',
+                           yes = end(cutSites) + 2, no = end(cutSites) + 1)
```

The object cutSites is a GRanges object that contains the start and end position of the MspI sequence length around the cut sites. These cut sites can now be checked if the sequence matches the MspI sequence.

msgbsR offer two approaches to checking the cut sites. The first approach is to use a BSgenome which can be obtained from Bioconductor. In this example, BSgenome.Rnorvegicus.UCSC.rn6 will be used.

```
> library(BSgenome.Rnorvegicus.UCSC.rn6)
> correctCuts <- checkCuts(cutSites = cutSites, genome = "rn6", seq = "CCGG")
```

If a BSgenome is unavailable for a species of interest, another option to checking the cut sites is to use a fasta file which can be used throught the checkCuts() function.

The correctCuts data object is in the format of a GRanges object and contains the correct sites that contained the recognition sequence. These sites can be kept within se by using the subsetByOverlaps function.

The incorrect MspI cut sites can be filtered out of datCounts:

```
> se <- subsetByOverlaps(se, correctCuts)
> dim(assay(se))

[1] 13983     6
```

se now contains the correct cut sites and can now be used in downstream analyses.

4 Visualization of read counts

Before any further downstream analyses with the data, the user may want to filter out samples that did not generate a sufficient number of read counts or cut sites. The msgbsR package contains a function which plots the total number of read counts against the total number of cut sites produced per sample. The user can also use the function to visulaise if different categories or groups produced varying amount of cut sites or total amount of reads.

To visualize the total number of read counts against the total number of cut sites produced per sample:

```
> plotCounts(se = se, cateogory = "Group")
```

This function generates a plot (Figure 1) where the x axis and y axis represents the total number of reads and the total number of cut sites produced for each sample respectively.

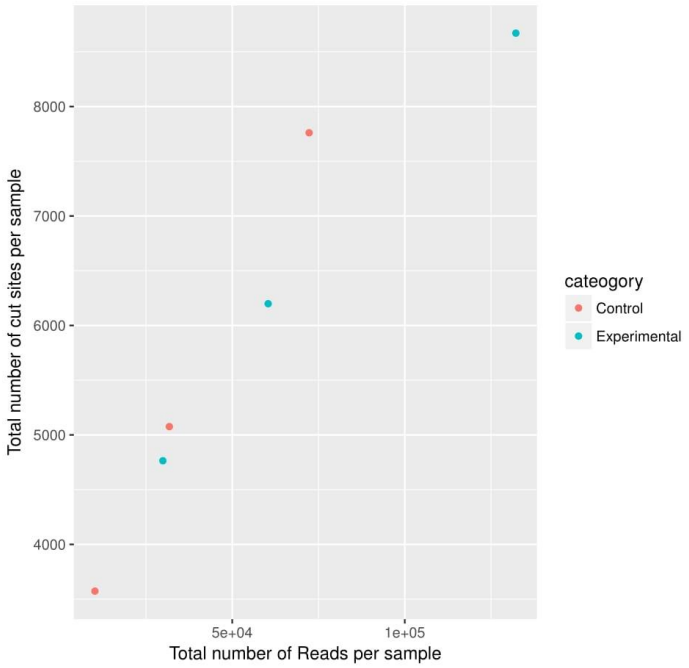


Figure 1: The distribution of the total number of reads and cut sites produced by each sample.

5 Differential methylation analysis

msgbsR utilizes edgeR in order to determine which cut sites are differentially methylated between groups. Since MS-GBS experiments can have multiple groups or conditions msgbsR offers a wrapper function of edgeR (Zhou et al., 2014) tools to automate differential methylation analyses.

To determine which cut sites are differentiallly methylated between groups:

```
> top <- diffMeth(se = se, cateogory = "Group",
+                 condition1 = "Control", condition2 = "Experimental",
+                 cpmThreshold = 1, thresholdSamples = 1)
```

The top object now contains a data frame of the cut sites that had a CPM > 1 in at least 1 sample and which cut sites are differentially methylated between the two groups.

6 Visualization of cut site locations

The msgbsR package contains a function to allow visualization of the location of the cut sites. Given the lengths of the chromosomes the cut sites can be visualized in a circos plot (Figure 2).

Firstly, define the length of the chromosome.

```
> ratChr <- seqlengths(BSgenome.Rnorvegicus.UCSC.rn6)["chr20"]
```

Extract the differentially methylated cut sites.

```
> my_cuts <- GRanges(top$site[which(top$FDR < 0.05)])
```

To generate a circos plot:

```
> plotCircos(cutSites = my_cuts, seqlengths = ratChr,
+            cutSite.colour = "red", seqlengths.colour = "blue")
```

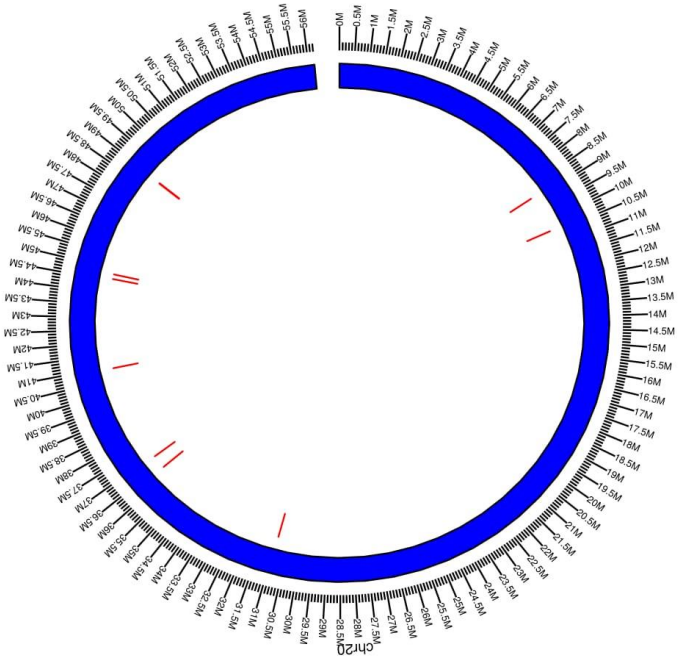


Figure 2: A circos plot of chromosome 20 representing cut sites defined by the user.

7 Session Information

This analysis was conducted on:

```
> sessionInfo()

R version 3.4.0 (2017-04-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.2 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.5-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.5-bioc/R/lib/libRlapack.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats4    stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] BSgenome.Rnorvegicus.UCSC.rn6_1.4.1 BSgenome_1.44.0
 [3] rtracklayer_1.36.0                  Biostrings_2.44.0
 [5] XVector_0.16.0                      SummarizedExperiment_1.6.0
 [7] DelayedArray_0.2.0                  matrixStats_0.52.2
 [9] Biobase_2.36.0                      msgbsR_1.0.0
[11] GenomicRanges_1.28.0                GenomeInfoDb_1.12.0
[13] IRanges_2.10.0                      S4Vectors_0.14.0
[15] BiocGenerics_0.22.0

loaded via a namespace (and not attached):
 [1] ProtGenerics_1.8.0            bitops_1.0-6
 [3] RColorBrewer_1.1-2            httr_1.2.1
 [5] tools_3.4.0                   backports_1.0.5
 [7] R6_2.2.0                      rpart_4.1-11
 [9] Hmisc_4.0-2                   DBI_0.6-1
[11] lazyeval_0.2.0                colorspace_1.3-2
[13] nnet_7.3-12                   gridExtra_2.2.1
[15] GGally_1.3.0                  compiler_3.4.0
[17] graph_1.54.0                  htmlTable_1.9
[19] labeling_0.3                  ggbio_1.24.0
```

```
[21] scales_0.4.1                checkmate_1.8.2
[23] genefilter_1.58.0           RBGL_1.52.0
[25] DESeq_1.28.0                stringr_1.2.0
[27] digest_0.6.12               Rsamtools_1.28.0
[29] foreign_0.8-67              genomeIntervals_1.32.0
[31] R.utils_2.5.0               base64enc_0.1-3
[33] dichromat_2.0-0             htmltools_0.3.5
[35] ensembldb_2.0.0             limma_3.32.0
[37] htmlwidgets_0.8             easyRNASeq_2.12.0
[39] RSQLite_1.1-2               BiocInstaller_1.26.0
[41] shiny_1.0.2                 hwriter_1.3.2
[43] BiocParallel_1.10.0         acepack_1.4.1
[45] R.oo_1.21.0                 VariantAnnotation_1.22.0
[47] RCurl_1.95-4.8              magrittr_1.5
[49] GenomeInfoDbData_0.99.0     Formula_1.2-1
[51] Matrix_1.2-9                Rcpp_0.12.10
[53] munsell_0.4.3               R.methodsS3_1.7.1
[55] stringi_1.1.5               yaml_2.1.14
[57] edgeR_3.18.0                zlibbioc_1.22.0
[59] plyr_1.8.4                  AnnotationHub_2.8.0
[61] grid_3.4.0                  lattice_0.20-35
[63] splines_3.4.0               GenomicFeatures_1.28.0
[65] annotate_1.54.0             locfit_1.5-9.1
[67] knitr_1.15.1                geneplotter_1.54.0
[69] reshape2_1.4.2              biomaRt_2.32.0
[71] XML_3.98-1.6                ShortRead_1.34.0
[73] biovizBase_1.24.0           latticeExtra_0.6-28
[75] data.table_1.10.4           httpuv_1.3.3
[77] gtable_0.2.0                reshape_0.8.6
[79] ggplot2_2.2.1               mime_0.5
[81] xtable_1.8-2                AnnotationFilter_1.0.0
[83] survival_2.41-3             OrganismDbi_1.18.0
[85] tibble_1.3.0                intervals_0.15.1
[87] GenomicAlignments_1.12.0    AnnotationDbi_1.38.0
[89] memoise_1.1.0               cluster_2.0.6
[91] LSD_3.0                     interactiveDisplayBase_1.14.0
```

8 References

Zhou X, Lindsay H, Robinson MD (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Research, 42(11), e91.