Using “Omics” to Discover Predictive Biomarkers in Women at High Risk of Spontaneous Preterm Birth

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Angharad Care

October 2019

Contents

ABSTRACT

Using "Omics" to Discover Predictive Biomarkers in Women at High Risk of Spontaneous Preterm Birth
Angharad Care

Spontaneous preterm birth (sPTB) is a complex pregnancy syndrome that remains poorly understood and is associated with significant perinatal morbidity and mortality worldwide. Current research suggests that there are multiple disordered physiological processes that trigger a final common pathway of early labour, rather than a single specific cause. It is this heterogeneity that has hindered the discovery of a single predictive biomarker and existing screening methods for sPTB prediction are insufficient to detect all women at risk. Consequently, our inability to identify women at risk inhibits efforts of prevention, which cannot be achieved without better understanding of causation or a more robust way of accurately discriminating those at high risk.

The development in "omics" technology has led to exciting breakthroughs in other areas of medicine and offers new avenues of investigation for sPTB prediction. The primary aim of the thesis was to establish a way of combining different types of 'omics' analysis from the same individual in a pilot study to identify candidate biomarker predictors or pathways.

Three different "omic" methodologies; genomics, transcriptomics and metabolomics, were used to analyse blood taken from asymptomatic women high-risk for sPTB at 16 and 20 weeks of pregnancy. Lastly, I investigated if there are distinct differences in biomarkers between PPROM and sPTB subgroups of spontaneous preterm birth. On an individual omics level only transcriptomics showed an association with sPTB. Gene set enrichment in this population demonstrates that the selenoamino acid pathway differentiates asymptomatic high-risk women. Hierarchical clustering in a non-linear distance matrix differentiated all but one of the sPTB and PPROM cases. More studies are required to validate the findings from our analysis.

Data from each omics discipline was combined together in a single data matrix and machine learning analyses applied. The area under the curve (AUC) of receiver operating characteristic (ROC) values for Linear discriminant analysis (0.90), Genetic expression programming (0.70), K-Means (1.00), Linear support vector machine (0.96), Support vector machine with a Gaussian Kernel (0.96), Probabilistic neural network (1.00) and Random Forest (0.96) demonstrate that most machine learning methods perform well on our dataset. Sample sizes needed to reach excellent (AUC = 0.9) vs. moderate (AUC = 0.7) prediction performance were found to be within realistic ranges.

This study provides a conceptual analytical framework for the prediction of sPTB. For a larger cohort prediction power is excellent, making individualized preterm prediction a realistic possibility.

ACKNOWLEDGEMENTS

It would not have been possible for me to write this thesis without the generous support of Lord and Lady Harris and the Wellbeing of Women charity who provided the funding for my clinical research fellowship.

It is difficult to describe the depths of my gratitude for my primary supervisor, Professor Zarko Alfirevic. You have been a tremendous mentor to me over the years. You have taught me more than I could ever give you credit for here, but particularly on how to conduct myself as a clinical scientist, doctor and teacher. Your ability to respond to a query, turn a draft around overnight or make the impossible possible will never fail to amaze me. Thank you for the support and advice on both my research and career; I am forever in your debt.

I am extremely fortunate to have not just one but three great professors as supervisors. Professor Ana Alfirevic and Professor Bertram Müller-Myhsok, please accept my deepest thanks for providing me with guidance and feedback over the last three years. You constantly challenged me to maintain the highest standards of scientific rigour with a warmth, kindness and humour that were always appreciated.

I would like to thank Dr Eunice Zhang, Dr. Till Andlauer, Juhi Gupta, Dr Marie Phelan and Professor Lu-Yun Lian for their time, assistance and teaching to help me complete analyses in GWAS, gene expression analysis and nuclear magnetic resonance for this thesis.

Huge thanks to my colleagues and friends who have assisted me with the arduous process of recruitment, trawled through patients notes to corroborate the phenotype classification system and still provided pep talks when my motivation was waning; particularly Dr Andrew Sharp, Dr Laura Goodfellow, Dr Borna Poljak and Dr Silvia Mammarella.

Thank you to the laboratory staff at both the Liverpool Women's and the Wolfson Centre for Personalised Medicine for their help and teaching, in particular Dr. Jane Harrold, who would always go above and beyond to help.

I would like to express my deepest love and gratitude to my parents, Graham and Mallt. Your thirst for learning and positive attitude to hard work have undoubtedly influenced my life choices, leading me to research. Although you have not been involved in the writing of my thesis, I am absolutely certain it would not have been completed without your constant belief in me and, more importantly, the hours of babysitting provided without complaint.

Thank you to my wonderful husband, Shannon, and our beautiful baby girl, Eleri. Without your love, support and joy, I would have almost certainly finished writing this thesis much faster. I love you both deeply.

A heartfelt thank you to all the women that participated in this research who truly understand the reality of 'spontaneous preterm birth'. I hope this thesis is another step in the journey to help prevent others being born too soon.

ABBREVIATIONS

ADHD – Attention Deficit Hyperactivity Disorder

AF – Amniotic Fluid

ANOVA – Analysis of variance

AUC – Area Under the Curve

BBC – Boston Birth Cohort

BMI – Body Mass Index

BV – Bacterial Vaginosis

CGR – Centre for Genomic Research

CI – Confidence Interval

CIN – Cervical intraepithelial neoplasia

CPMG – Carr-Purcell-Meiboom-Gill

CRH – Corticotrophin releasing hormone

CS – Caesarean section

CST – Community State Types

dbGaP - database of Genotypes and Phenotypes

DNA – Deoxyribonucleic acid

ELISA - Enzyme linked immunosorbent assay

fFN – Fetal Fibronectin

GEP – Genetic Expression Programming

GSEA – Gene Set Enrichment Analysis

GWAS – Genome Wide Association Study

HDMB – Human Metabolome Database

HPA – Hypothalamic Pituitary Adrenal

HRC – Haplotype Reference Consortium

HWE – Hardy Weinberg Equilibrium

IGFBP1 - Insulin-like growth factor binding protein-1

IVF – In vitro fertilization

IVT – In vitro transcription

FDR – False discovery rate

LLETZ – Large loop excision of the transformation zone

MAF – Mean Allele Frequency

MIAC – Microbial Invasion of the Amniotic Cavity

mQTL – methylation-quantitative trait loci

mtDNA – mitochondrial DNA

NEC – Necrotising enterocolitis

NGS – Next Generation Sequencing

NMR – Nuclear magnetic resonance

OPLS-DA - Orthogonal projections to latent structures-discriminant analysis

OR – Odds ratio

OTR – Oxytocin receptor

PAMG-1 - Placental alpha microglobulin 1

phIGFBP1 – Phosphorylated insulin-like growth factor binding protein-1

PCA – Principal component analysis

PLS-DA – Partial least square – discriminant analysis

ppm – parts per million

PPROM – Preterm prelabour rupture of the membranes

PPV – Positive predictive value

PQN – Probabalistic quotient normalization

PTB – Preterm birth

PTL – Preterm labour

PVL – Periventricular leukomalacia

QC – Quality Control

RDS – Respiratory distress syndrome

RFM – Reduced fetal movements

RNA – Ribonucleic acid

ROC – Receiver Operating Characteristic

ROS – Reactive oxygen species

SNP – Single Nucleotide Polymorphism

sPTB – Spontaneous preterm birth

sPTL – Spontaneous preterm labour

sPTL-IM – Spontaneous preterm labour with intact membranes

TNF – Tumour necrosis factor

WES – Whole Exome Sequencing

WGS – Whole Genome Sequencing

XHE - X chromosome homozygosity estimate

17-OHP – 17 α hydroxyprogesterone caproate

Chapter 1: Introduction

1.1 Epidemiology of Preterm Birth

Definition

Preterm birth (PTB) is usually defined as delivery at any gestation before 37 completed weeks of pregnancy ($<37^{+0}$ weeks, <259 days) (Spong et al. 2013). The lower limit of preterm birth and upper limit of late spontaneous miscarriage are blurred as the limit of viability varies with differences in healthcare settings. The World Health Organisation recommends using 28 weeks completed gestation as a cut off for viability, whereas neonates reaching 23 completed weeks have been successfully resuscitated in the UK. If the baby is born before compatibility with life, spontaneous delivery is termed "miscarriage". Despite the difference in semantics, both late spontaneous miscarriage and early spontaneous birth are considered to share the same pathophysiological triggers.

Incidence

Worldwide, an estimated 15 million babies are born before 37 completed weeks of pregnancy (Howson et al. 2012). Across 184 countries, the rate of PTB ranges from 5% to 18% and in nearly all countries reporting reliable data, the rates of prematurity are increasing (Blencowe et al. 2013, Chawanpaiboon et al. 2019). Being born preterm results in insufficient time in utero for complete organ maturation and three quarters of all perinatal mortality and over half of long-term morbidity is attributable to PTB (Goldenberg et al. 2008). Many survivors face a lifetime of disability, including cognitive impairment, motor disability, poor respiratory health, behavioural disturbance and visual and hearing deficiency. The severity of these risks is inversely proportional to the gestational age at birth. (Figure 1.1).

Complications of prematurity are the leading cause of death among children under five years of age and were responsible for nearly one million deaths worldwide in 2013 (Blencowe et al. 2013). The emotional and economic implications of PTB remain a burden to healthcare systems and societies. Parents of neurologically disabled children report social exclusion from parents with normal children, anxiety, relationship breakdown, reduced quality of life and ending careers to become carers (McCormick. 1985). One UK study (Mangham et al. 2009) estimated the total cost to the public sector for the care of children born prematurely up to 18 years old to be £2.95 billion annually.

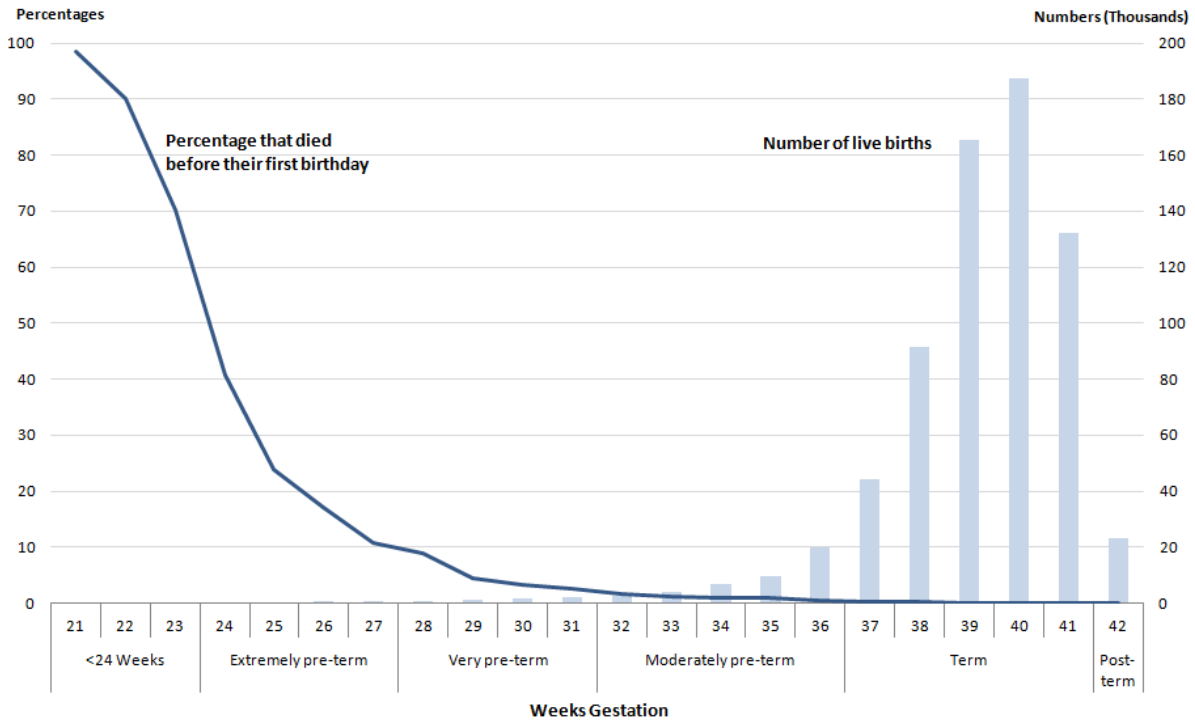


Figure 1.1 Percentage of infant deaths and number of live births per week of gestation in England and Wales. Data from Office of National Statistics. Pregnancy and ethnic factors influencing births and infant mortality: 2013-2015

Classification

As neonatal morbidity and mortality are inversely proportional to the gestational age of birth, preterm birth is often sub classified into:

- Late preterm (34^{+0} to 36^{+6} weeks)
- Moderate preterm (32^{+0} to $<34^{+0}$ weeks)
- Very preterm (28^{+0} to $<32^{+0}$weeks)
- Extremely preterm ($<28^{+0}$ weeks)

This type of classification groups neonates by gestation specific morbidity and mortality risks, helping us to discuss prognosis in general terms. However, a classification based on gestation tells us little about the phenotype or cause of these births. Preterm birth is the only pathology defined by a specific time point rather than a common collection of signs and symptoms. There are multiple causes and pathologies under this umbrella term, therefore preterm birth can also be classified based on phenotype:

1. Medically indicated or "iatrogenic" preterm birth (30%)
2. Spontaneous preterm labour with intact membranes (sPTL-IM) (45%)
3. Premature prelabour rupture of the membranes (PPROM) (25%)

   followed by either i) medically indicated delivery or ii) spontaneous labour

Medically indicated, iatrogenic or elective preterm births aim to prevent severe maternal or fetal morbidity and mortality from conditions such as pre-eclampsia, intrauterine growth restriction, fetal distress, placental abruption or maternal medical disease. The remaining 70% are due to a spontaneous onset of labour with regular uterine contractions and progressive dilatation of the cervix or secondary to spontaneous rupture of the membranes that is not immediately followed

by regular uterine contractions and labour (Goldenberg et al. 2008). Pregnancies affected by PPROM frequently labour spontaneously, but it is difficult to predict at what gestation in the future labour will occur; from hours to months later. If there are concerns regarding infection or fetal wellbeing, labour may be induced before it can begin spontaneously. In studies of preventative treatment sPTL-IM and PPROM have frequently been lumped together as "spontaneous preterm birth". Spontaneous PTB (sPTB) is currently viewed as a "syndrome" rather than a disease entity encompassing multiple disease mechanisms into a final common pathway of delivery (Villar et al. 2012).

Clinically, PPROM is generally defined as the onset of amniotic fluid (AF) leakage from the vagina prior to the onset of labour at less than 37 weeks of gestation. However, there is no universally accepted classification of PPROM for use in the research studies. Difficulty arises trying to ascertain a timepoint at which labour commences that is measurable and objective, particularly as women can report irregular symptoms of labour prior to membrane rupture. Therefore, a challenge arises in appropriately classifying women with PPROM from those classified as spontaneous preterm labour (sPTL). Delivery of the baby is an objective event and better recorded than the onset of labour. Some studies assign an arbitrary time from rupture of membranes to birth beyond which a spontaneous preterm birth would be classified as PPROM and not sPTL-IM. Definitions from several published studies of PPROM are shown in Table 1.1. There is a range of inclusion criteria and time specifications depending on what is being studied. Only one study listed here (Hadley et al. 2017) has included criteria for indicated or sPTB classification. Additionally, the diagnosis of ruptured membranes can in itself be challenging for clinicians. Additional bedside testing for the presence of AF in the posterior fornix of

the vagina can assist in making the diagnosis if the clinical history is unclear, but like all clinical tests can be subject to false positives and negatives.

Table 1.1 Inclusion criteria for PPROM cases in research studies published in 2017/8

Author	Year	Sterile Speculum	Clinical Test	Min. time ROM to Labour Onset	Min. Time ROM to Delivery	Other
Sak *et al.*	2017	✓	✓	NS	NS	
Zhang *et al*	2017	NS	NS	NS	NS	
Sung *et al*	2017	✓	✓	48 hours	NS	
Dundar *et al*	2017	✓	✓	NS	NS	
Vanderbroucke *et al*	2017	✓	✓	NS	72 hours	
Hadley *et al*	2017	✓	✓	NS	>2 hours	Did not meet criteria for indicated PTB or PTL (6 contractions an hour or equal to 4cm dilated)
Toprak *et al*	2017	✓	✓	NS	NS	
Musilova *et al*	2017	✓	✓	NS	NS	
Roberts *et al*	2017	NS	NS	NS	NS	Diagnosis made by attending medical practitioner
Shree *et al*	2017	NS	NS	NS	>12 hours	
Hromadnikova *et al*	2017	✓	✓	2 hours	NS	
Radochova *et al*	2017	✓	✓	NS	NS	
Patel et al	2017	✓	✓	NS	NS	
Pharande et al	2017	NS	NS	NS	>24 hours	Clinical diagnosis
Wang et al	2018	✓	✓	NS	NS	pH strip, AF crystallization and sICAM-1 positive

Risk Factors

The triggers of spontaneous labour, both at term and preterm, are still poorly understood – and a precise mechanism is not established in most cases. Therefore, factors associated with preterm birth have been sought to try and identify the most at-risk populations. Women with a previous preterm birth have a recurrence risk of 15% to 50% depending on the number, gestational age and characteristics of previous deliveries (Goldenberg et al. 2008). The risk of a recurrent preterm birth is inversely related to the gestational age at the first delivery (Kazamier et al. 2014). Most risk factors cannot be altered between pregnancies such as genetic influences or uterine abnormalities. Even modifiable risk factors such as social status and body mass index (BMI) are difficult to change. Identifying at-risk women for spontaneous preterm birth in their first pregnancy is particularly challenging. Table 1.2 lists some of the common risk factors associated with spontaneous preterm birth.

Table 1.2 Risk factors for spontaneous preterm birth

Risk Factors for Spontaneous Preterm Birth	
Medical/Obstetric History	Previous Preterm Birth
	Anatomical abnormalities of the uterus
	Conceiving through *In Vitro Fertilisation* (IVF)
	Thrombophilia
	Chronic medical conditions such as diabetes and high blood pressure
	Excisional cervical surgery (e.g. knife cone biopsy, LLETZ)
	Other cervical damage – in previous delivery, recurrent second trimester surgical terminations
	Family history of spontaneous PTB (maternal side only)
Maternal	Extremely Low (<19) Body Mass Index (BMI)
	Short Inter Pregnancy Interval (<6 months)
	Higher social deprivation
	Smoking
	Increasing age
	Domestic Violence
Fetal	Congenital Abnormality
	Chromosomal Abnormality
Current Pregnancy	Short cervical length for gestational age
	Positive fetal fibronectin between 22 and 34 weeks
	Certain congenital abnormalities of the fetus
	Vaginal bleeding in pregnancy
	Infections – urinary tract, sexually transmitted, bacterial vaginosis, periodontal disease
	Overdistension of the uterus – multiple pregnancy, polyhydramnios, macrosomia
	Recurrent Antepartum haemorrhage
	Pre-eclampsia, uteroplacental insufficiency

Neonatal Outcomes following Preterm Birth

Two UK cohort studies comparing outcomes of babies born between 22 and 26 weeks in 1995 and 2006 show survival rates of extremely premature infants have been improving (40% to 53%) (Costeloe et al. 2012). Overall, more babies are now being admitted for care at earlier gestations, although healthy survivor numbers are increasing so are the total number of neonates with moderate or severe disability. The Epicure 2 study showed that in infants born before 26 weeks in 2006 approximately 44% will survive to three years of age and of those 15% will have a severe disability. Neonatal clinical networks have now increased centralisation of care for babies born less than 26 weeks as survival is greatest in hospitals that can provide neonatal intensive care (Level 1 service) (Marlow et al. 2014).

Short Term Morbidity

Cells in the lung alveoli (Type 2 pneumocytes) begin to produce surfactant from 30 weeks of gestational age decreasing the surface tension within the alveoli. Babies born below this gestation are at highest risk of respiratory distress syndrome (RDS). Although rates of RDS have decreased over the last few decades due to the use of antenatal corticosteroids, increased use of surfactant and improvements in lung ventilation, some preterm neonates treated with oxygen and positive-pressure ventilation will ultimately develop bronchopulmonary dysplasia (BPD), a chronic lung injury.

Preterm neonates are particularly susceptible to intraventricular haemorrhage (IVH) due to the high levels of vascularisation that is occurring during brain development before 34 weeks. Severe haemorrhage predisposes the child to impairment of cognitive, motor and visual functions. Periventricular leukomalacia (PVL) is a white matter brain injury that has long term sequelae including cerebral

palsy and a low IQ (Back et al. 2007). Even in the late preterm birth group between 32-36 weeks there are increased levels of autism, attention deficit hyperactivity disorder (ADHD) and school difficulties when compared with children born at term (Guy et al. 2015).

The pathogenesis leading to necrotising enterocolitis (NEC) remains unknown. It is a gastrointestinal disorder leading to ischemic injury and abnormal bacterial colonisation. In preterm infants NEC usually presents after the commencement of feeds and may appear after two to three weeks of life once preterm babies have survived the early neonatal period. The mortality for NEC can be as high as 50% and operative intervention is necessary in almost 20% to 40% of cases (Yee et al. 2015).

Long Term Morbidity

Studies examining long term outcomes for preterm babies show the same inverse relationship with gestational age for both morbidity and mortality. In the UK, 39% of deaths under the age of five years are directly caused by prematurity (WHO. 2015). A risk of ill health during childhood exponentially increases for very preterm and moderately pre-term (32 to 36 weeks) neonates when compared to term counterparts (Boyle et al. 2012). Long term adverse outcomes include delayed behavioural development at six years of age, decreased lung function at eight to nine years of age, increased hospitalisation, lower exercise capacity into early adulthood and increased risk of poor metabolic and cardiovascular health (Bartha et al. 2012, Lapillonne et al. 2013., Roggero et al. 2013).

1.2 Causes of Spontaneous Preterm Labour

The mechanisms that lead to human term labour, let alone preterm parturition are not yet fully understood. It is, therefore, unsurprising that effective strategies to prevent preterm birth remain inadequate on an individual level. Better identification of the cause may help focus use of the correct preventative interventions or lead to development of more effective treatments. Various pathological processes linked to preterm birth include inflammation triggered by infection, pathological uterine distension (multifetal pregnancy, polyhydramnios, uterine anomalies), cervical insufficiency and fetal and maternal stress.

Inflammation and Infection

The role of inflammation and infection in preterm birth has been recognised for many decades. Infection is frequently associated with preterm labour (PTL) in both humans and animal models. In pregnant mammals, systemic administration of a microbial load can induce PTL (McDuffie et al. 1992). In humans, 25%- 40% of sPTB have evidence of intrauterine infection, particularly in PPROM (Agarwal and Hirsch. 2012). However, in these cases it can be difficult to elucidate if intrauterine infection preceded ruptured membranes or occurred following the loss of the protective membrane barrier. Ascending infection from the genital tract causing inflammation of the lower uterine segment and triggering cervical shortening and labour has been proposed as one mechanism of PTL.

The AF cavity is a sterile environment, therefore positive cultures of AF are considered pathological. The reported rates of positive culture of AF detected in women presenting in PTL with intact membranes is 13%, this is higher than rates of non-labouring preterm patients and term labourers (Goncalves et al. 2002). In women with PPROM this rises to 32% and again to 75% by the time these women

subsequently labour, demonstrating colonisation of microbes both before and during the latent period (Goncalves et al. 2002). A third mechanism of infection unrelated to the cervix is haematological spread through the placenta causing microbial invasion of the amniotic cavity (MIAC) (Goncalves et al. 2002). This route of infection is supported by the fact that extrauterine infection such as asymptomatic bacteriuria and pyelonephritis (Wing et al. 2014), periodontal infection (Parthiban and Mahendra. 2015) and malaria (McDonald et al. 2015) are all associated with an increased risk of preterm birth.

Changes in cervical ripening during labour have been associated with increased production of inflammatory cytokines such as interleukins-1, -6, -8, tumour necrosis factor (TNF) and prostaglandins (Chandiramani et al. 2012, MacIntyre et al. 2012). Influx of inflammatory cells into the cervix release matrix metalloproteins contributing to collagen breakdown and ultimately a softening or ripening of the cervix.

Uterine Stretch

Overdistension over the uterus as a PTB risk has been exemplified clinically by the decrease in the mean age of spontaneous delivery to 35 weeks in twins and 30 weeks in quadruplets. Additionally, women diagnosed with polyhydramnios and unicornuate uterus are also at increased risk of spontaneous PTB. In vitro, induced mechanical stretch of uterine myometrium causes increases in gap junction proteins such as CX 26 and CX 43 (Xu et al. 2013), IL-8 and oxytocin receptor (OTR) expression leading to MAPK pathway activation (Kim et al. 2015), and upregulating COX-2 activity (Sooranna et al. 2004) ultimately terminating with local prostaglandin release causing increased contractility of the uterine muscle.

Additionally, stretch also upregulates calcium signalling in the myometrium leading to muscle contraction (Li et al. 2009).

Maternal / Fetal Stress

Stress can be difficult to quantify and disassociate from other risk factors such as smoking, poor nutrition and low socioeconomic status. Stress in the fetus is thought to arise secondary to abnormal placentation, may present with growth restriction and lead to maturation and activation of the fetal hypothalamic-pituitary-adrenal (HPA) axis. How the HPA axis is exactly activated is not yet understood, but placental corticotrophin-releasing hormone (CRH) is currently thought to play an important role (Gravett et al. 2010). Moreover, maternal stress increases biological effectors, including cortisol and adrenaline which have been postulated to activate placental CRH gene expression (Sandman and Davis. 2012). CRH is a neuropeptide of predominantly hypothalamic origin, it is also expressed in human placenta and membranes and released in increasing amounts over the course of pregnancy (Gravett et al. 2010). The exponential rise of CRH has been associated with the length of gestation (McLean and Smith. 1999). These findings have led some researchers to suggest that placental CRH may act as a "placental clock" and regulate the length of gestation (McLean and Smith. 1999). Therefore, one current theory behind preterm labour is the premature senescence of the placenta leading to an earlier trigger for birth or PPROM. There is some suggestion this may be mediated by imbalances in reactive oxygen species (ROS) causing damage mediated by p38 mitogen activated kinase (p38MAPK) pathways (Polettini et al. 2015).

Cervical Function

Gradual softening and effacement of the cervix occur in the weeks before labour. Cervical ripening involves a breakdown of collagen, changes in proteoglycan

concentration and increase in water content that occur in response to increased local prostaglandin release or partial antagonism to progesterone receptors (i.e. action of mifepristone) (Bennett. 2007).

The role of the cervix in maintaining pregnancy remains undefined and is probably multifactorial, with two key roles i) prevention of ascending infection and ii) physical support to keep the pregnancy in utero.

Maintenance of a healthy mucus plug and adequate length to the cervix may act to prevent ascending infection that triggers production of local inflammatory cytokines and prostaglandin release.

The quality of strength of the cervix to support the pregnancy in situ against gravitational pressure is required to prevent premature cervical dilatation. Recognised cervical weakness also called 'cervical insufficiency', causes women to suffer recurrent mid trimester loss usually with a history of painless dilatation of the cervix. Funnelling is a feature associated with cervical weakness and can be seen on transvaginal ultrasound (TVUSS) as the membranes prolapsing through the endocervical canal of the cervix. Women who have excisional cervical surgery for cervical intraepithelial neoplasia (CIN) or cervical cancer also have an increased risk of spontaneous PTB. This contributes to the argument for the function of the cervix being related to its length. In the treatment of CIN the proportions of the volume/length excised at large loop excision vary substantially, and the latter has been shown to correlate with the pregnancy duration (Kyrgiou et al. 2015).

Vaginal Microbiome

Micro-organisms, particularly pathogenic microbes have long been hypothesised to play a role in the onset of early labour. The recent advances of sequencing approaches have allowed for broad unbiased surveys of bacterial

communities and the discoveries of new bacterial species and different taxa. In a seminal paper by Ravel *et al*. (2011) a study of women of reproductive age showed that individuals could be categorised by the dominant species of lactobacillus within their vaginal microbiome. Lactobacillus species typically dominated more than 90% of their entire community and were divided into "community state types" (CST). The most common species of lactobacilli were *L.crispatus (CST I), L.gasseri (CST II), L.iners (CST III)* and *L.jensenii* (CST V) (Ravel J et al. 2011). Some women were also found to have microbiomes with low levels of lactobacilli and increased diversity of other species such as *Atopobium*, *Gardnerella*, *Prevotella* and *Megasphera (CST IV – diverse group).*

In pregnancy, the vaginal microbiome is characterised by a stable, low richness, low diversity community and is generally dominated by *Lactobacillus* species (Aagaard et al. 2012. McIntyre et al. 2015. Romero et al. 2014). *L.species* are thought to be protective against pathogenic organisms (Reid et al. 2011) and prevent inflammatory shifts in the vaginal environment. Preterm birth has been shown to correlate with a Lactobacillus-poor, high-diversity vaginal community (DiGiulio et al. 2015) with bacterial vaginosis (BV)-taxa such as *Gardnerella* and *Ureaplasma* implicated as causative agents. Empiric treatment of BV to try and prevent sPTB has failed to show benefit across multiple studies and is not currently recommended (Brocklehurst et al. 2013).

As methods to look at individual types of lactobacillus has developed, more recently there is evidence to show that vaginal microbiota dominated specifically by *L.iners* are a risk factor for preterm birth (Petricevic et al. 2014, Kindinger et al. 2017). This has only been consistently demonstrated in Caucasian populations. Studies including predominantly African American and Hispanic pregnancy

populations could not replicate the same findings, as these populations tend towards a normal vaginal microbiome that is lactobacillus-poor and more diverse (Fetweiss et al., 2014, Ravel et al., 2011). Stability of a diverse vaginal microbiome is associated with term delivery whilst a significant decrease of community richness and diversity particularly between the first and second trimesters is associated with preterm birth (Stout et al. 2017). When Caucasian and African American populations have been compared, very few taxa associated with preterm birth were found in both Caucasian and African American populations (Callahan et al. 2017). Only the predominance of *L.crispatus* may be protective and increased proportions of *Prevotella* species may confer risk for sPTB across populations (Callahan et al. 2017).

Interestingly, dominance of *L.crispatus* has also been shown in other studies to be protective against sPTB and reduce early onset neonatal sepsis (Verstraelen et al. 2009, Brown et al. 2018). Brown *et al*. (2018) took temporal samples through pregnancy of the vaginal microbiome in women experiencing PPROM. Only one-third of women with PPROM demonstrated a dysbiotic vaginal microbiome associated with subsequent chorioamnionitis or funisitis, suggesting two-thirds of women with PPROM are likely to have a non-infective cause.

This is a promising area of research in the understanding of sPTB. However, more research needs to be done to understand the differences seen in “normal” vaginal microbiome populations in pregnancy before this knowledge can be made clinically useful in the prevention of sPTB. Exploration is required of the interplay between the vaginal microbiome and genetics to elicit racial population differences and impact of different species strains on the vaginal and cervical epithelium.

Genomics

Genetics is estimated to play a role in up to 40% of PTB (Clausson et al. 2000). In singletons, genetic susceptibility to PTB is based on the evidence of familial aggregation, identification of disease-susceptibility genes and racial disparity in PTB rate that may be related to differences in risk-predisposing allele frequencies. PTB rates are higher in sisters of women with a history of PTB compared to their sisters-in-law (16% vs. 9%). Mothers who were born preterm are more likely to deliver preterm by almost 20% (Porter et al. 1997). This suggests that PTB is inherited in a matrilineal manner across generations and is unlikely to be affected by patterns of PTB in the father's family (Boyd et al. 2009). Several studies have confirmed a two-fold increase in risk of sPTB for black American women compared to white American women, even after controlling for socio-economic factors associated with PTB.

The most commonly studied pathways for potential candidate genes are those involved in infection and inflammation. A recent pathway analysis of published studies of different polymorphisms in 274 genes suggested that there may be different gene pathways for women presenting with sPTB and PPROM (Capece et al. 2014). An autoimmune or hormonal regulation axis may exist for sPTB, whilst pathways implicated in the etiology of PPROM include hematologic/coagulation function disorder, collagen metabolism, matrix degradation and local inflammation (Capece et al. 2014). This topic will be discussed in greater depth in the next chapter focussing on "omics" and sPTB.

1.3 Prediction of Spontaneous Preterm Birth

Obstetric History

There is particular difficulty identifying women at risk of sPTB in their first pregnancy. Amongst singletons a history of sPTB remains the most powerful predictor, whilst twin pregnancies are at 40% risk of delivering spontaneously before 37 weeks. A meta-analysis quantifying the risk of recurrence of sPTB based on different subtypes of subsequent pregnancy is summarised in Table 1.3.

Women with a previous PTB at <37 weeks of gestation are at an increased risk for recurrent PTB compared with women who have a previous term birth (OR 5.43, 95% CI 4.03-7.31). Risk of subsequent PTB is increasing with decreasing gestational age in the previous pregnancy (Kazemier et al. 2014). This data is taken from 13 relevant studies from 9104 identified publications, including n = 760, 937 women.

Table 1.3. Effect of past obstetric history upon absolute risk of PTB. Adapted from Kazemier et al. 2014.

First Delivery	Second Delivery	Absolute Risk of PTL
Term Singleton	Preterm Singleton	4.0% (95% CI 3.9-4.0)
Term Twin	Preterm Singleton	1.3% (95% CI 0.7-2.0)
Preterm Singleton	Preterm Singleton	20.2% (95% CI 19.9-20.6)
Preterm Twin <30 weeks	Preterm Singleton	10.0% (95% CI 8.0-12.1)
Term Singleton	Preterm Twin	25.4% (95% CI 24.3-26.5)

Ultrasound Measurement of Cervical Length

Transvaginal ultrasound (TVUSS) measurements of cervical length (CL) are used for prediction of PTB in two broad populations; 1) women with symptoms of preterm labour and 2) asymptomatic populations. The asymptomatic populations can be further subdivided into high-risk (women with known risk factors for sPTB) and

low-risk (no risk factors for sPTB). Transvaginal imaging shown in Figure 1.2 is more accurate than transabdominal USS measurement, with clearer images of the cervix obtained in this view.

There is an increased likelihood of sPTB as the CL decreases. When using a short CL for risk prediction in an asymptomatic cohort, prediction is improved when screening a predefined high-risk population. For a 10% false-positive rate, the detection rate of spontaneous delivery before 32 weeks was 38% for maternal factors (obstetric history, smoking etc.), 55% for CL measurement alone and 69% for combined testing (To et al. 2006). A systematic review of TVUSS measurement during the second trimester found that, using ROC curves, the test performs best when a cut off of ≤20 mm at ≤24 weeks is used and PTB is defined as <35 weeks gestation.

CL can also be plotted on a nomogram of cervical length in pregnancy as shown in Figure 1.3 (Salomon et al. 2009). This allows for temporal or serial plotting to ensure cervical length centiles are being maintained and there is no acute or rapid shortening. However, population centiles should be designed specifically for local populations as there is significant variation in the prevalence of women with a short cervix when screening low risk cohorts.

In women who present with threatened preterm labour (PTL) using a CL cut-off of ≤15 mm appears most accurate in predicting spontaneous delivery within 7 days with a sensitivity and specificity of 74% (95% CI, 58%–85%) and 89% (95% CI, 85%–92%), respectively (Boots et al. 2014).

Figure 1.2. Transvaginal Ultrasound Image of Cervical Length measurement. Yellow crosses represent callipers placed at the internal and external cervical ostia. The dashed yellow line between these points represents the distance measured shown in mm at the bottom right of the image. 41 mm is considered to be a 'long' cervix and the true cervical measurement is likely to be longer as the cervical canal curves below the straight line measurement on the image.

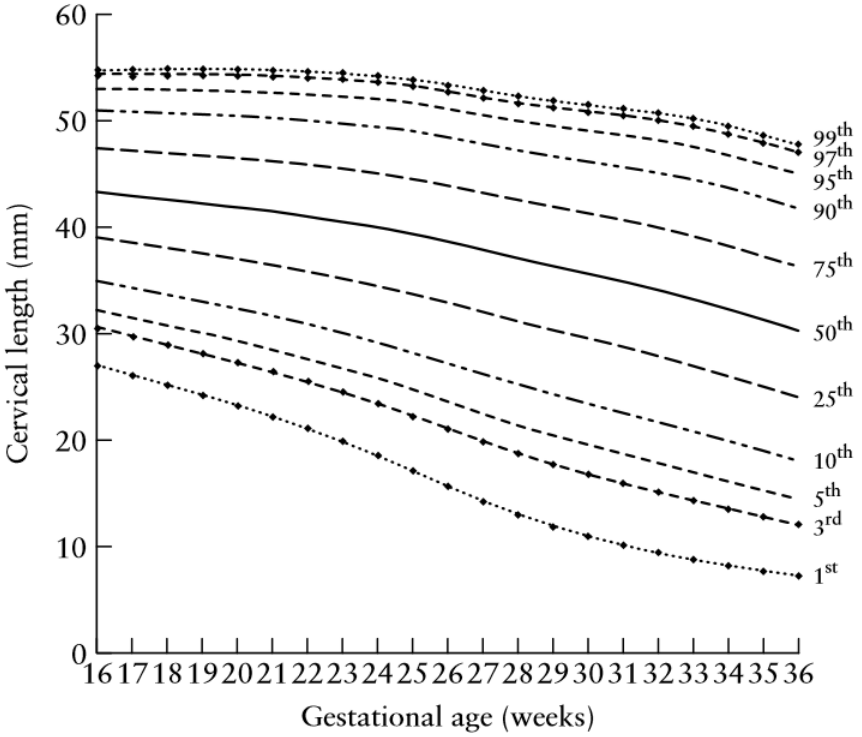


Figure 1.3. Reference ranges for cervical lengths across gestations. From Salomon LJ et al. 2009.

Cervico-vaginal biomarkers of preterm birth

Biological fluids such as amniotic fluid (Menon et al. 2014, Baraldi et al. 2016), blood (Saade et al. 2016) and saliva (Lachelin et al. 2009) are a rich source of biomarkers. Researchers have looked for hundreds of predictive markers, but early significant results are not often reproducible in validation studies. Measurement of two serum proteins, insulin-like growth factor – binding protein 4 (IBP4) and sex hormone binding globulin (SHBG), have shown to have some prediction of preterm birth (Saade et al. 2016). A private American company (Sera Prognostics®) currently offer measurement by mass spectrometry but this service is not currently used in the UK. More success has been obtained with biomarkers in cervico-vaginal fluid, with three biomarker bedside tests currently used in clinical practice in the UK for women symptomatic of preterm labour.

Detection of fetal fibronectin (fFN), phosphorylated insulin-like growth factor binding protein-1 (IGFBP1) and placental alpha microglobulin 1 (PAMG-1) in cervicovaginal fluid are all used as predictive tests of spontaneous PTL (Care et al. 2018).

Lockwood *et al.* (1991) were the first to report an association between fFN and PTB. fFN is a glycoprotein that has been described as a biological "glue" that binds chorion with maternal decidua in the extracellular matrix. After complete fusion of the chorion and decidua at 20 weeks, fFN levels are low (<50 ng/ml) in cervicovaginal secretions and are thought to be released through mechanical or inflammatory mediated damage to the membranes before birth (Lockwood et al. 1991). Lockwood *et al*. (1991) used an enzyme linked immunosorbent assay (ELISA) test against the monoclonal antibody FDC-6 (originally discovered by Matsuura H and Hakomori. (1985)) to detect fFN, with increasing concentrations

associated with increased likelihood of PTB. Although fFN has been recognized as the best predictor for spontaneous PTB <32 weeks, even when compared to a short CL (<25 mm) in the asymptomatic population (Goldenberg et al. 1996), it has had limited impact as a screening tool. Unlike a short cervix, there are currently no preventative treatments tested in clinical trials that have known benefit once a high fetal fibronectin is detected. More recent systematic reviews have challenged its use in the asymptomatic high risk and low risk population at all. It has been suggested that fFN may only be a clinically useful test in symptomatic patients where the biggest difference can be seen between pre and posttest probabilities of PTL. However, in all clinical settings considered within this review, none resulted in a positive summary likelihood ratio (sLR) above 10 and a negative sLR of less than 0.1, indicating, at best, a moderate predictive performance regardless of reference outcomes considered, clinical conditions, or the type of population tested (Faron et al. 2018).

In symptomatic populations, the most advantageous feature of qualitative fFN is its high negative predictive value (NPV) (0.93, CI 95% 0.92 – 0.95) which helps prevent overtreatment with unnecessary antenatal corticosteroids, reduces anxiety and returns women to normal care pathways (Melchor et al. 2018). The positive predictive value (PPV) is low (19.7%) using a qualitative test (any result above >50 ng/ml is considered a positive result), this can be increased to 37.0% or 46.2% using a quantitative test, and thresholds of 200 ng/ml or 500 ng/ml respectively (Abbott et al. 2013).

A systematic review demonstrated that the test is most accurate in predicting sPTB within 7-10 days among women with threatened PTB before advanced cervical

dilatation, with median likelihood ratios of 5.4 (95% CI 4.4-6.7) (De Franco et al. 2013).

The ability to predict sPTB using transvaginal USS and fetal fibronectin together is improved by concurrent usage (Gomez et al. 2005). In a symptomatic population, women with a CL of at least 30 mm or with a CL between 15 and 30 mm with a negative fibronectin result are at low risk (<5%) of spontaneous delivery within 7 days (Van Baaren et al. 2014).

IGFBP1 is a 25kDa protein that is secreted by maternal decidual cells as a highly phosphorylated isoform, phIGFBP1 (Martina et al. 1997). Similar to FFN, detection of phIGFBP1 in the cervicovaginal fluid of the posterior fornix indicates a disruption of the choriodecidual interface. The phIGFBP1 test has a comparable NPV to the test in predicting spontaneous PTB within 7 days in symptomatic women (phIGFBP1 92% vs. fFN 97%) (Ting et al. 2007), and a more recently published meta-analysis suggests it is in fact better (NPV phIGFBP1 0.99 vs. fFN 0.93) (Melchor et al. 2018). The advantage of this test is that it can be used in women who have had recent sexual activity or bleeding, which is a contraindication to fFN use. Unfortunately, it has poor performance in the asymptomatic population.

PAMG-1 is a glycoprotein discovered in 1976 that is produced by the decidua. It exists in high concentrations in amniotic fluid, but low concentrations in the cervicovaginal discharge (Petrunin et al. 1976). Originally, PAMG-1 was used to develop a bedside test for PPROM called Amnisure ROM® test. However, false positive findings in women with intact membranes led to the discovery that it also had predictive ability for PTL within 7 days. Based on this finding, the PartoSure® bedside test was later developed by the same company. Two potential theories have been established by *Lee et al.* (2009) for this additional ability to predict PTL. Firstly

imminent onset of PTL may result in the transudation of PAMG-1 through chorioamniotic pores in fetal membranes during uterine contractions and/or secondly, through the degradation of the extracellular matrix of fetal membranes due to the inflammatory process of labour and/or infection allowing PAMG-1 to permeate.

Comparison of these tests has been complicated by changing prevalence of PTL between study populations affecting positive and negative predictive values despite stable sensitivities and specificities. A direct comparison of PAMG-1 and phIGFBP1 independently and in combination with the gold standard cervical length measurement in 383 patients across three hospitals demonstrated that PAMG-1 has a significantly higher PPV and specificity compared with phIGFBP1 for the prediction of sPTB at 7days ($P<.01$). Both tests had comparable sensitivity and negative predictive value (Nikolova et al. 2018).

Risk Stratification – QUIPP app

Predictions of imminent PTL in symptomatic and asymptomatic women is reliant on picking up end stage physiological processes to allow for appropriate management. It does not aid with prevention of preterm birth by identifying an at-risk population early enough to benefit from preventative treatment strategies. The app does not recommend treatment types or thresholds, but alerts users to those most at risk. This makes for a difficult screening tool as no treatment is available in the identified populations unless a short cervix is identified.

In view of the multiple pathophysiology of PTB, it is unrealistic to expect a single biomarker to be able to predict sPTB in early gestation. The ideal biomarker test or predictive model should try to incorporate the fewest numbers of biomarkers to be measured, be highly sensitive and specific, exist in a biological fluid that is

without risk to obtain, and be detectable early enough in pregnancy to allow for preventative measures to be taken.

Currently the development of the "QUIPP" electronic application that can be used on mobile phones gives a risk prediction in either asymptomatic or symptomatic populations (Watson et al. 2020, Carter et al. 2020). A screenshot of the application is shown in Figure 1.4. This application requires input of obstetric history, current gestation, cervical length and fetal fibronectin to give a risk prediction score for PTL within 1, 2 and 4weeks and also, risk of delivery before 30, 34 and 37 weeks. This app has started to be used in preterm birth prevention clinics in the UK (Care et al. 2019), but risk thresholds for treatment remain unclear. Unpublished data from the Liverpool Harris-Wellbeing Preterm Birth prevention clinic retrospectively applied the app data to the cohort of 119 women recruited to a biomarker study. Clinicians were blinded at the time of taking fetal fibronectin. Using a treatment threshold of a QUIPP risk of PTB < 34 weeks >10% (referred to as "QUIPP positive") would more than double treatment rates from 20% (24/119 treated CL alone) to 42% (51/119 women treated) and 43 of 51 QUIPP positive women were still pregnant at 34 weeks (false positive rate) (Goodfellow et al. 2019). Fifteen of the 119 women (13%) had PPROM or sPTB <34 weeks with 8/15 women (53%) identified using the QUiPP app. This 10% treatment threshold gave a positive likelihood ratio (LR) of 1.3 (95% CI 0.76-2.18), and negative LR of 0.8 (95% CI 0.45-1.40). Modification of the treatment threshold could not improve on this. Existing publication suggests that clinicians are comfortable with a risk threshold of 5% (Carter et al. 2020, Carter et al. 2016), if this was used clinically as a treatment threshold for preterm birth preventative therapy then it would result in many more women being treated with an even higher false positive rate.

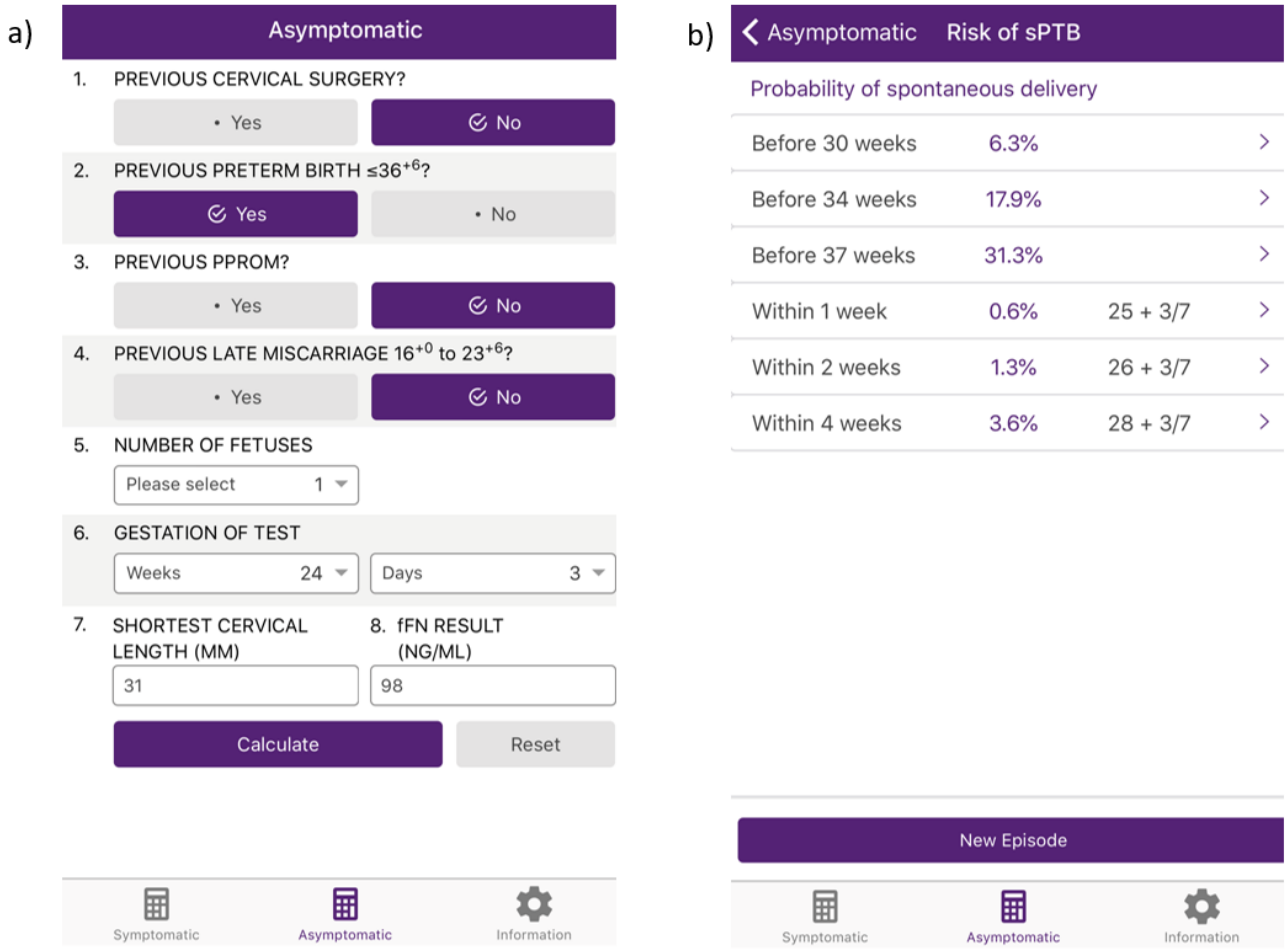


Figure 1.1 a) Screenshot of QUIPP app which requires data on previous obstetric history, cervical length and fFN result b) Screenshot of resulting risk prediction scores.

The QUiPP algorithm, as a concept, is a positive advance in the field of preterm birth prevention. However, in women with a previous preterm birth or PPROM the pre-test probability of a recurrent event is often too high to reassure the clinician or patient and use of this risk prediction score may result in over treatment without affecting preterm birth rates.

1.4 Prevention of Spontaneous Preterm Birth in Singleton Pregnancies

Although prediction is key, prevention remains the ultimate goal. Prevention can be classified into primary and secondary preventative strategies. The aim of primary prevention is to lower the incidence of PTB by improving physical and mental wellbeing and avoiding modifiable behavioural factors associated with PTB. For example, smoking cessation lowers the risk of sPTB by 16% (OR 84%, 95% CI 72%-98%) (Vanderhoeven and Tolosa. 2010).

Secondary preventions are interventions targeted to an at-risk population identified from the general population. A recently published Cochrane review summarised all evidence for interventions relevant to the prevention of PTB as reported in Cochrane systematic reviews (SRs) (Medley et al. 2018b). Four systematic reviews reported clear evidence of benefit from: 1) midwife-led continuity models of care versus other models of care for all women; 2) screening for lower genital tract infections for pregnant women less than 37 weeks' gestation and without signs of labour, bleeding or infection; and 3) zinc supplementation for all pregnant women without systemic illness. The fourth showed that cervical cerclage showed clear benefit for women with singleton pregnancy and high risk of PTB only.

At present there are only effective preventative treatments for women identified with a short cervix. As a result, screening clinics have been set up across the UK to perform transvaginal USS women at high risk of preterm labour (Sharp & Alfirevic. 2014, Care et al. 2018, NHS England 2019).

Omega 3

A recent Cochrane review showed omega 3 supplementation to be associated with a reduction in PTB <37 weeks (13.4% versus 11.9%; risk ratio (RR) 0.89, 95% CI 0.81 to 0.97; 26 RCTs, 10,304 participants; high-quality evidence) and early

preterm birth <34 weeks (4.6% versus 2.7%; RR 0.58, 95% CI 0.44 to 0.77; 9 RCTs, 5204 participants; high-quality evidence). This evidence was consistent across participants with a range of baseline risks for preterm birth (Middleton et al. 2018). However, it probably increases the risk of post term pregnancies. Prolonged gestation > 42 weeks was increased from 1.6% to 2.6% in women who received omega-3 long chain polyunsaturated fatty acids (LCPUFA) compared with no omega-3 (RR 1.61 95% CI 1.11 to 2.33; 5141 participants; 6 RCTs; moderate-quality evidence).

Although the exact mechanism by which omega-3 reduces sPTB is not exactly certain, it is known to have anti-inflammatory properties. Regulatory signalling by omega-3 polyunsaturated fatty acids (n-3 PUFAs) has been reported via, among others, the selective FFAR4/GPR120 (free fatty acid receptor 4) protein leading to reduced activity of the NFKB (nuclear factor kappa B) complex and the inflammasome (Oh et al. 2010, Liu et al. 2014). It may therefore play a role in increasing the threshold to transition into the inflammatory labouring state.

It remains unclear whether omega supplements to the general obstetric population may reduce sPTB rates without increasing the complications of post term pregnancies, particularly increased risk of stillbirth.

Antibiotics

Although the role of inflammation in sPTB pathophysiology is well documented, and antibiotics are recommended in some PTB guidelines (Medley et al. 2018a), there is no evidence to show that the use of antibiotics in the management of patients in threatened preterm labour reduces the incidence of preterm birth. The ORACLE II trial randomised women (n = 6295) in PTL with intact membranes to

one of three antibiotic groups or a placebo group taken four times a day. Although the antibiotics were associated with a lower risk of maternal infection, none of the regimes associated with a lower risk of sPTB (Kenyon et al. 2001). These findings were also echoed in a meta-analysis of fourteen trials (n = 7837 women) although predominantly consisting of ORACLE II data (Flenady et al. 2013).

The Cochrane systematic review of systematic reviews to highlight areas for further investigation and development found no effect for antibiotic prophylaxis in the second and third trimester (Medley et al. 2018b). However, antibiotics for women with asymptomatic bacteriuria shows a possible benefit in reduction of preterm birth (Medley et al. 2018b).

Probiotics

With interest increasing in the effect of the vaginal microbiome on preterm labour physiology, probiotics are a topic of discussion for preterm birth preventions. Unfortunately, at present there is insufficient data to recommend probiotics due to small trial numbers (Othman et al. 2007). A Cochrane systematic review evaluated only three trials, and the effect of probiotics on vaginal infection was assessed in only two trials with 88 women. Although there was an 81% reduction in the risk of genital infection with the use of probiotics (RR 0.19; 95% CI 0.08 to 0.48), effects on preterm birth could not be supported as confidence intervals were very wide and included no effect. Effect on PTB <37 weeks (one trial; 238 women) had a relative risk of 3.95 (95% CI 0.36 to 42.91) (Othman et al. 2007).

With the advancement in the assessment of the vaginal microbiome, targeted probiotics for high risk women or women with dysbiotic vaginal microbiomes may be one area of investigation for the future.

Prevention Therapies for Women with Short Cervix and/or History of sPTB

In singleton pregnancies, if asymptomatic women are identified to have a short cervix preventative therapy can be offered to reduce risk of sPTB. Vaginal progesterone (Fonseca et al. 2007, Hassan et al. 2011, Dodd et al. 2013) and, to a lesser extent, Arabin pessary (Goya et al. 2012) have demonstrated a reduction in risk of sPTB risk in an unselected obstetric population with a short cervix (<25 mm). However, doubts about the feasibility of adequate quality control have prevented universal screening programmes being introduced for an obstetric population.

In women with a previous sPTB and a short cervix (<25 mm before 24 weeks) a meta-analysis of five trials of cervical cerclage (n=504) has shown a reduction in risk of sPTB risk <37 weeks by 36% (Berghella et al. 2011) but no trial has shown benefit in an unselected population with short cervix (Rust et al., 2000. To et al. 2004). There is also supportive data for vaginal progesterone in this same population (Romero et al. 2018). A meta-analysis of five trials showed a risk reduction of PTB ≤ 34 weeks of gestation or fetal death compared to placebo (18.1% vs 27.5%; RR, 0.66 (95% CI, 0.52–0.83); P = 0.0005; five studies; 974 women) (Romero et al., 2018) and based on indirect comparison meta-analysis is considered as effective as cerclage. Clinical trials to directly compare available treatments in a high-risk population with a short cervix are currently ongoing (Hezelgrave et al. 2016, Pacagnella et al. 2019).

IM progesterone or 17α hydroxyprogesterone caproate (17-OHP) is a synthetic progesterone which has shown reduction of PTB < 35 weeks (RR 66%; 95% CI 54%-81%) in a trial of 463 women with a history of sPTB. (Meis et al. 2003) It is currently given routinely to women with a history of sPTB in the United States.

In view of cervical length screening and multiple available treatments for high risk populations with a short cervix; PTB prevention clinics have been set up throughout the UK. A recent survey of clinics found that a combination of treatments are currently being used in practice, but no predictive tests are used to tailor treatments individually or decide between treatment combinations (Care et al. 2018).

1.5 Rationale for Thesis

Preterm birth prevention antenatal clinics are the current UK screening model which combines predictive factors of previous obstetric history of sPTB and short cervix on transvaginal ultrasound (Care et al. 2018). However, the majority of sPTB will occur in the low risk population and there is currently no good screening service for these women. The introduction of cervical length screening across the whole population has been considered, but there are still reservations as to the number of women with short cervix that will be detected and ultimately the number of preterm births prevented. This has to be weighed up against the cost and feasibility of training for ultra-sonographers across the UK. Even in a high-risk population there remains a proportion of women who have recurrent sPTB but do not present with a short cervix <25 mm prior to 24 weeks. In our study 9% of women with a short cervix between 20-24 weeks will still deliver before 34 weeks despite screening (Care et al. 2014).

There is a need for a screening test that is acceptable to the population and can be used in both the high and low risk populations. Lack of understanding of the natural history of the disease is hindering progress in this field as multiple pathophysiological pathways to the same endpoint of preterm labour are in existence. Novel high throughout technologies now mean that we are able to obtain huge amounts of information from a single woman. Millions of data points about genes, transcripts and metabolites can be combined with clinical data to establish activity at a cellular level.

This thesis will address using these novel high throughput techniques in an exploratory pilot study to combine multiple layers of omics data from the same individual to identify candidate biomarker predictors or pathways. Figure 1.5 details the thesis structure.

Aims

1. The primary aim of the thesis was to establish a method of combining three 'omics' analysis used in this pilot study for the prediction of sPTB. This study's results will require replication in a larger validation study as this will not be performed as part of this thesis.
2. To use three different "omic" methodologies; genomics, transcriptomics and metabolomics, to analyse blood taken at 16^{+0} and 20^{+0} from women at high-risk of sPTB based on a previous history of sPTB or PPROM between 16^{+0}-33^{+6}.
3. Lastly, I aim to establish if there are distinct differences in biomarkers between PPROM and sPTB subgroups of spontaneous preterm birth.

Objectives

1. To review published studies examining genomic, transcriptomic and metabolomics analysis of women with sPTB or PPROM to establish existing biomarkers and/or biomarkers requiring validation.
2. Perform genome wide association study (GWAS) to investigate genetic factors that may correlate with spontaneous PTL in women who experience multiple spontaneous preterm birth and provide genomic data for combined omic analysis.
3. Perform transcriptomic analysis on extracted RNA to investigate gene expression correlating with sPTB or PPROM.
4. Perform nuclear magnetic resonance (NMR) analysis to establish correlation with known metabolites to be used as part of a combined 'omic' analysis.
5. Use a bioinformatic pipeline to combine all three layers of omic analysis for sPTB prediction.

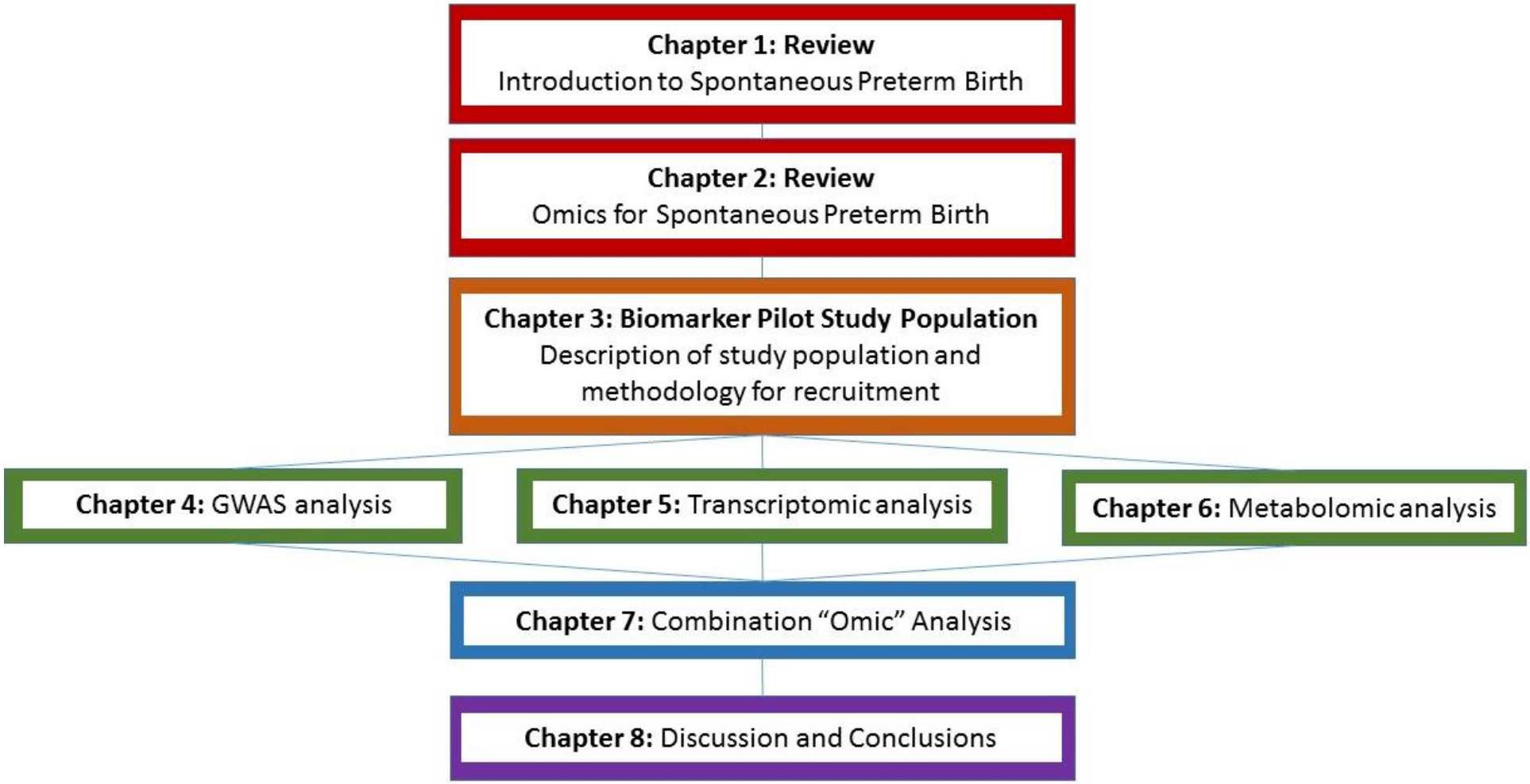


Figure 1.2. Thesis Structure

Chapter 2: Using “Omics” for spontaneous preterm birth prediction

2.1 Introduction

Chapter 1 highlights that our ability to predict sPTB is poor and mechanisms of this syndrome remain poorly understood. Current screening methods for prediction are insufficient to detect all women at risk (Care et al. 2014). Women identified as at risk do not always receive a treatment that is effective in preventing sPTB. The sequelae of preterm birth can lead to significant mortality and morbidity (Costeloe et al. 2012). There is a clear need for improvements in prevention, which cannot be achieved without a better understanding of causation or a more robust way of accurately discriminating those at high risk.

One promising area that has revolutionized personalised medicine is the use of "omics" in healthcare. At present they have contributed to medical advances on individual omic levels, such as exome and genome sequencing have improved diagnosis of rare disease (Worthey et al. 2011, Waggoner et al. 2018). In this chapter I will review the literature and various approaches taken related to the use of "omics" in the context of sPTB and discuss the progress of combining omics platforms as well as the limitations of these methodologies.

The suffix *-omic*, derived from the ancient Greek, refers to in-depth knowledge. Currently, we have over 30 such disciplines with the *-omics* suffix (Figure 2.1) (Kumar, D. 2015). Omics aims to characterize and quantify biological molecules that combined provide the knowledge of the structure, function, and behavioral phenotype of an organism. The topic is too vast to consider all omics platforms and therefore my thesis will be focused around the discussion of genomics, transcriptomics and metabolomics in the prediction of sPTB; this reflects my study design discussed in the next chapter.

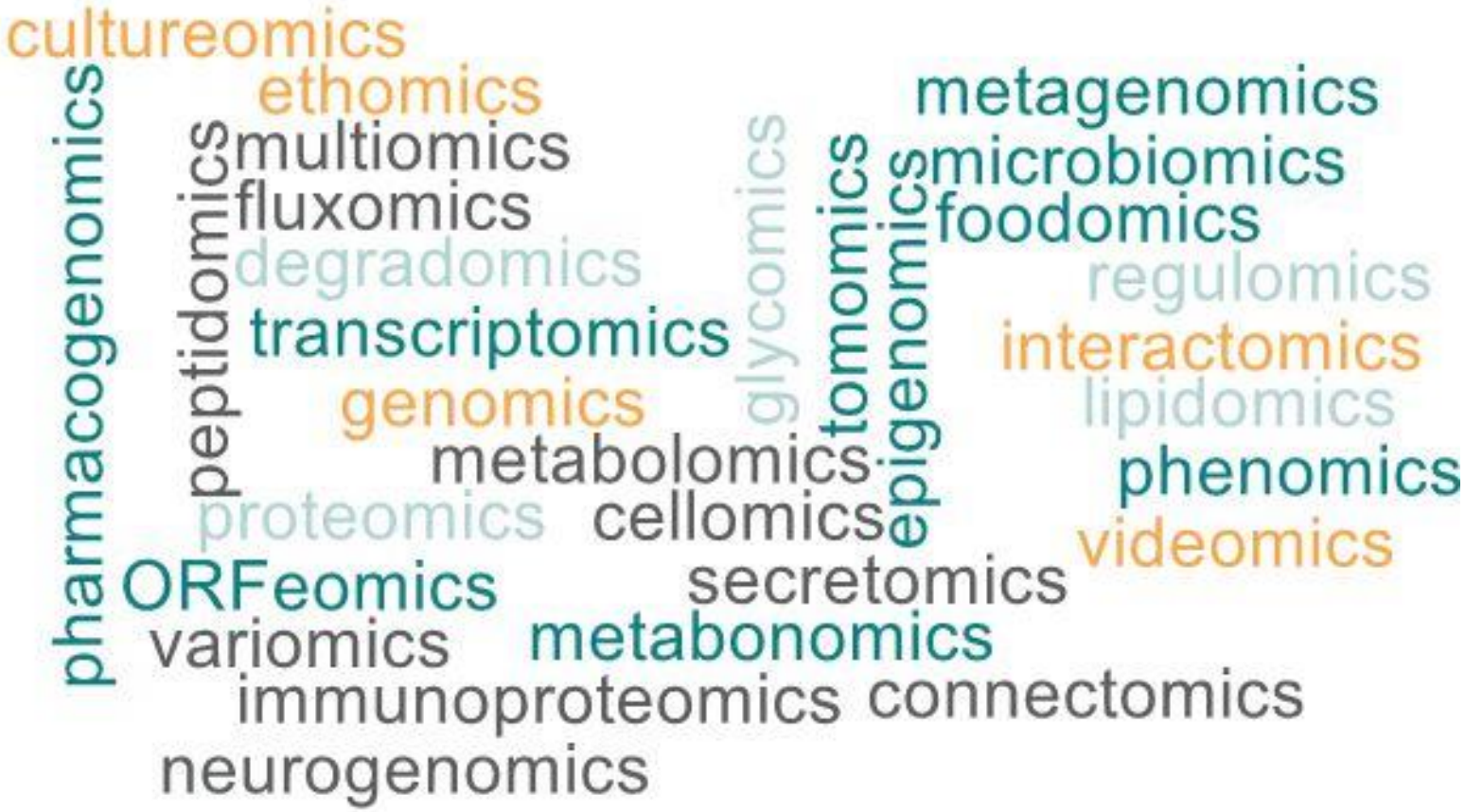


Figure 2.1. Omics word cloud demonstrating range of omic disciplines currently being researched

2.2 Genomics

Genomics is the science of understanding the genomes of any given organism or species. The genome refers to the complete genetic code of a unique individual and is comprised of two complimentary strands of deoxyribonucleic acid (DNA) (Watson and Crick. 1953). The backbone of these strands consist of a phosphate and a sugar molecule attached to one of four bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Due to the specific chemical structure of each base; A always pairs with T and C pairs with G. The length of the human genome is approximately 3 billion base pairs (Mattick. 2003). This DNA strand is tightly coiled and along its length surround proteins called histones which help define its structure and level of activity. A single histone cluster surrounded by DNA is called a nucleosome. Several nucleosomes together form chromatin and tightly coiled chromatin form larger structures called chromosomes. (Annunziato. 2008)

In humans there are 46 chromosomes arranged into 23 pairs, with one of each pair inherited from each parent. Twenty-two pairs are called autosomes with one pair called sex chromosomes consisting of the X and Y chromosome which determine gender. A female will have two X chromosomes (46, XX) and a male with have an X and Y (46, XY).

Interestingly the average human genome contains around 5-10 million genetic variants including 20,000 “coding” variants which are located within transcribed genes. A genetic variant refers to an alteration of the most common genetic sequences. Clearly not all these variants are clinically significant (Levy et al. 2007). Single nucleotide polymorphisms (SNPs) are the most common variation in the genome with an estimated 10 million SNPs occurring in the human population (International HapMap Consortium. 2003). A SNP is a single base pair substitution

at a particular location (locus) of the genome. Interestingly humans only differ from one another by 0.1% of their genetic make-up (Dolan and Christiaens. 2013) but this 0.1% may determine a given individual's disease susceptibility.

There is a good body of evidence that supports the premise that genetic factors residing in the maternal genome contribute up to 40% of the variation in birth timing and preterm birth. Evidence supporting this claim comes from epidemiological studies, twin studies and segregation analysis of pedigrees (Clausson et al. 2000, Kristka et al. 2008, Boyd et al. 2009, Plunkett et al. 2009). See section 2.1. Here I will explore the different techniques and study designs available to analyse the genome and its expression and review recent literature related to sPTB prediction in "omic" studies.

Candidate Gene Studies

Candidate gene studies have been at the very forefront of genetic association studies. They involve the selection of genes that are biologically plausible candidates for a condition such as sPTB. Therefore, require some prior knowledge about gene function and probable mechanisms of disease. Once genes are identified, assessment and selection of polymorphisms occur. The detection of genetic variants is usually assessed by selecting single nucleotide polymorphisms (SNPs) within that gene (Figure 2.2). The gene variant is then tested for its occurrence in a population with the disease or trait of interest (cases) and participants without the condition (controls). If the case and control cohorts are adequate based on a sample size calculation it can show statistically different polymorphisms in relevant genes. Detection of these associations can then lead to evaluation of the effectiveness in prognosis, diagnosis and usefulness as a potential biomarker.

A key limitation to these studies is that they are biased to current theories or presumed knowledge and cannot identify novel genes. However, these studies are relatively quick and cheap to perform and do not require studies of large families with a combination of unaffected and affected members.

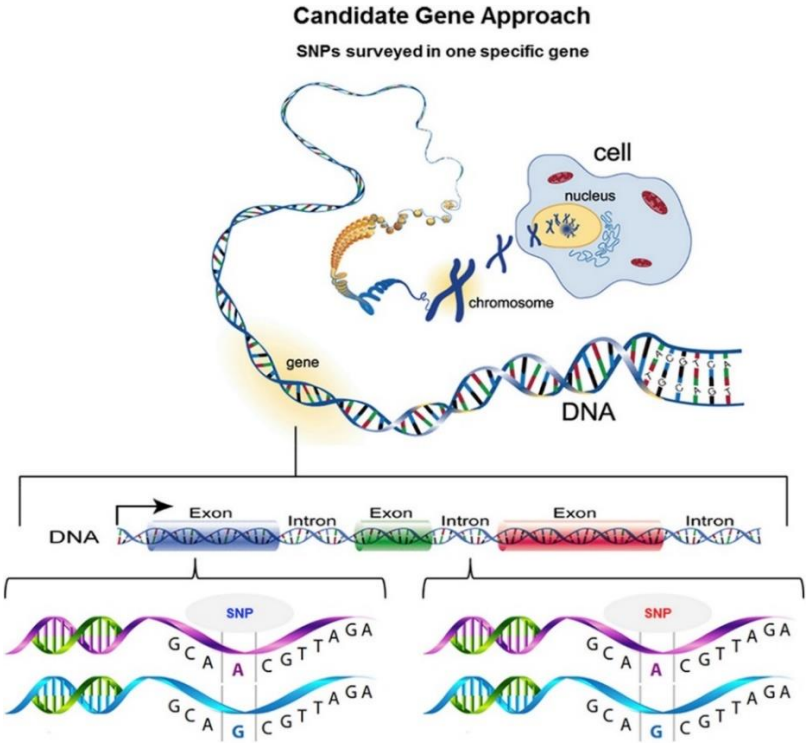


Figure 2.2 Single Nucleotide Polymorphisms refer to one base pair change and may fall within coding regions of the gene (exons; blue SNP), non-coding regions of genes, or the intergenic regions (introns; red SNP). (Image from Strauss et al. 2018.)

For sPTB candidate gene studies have shown that some polymorphisms in genes coding for components of the innate immune system (Table 2.1) are significantly associated with PTB, but most positive results have not been replicated or validated in a suitably large cohort.

Table 2.1. List of genes that have been studied for single nucleotide polymorphisms associated with the risk for preterm birth. (Adapted from Sheikh et al. 2016) The genes highlighted in bold have been identified in more than one.

Systems	Candidate Genes (nomenclature)	Reference
Endocrine system related genes	Corticotropin receptor 1 (CRHR1)	Bream *et al*. 2013
	Follicle stimulating hormone receptor (FSHR)	Plunkett *et al.* 2011. Chun *et al.* 2013
	Glucocorticoid receptor (NR3C1)	Bream *et al*. 2013
	Insulin-like growth factor 2 (IGF2)	Romero *et al.* 2010a, Romero *et al.* 2010b
	Insulin-like growth factor receptor 1 (IGF1R)	Hataaja *et al.* 2011, Bream *et al*. 2013
	Leucyl/cysteinyl aminopeptidase (LNPEP)	Kim *et al.* 2013
	Oxytocin (OXT)	Kim *et al.* 2013
	Oxytocin receptor (OXTR)	Bream *et al*. 2013, Kim *et al.* 2013, Kuessel *et al.* 2013.
	Progesterone receptor (PGR)	Ehn *et al.* 2007, Guoyang *et al*. 2008, Manuck *et al.* 2010, Oliveira *et al.* 2011, Bream *et al*. 2013, Mann *et al* 2013. Ryckman KK *et al.* 2010, Jeffcoat MK *et al.* 2014.
	Prostaglandin E receptor (PTGR3)	Liu *et al.* 2012 Bream *et al*. 2013
	Prostaglandin E synthase 2 (PTGES2)	Grisaru-Granovsky *et al.* 2010. Vogel *et al.* 2009. Rocha *et al.* 2013.
	Prostaglandin G/H synthase 1 (PTGS1)	
	Prostanoid DP receptor (PTGDR)	
	Relaxin 2 gene (RLN2)	
Tissue remodelling and biogenesis related genes	**Collagen type I (COL1A2)**	Manuck *et al.* 2011. Romero *et al.* 2010. Romero *et al.* 2012. Ryckmann *et al.* 2010. Myking *et al*. 2011.
	Collagen type IV (COL4A2)	Romero *et al.* 2010. Romero *et al.* 2012.
	Collagen type IV (COL4A3)	Romero *et al.* 2010. Romero *et al.* 2012.
	Collagen type IV (COL4A4)	Romero *et al.* 2010. Romero *et al.* 2012.
	Collagen type IV (COL4A5)	Romero *et al.* 2010. Romero *et al.* 2012.
	Collagen type IV (COL4A6)	Romero *et al.* 2010. Romero *et al.* 2012.
	Collagen type V (COL5A2) alpha-2	Romero *et al.* 2010. Romero *et al.* 2012.
	Intercellular adhesion molecule-1 (ICAM1)	Romero *et al.* 2010. Romero *et al.* 2012.
	Matrix Metalloproteinase 1 (MMP-1),	Myking *et al*. 2011.
	Matrix Metalloproteinase 8 (MMP-8),	Kwon *et al.* 2009.
	Matrix Metalloproteinase 9 (MMP-9),	Pereza *et al.* 2014
	Matrix Metalloproteinase 10 (MMP-10),	Ryckmann *et al.* 2010. Pereza *et al.* 2014. Jones *et al.* 2012.

	Matrix Metalloproteinase 16 (MMP-16), **Tenascin-R (TNR),** **TIMP metallopeptidase inhibitor 2 (TIMP2)**	Romero *et al.* 2010. Romero *et al.* 2012. Romero *et al.* 2010. Romero *et al.* 2012. Romero *et al.* 2010. Romero *et al.* 2012. Romero *et al.* 2010. Romero *et al.* 2012.
Vascular and angiogenesis related genes	Alpha adducing (ADD1) Angiopoietin 1 (ANGPT1) **Angiotensin converting enzyme (ACE)** Angiotensin II receptor type 1 (AT1) **Angiotensinogen (AGT)** **Beta-2 adrenergic receptor (ADBR2)** Complement receptor 1 (CR1) Cyclin-dependent kinase 4 inhibitor (CDKN2) **Endothelial nitric oxide synthase (NOS3)** Endothelin 1 (EDN1) Factor V (F5) **Inducible nitric oxide synthases (NOS2)** Kinase insert domain receptor (KDR) Peroxisome proliferator- activated receptor gamma (PPARG) **Plasminogen activator inhibitor-1 (SERPINE)** Renin (REN) Small conductance calcium activated potassium channel 3 (KCNN3) Thrombomodulin (THBD) Vascular endothelial growth factor (VEGFA)	Gibson *et al.* 2007 Andraweera *et al.* 2012 Valdez-Velazquez *et al.* 2007, Uma *et al.* 2008. Valdez-Velazquez *et al*. 2007. Valdez-Velazquez *et al.* 2007, Gargano *et al.* 2009 Gibson *et al.* 2007, Suh *et al.* 2013. McElroy *et al.* 2013 Falah *et al.* 2013 Gibson *et al.* 2007. Suh *et al.* 2013. Romero *et al.* 2010. Romero *et al.* 2012. Yu *et al.* 2009. Gargano *et al.* 2009 Gibson *et al.* 2007. Suh *et al.* 2013. Andraweera *et al.* 2012. Meirhaeghe *et al.* 2007. Gibson *et al*. 2007, Chen *et al.* 2007. Valdez-Velazquez *et al.* 2007. Mann *et al.* 2013. Bream *et al.* 2013. Day *et al*. 2011. Gibson *et al.* 2007. Andraweera *et al.* 2012. Andraweera *et al.* 2012.
Metabolism related genes	Apolipoprotein A-I (APOA1) Apolipoprotein C (APOC) Apolipoprotein E (APOE) ATP-binding cassette transporter (ABCA1) Cholesteryl ester transfer protein (CETP) Cytochrome P4501A1 (CYP1A1) Dehydrocholesterol reductase (DHCR24) FC alpha receptor (FC*a*R) Glutathione S-transferase mu 1 (GSTM1)	Steffen *et al.* 2007 Steffen *et al*. 2007 Steffen *et al.* 2007 Steffen *et al.* 2007 Steffen *et al*. 2007 Lewinska *et al.* 2013, 78, 81-85 Steffen *et al*. 2007 Sugita *et al.* 2012 Suh *et al.* 2008, Lee *et al.* 2010, Mustafa *et al.* 2013, Luo *et al.* 2012

	Glutathione S-transferase theta 1 (GSST1) Glutathione S-transferase theta 2 (GSST2) Glutathione S-transferase theta pseudogene (GSTTP1) Hepatic lipase (LIPC) Hydroxy methyl glutaryl CoA reductase (HMGCR) Lipoprotein lipase (LPL) Mannose binding lactin (MBL) Methionine synthase (MTR) Methionine synthase reductase (MTRR) Methylene tetrahydrofolate reductase (MTHFR) Methylenetetrahydrofolate dehydrogenase 1 (MTHFD1) Serine hydroxy methyltransferase 1 (SHMT1) Serum paraoxonase/arylesterase 1 (PON1) Vitamin D receptor (VDR)	Suh *et al.* 2008, Tsai *et al.* 2008, Zhang *et al.* 2008. Luo *et al.* 2012. Zheng *et al.* 2013 Zheng *et al.* 2013 Zheng *et al.* 2013 Steffen *et al.* 2007 Steffen *et al.* 2007 Falah *et al*. 2013 Falah *et al.* 2013 Gargano *et al.* 2009 Gargano *et al.* 2009 Gargano *et al.* 2009, Engel *et al*. 2006. Christensen *et al*. 2014. Gargano *et al.* 2009 Ryckman *et al.* 2010, Myking *et al.* 2011, Harley *et al*. 2011 Bream *et al*. 2013
Innate immunity and inflammation related genes	Colony-stimulating factor 2 (CSF2) **Defensin alpha 5 (DEFA5)** Fms-like tyrosine kinase 1 (FLT1) HLA class II histocompatibility antigen, DR alpha chain (HLA-DRA) HLA class II histocompatibility antigen, DRB1-9 beta chain **Interferon ɣ (IFN-ɣ)** Interferon ɣ receptor 2 **Interleukin 1 alpha (IL1α)** **Interleukin 1 beta (IL1β)** Interleukin 1 receptor 2 (IL1R2) **Interleukin 1 receptor antagonist (IL1RN)** Interleukin 1 receptor-associated kinase 1 (IRAK1) **Interleukin 2 (IL2)** Interleukin 2 (IL2) Interleukin 2 receptor beta (IL2Rβ) **Interleukin 4 (IL4)** **Interleukin 6 (IL6)**	Harmon *et al.* 2013 Romero *et al*. 2010. Romero *et al.* 2010b. Gomez *et al*. 2010 Falah *et al.* 2013 Falah *et al.* 2013 Moura *et al.* 2009, Devi *et al.* 2014 Harmon *et al.* 2013 Ryckman *et al.* 2010. Sata *et al.* 2009, Yilmaz *et al.* 2012. Jones *et al.* 2010, Hollegaard *et al.* 2008, Yilmaz *et al*. 2012, Schmid *et al.* 2012 Ryckman *et al*. 2010. Chaves *et al.* 2008, Kalinka *et al.* 2009, Jones *et al.* 2012. Karody *et al.* 2013

	Interleukin 6 receptor (IL6R) Interleukin 10 (IL10) Interleukin 12 (IL12) Interleukin 12 receptor (IL12Rβ) Interleukin 12 alpha (IL12α) **Interleukin 13 (IL13)** Interleukin 15 (IL15) Interleukin 23 receptor (IL23R) Killer cell immunoglobulin-like receptor three domain long cytoplasmic tail 2 (KIR3DL2) **Lactotransferrin (LTF)** Low-affinity receptor for immunoglobulin G (FcγRIIB) Major histocompatibility complex, class II (HL-DQA1) Nuclear factor-kappa B1 (NFkB1) Protein kinase C alpha (PRKCA) Selenoprotein S (SEPS1) Surfactant, pulmonary associated protein D (SFTPD) TIR domain receptor associated protein (TIRAP) **Tumour necrosis factor alpha (TNF α)** **Tumour necrosis factor receptor 2 (TNFR2)** TNF receptor associated factor 2 (TRAF2) Toll-like receptor 2 (TLR*2*) Toll-like receptor 4 (TLR*4*) Toll-like receptor 5 (TLR*5*) Toll-like receptor 9 (TLR*9*) Toll-like receptor 10 (TLR*10*) Transforming growth factor beta1 (TGF-*β1*)	Romero *et al*. 2010. Romero *et al.* 2010b. Velez *et al.* 2009 Ryckman *et al.* 2010, Heinzmann *et al.* 2009, Karjalainen *et al.* 2012 Moura *et al.* 2009, Kalinka *et al*. 2009, Velez *et al.* 2007 Velez *et al 2008a* Romero *et al.* 2010. Romero *et al*. 2010b. Ryckman *et al*. 2010. Velez *et al.* 2007 Stonek *et al.* 2008. Heinzmann *et al.* 2009. Velez *et al.* 2009 Karjalainen *et al*. 2012 Heinzmann *et al.* 2009. Karjalainen *et al.* 2012 Velez *et al.* 2009 Falah *et al.* 2013 Harmon *et al.* 2013 Romero *et al.* 2010. Romero *et al.* 2010b. Iwanaga *et al*. 2011 Falah *et al.* 2013 Karody *et al*. 2013 Gomez *et al*. 2010 Wang *et al.* 2013 Karjalainen *et al*. 2012 Karody *et al*. 2013 Moura *et al.* 2009, Jones *et al*. 2010, Jones *et al.* 2012, Pu *et al*. 2007, Hollegaard *et al*. 2008, Liang *et al*. 2010, Harper *et al.* 2011, Yilmaz *et al.* 2012, Drews-Plasecka *et al*. 2014., Jafarzadeh *et al.* 2013. Jones *et al*. 2012, Pu *et al*. 2007. Bream *et al*. 2013. Karody *et al*. 2013

		Rey *et al*. 2008, Bitner *et al*. 2013, Karody *et al*. 2013 Karody *et al.* 2013 Karody *et al.* 2013 Heinzmann *et al.* 2009 Devi *et al*. 2014
Miscellaneous genes	Catechol-o-methyltransferase (COMT) Early growth response 1 (EGR1) FERM domain containing protein 7 (FRMD7) Mitochondrial genome variants Transcription factor AP2A (TFAP2A) Specificity protein 3 (SP3)	Thota *et al.* 2012 Enquobahrie *et al.* 2009 Myking *et al*. 2013 Velez *et al.* 2008b Enquobahrie *et al.* 2009 Enquobahrie *et al.* 2009

Family Based Linkage Studies

Families have been used in the design of genetic studies dating as far back to Mendel's study to examine the concept of inheritance of traits in plants. They use designs based on related individuals, which could be sibling pairs, parents and offspring, or more complex family trees. Historically, they have been well suited to investigate conditions that have genes of major effect. For linkage studies, this is a family-based approach to identifying susceptibility genes. Linkage refers to the tendency for alleles at certain chromosome positions (loci) that are close together to be also transmitted/inherited together. The further apart genes are from each other, the more likely they are to be split apart by a recombination event during meiosis. Using known genetic markers that define the inheritance of the same chromosomal region among different family members, the approximate location of the region of causation is identified. Two studies of Finnish families identified three markers with the highest linkage on 15q26.3 with insulin-like growth factor 1 (IGF1R) using autosomal chromosomal markers (Haataja R et al. 2011). When the same families were used to identify X chromosomal markers, an additional two genes, androgen

receptor gene at Xq12 and interleukin-2 receptor gamma subunit (IL2RG) located on Xq13, were implicated with preterm birth (Karjalainen et al. 2012).

Genome Wide Association Study (GWAS)

Unlike candidate gene studies, GWAS aims to identify common genetic variation (>5% frequency) in hundreds of thousands of SNPs across the entire genome without bias of pre-existing knowledge. Complex or multifactorial diseases, such as sPTB, are thought to be polygenic, in contrast to highly penetrant single gene disorders. As GWAS is looking to identify common genetic variation with small effect sizes, GWAS requires a large study sample for the discovery of statistically significant contribution to a trait.

Recently, a GWAS examining a large population of >40,000 women of European ancestry and a replication cohort of >8,000 women from a Nordic dataset have identified several maternal loci (EEFSEC, EBF1 and AGTR2) that may contribute to the length of gestation and preterm birth in Caucasian women (Zhang et al. 2017). EEFSEC encodes a protein involved in the production of selenoproteins, EBF1 encodes a protein implicated in B-cell development and AGTR2 encodes the type 2 angiotensin II receptor. The study provided new lines of investigation away from the increasing evidence for inflammation and innate immunity. Until this study, other individual GWAS of sPTB had not replicated loci with genome wide significance (Monangi et al. 2015). The possible reasons for this include:

- variation in definitions of preterm birth across studies (i.e. not strictly spontaneous preterm birth inclusion and variation in gestational age definitions),
- insufficient study numbers to detect small-effect sizes across the entire allele frequency spectrum,

- sPTB is caused by rare variants which had not been tagged by conventional genome wide arrays.
- Paternal genes not considered in this method of analysis or their contribution to fetal genetic effect and may play more significant role.

Additionally, the complexity of the sPTB phenotype suggests that there may be a genetic predisposition to different phenotypes of sPTB. There may be a small genetic contribution to each pathophysiology or phenotype and a combination of genetic, transcriptomic and environmental interaction analysis is required to examine biological pathways tracing back to the key genes.

Apart from just conducting a GWAS, researchers are now interested in identifying SNPs that may significantly interact with environmental influences such as pre-pregnancy body mass index (BMI). Hong *et al.* (2017) identified and replicated a significant interaction between maternal genotype rs11161721 in the *COL24A1* gene and pre-pregnancy BMI category on overall PTB risk in an African American population (n=1733; 698 mothers of PTB, 1035 term birth) from the Boston Birth Cohort (BBC) and in independent GWAS data sets deposited in the database of Genotypes and Phenotypes (dbGaP), respectively (African American n=780, Caucasian n=683). Although three other SNPs were found to be significantly associated with different PTB outcomes in the BBC, the associations were not confirmed in either of the two independent data sets (Hong et al. 2017).

Mitochondrial Genetics

Following the positive results from maternal GWAS data the risk of preterm birth may be associated with the maternal genome, a logical possibility is that this may occur through maternal transmission of the mitochondrial genome via mitochondrial DNA (mtDNA). The mtDNA is of interest in sPTB because oxidative stress is likely to play an important role in labour. (Velez et al., 2008b). Additionally, there is good evidence that mitochondria represent a major source of reactive oxygen species in aging tissues (Cadenas et al. 2000), and mitochondria have a protective role against inflammation. Inflammation itself may also have a role in the mediation of aging; a process called inflammaging, with both processes sharing many of the same pathways including oxidative stress and DNA damage (Salminen et al. 2012). In one study of preterm birth, Velez *et al.* (2008) used previously established mtDNA variants and intersected them with smoking, a known risk variable for PTB which increases oxidative stress. Marginal significance was shown for two of the mutations, A4917G and T4216C, however this has not been validated.(Velez et al. 2008b) Contrary to this, Alleman *et al.,* (2012) examined the association between mitochondrial genotypes and preterm delivery using a meta-analysis to combine two large GWAS studies and tested for associated 135 mitochondrial genome SNPs (mtSNP). No single mtSNP reached genome wide significance and they did not support the theory that mitochondrial genetics contributes to maternal transmission of PTL and related outcomes. (Alleman et al. 2012) More recently, Crawford *et al*., (2018) discovered that infants with increasingly divergent mitochondrial and nuclear genome were more at risk of sPTB. This might go some way to explaining the higher rates of sPTB in the African American population when compared to a Caucasian American population even after controlling for deprivation. This is an area of

interest that requires further study to elicit the role of mtDNA and the interaction with nuclear DNA leading to this phenomenon.

A study of fetal membranes examining the mRNA expression of mitochondrial enzyme manganese superoxide dismutase (Mn SOD), that scavenges ROS and is upregulated in sites of inflammation, found upregulation in term labour and in preterm labour only in the presence of histological chorioamnionitis (Than et al. 2009). Therefore, there is inconclusive evidence regarding the involvement of mitochondrial genetics, but this remains a promising area for future work.

Whole Exome Sequencing (WES)/Whole Genome Sequencing (WGS)

GWAS, using microarray technology, has only very recently made better progress in PTB genetics. The lack of reproducible results is potentially explained by the possibility of sPTB being caused by rare rather than common variants. With high throughput technologies becoming increasingly affordable and accessible, it has become easier to perform whole exome and whole genome sequencing using next-generation sequencing (NGS) strategies searching for new or rare variants. NGS is the catch all term to describe several different modern sequencing technologies capable of interrogating the entire genome or transcriptome, not depending on pre-chosen targets like microarray. NGS is based on synthesis, i.e. the incorporation of nucleotides by a DNA polymerase while microarrays are based on hybridization to strands of complimentary mRNA probe stored in wells of a chip. Even for the newest genome-wide microarrays, it is virtually impossible to include probes against every single nucleotide position.

Although the exome refers to less than 2% of the genome as a whole, it contains approximately 85% of known disease-causing variants. A whole exome sequencing (WES) of genomic DNA of neonates born to African-American mothers

whose pregnancies were complicated by PPROM (n=76) compared to normal term pregnancies (N=43) identified rare heterogenous nonsense and frameshift mutations only present in PPROM cases in several candidate genes (CARD6, DEFB1, FUT2, MBL2, NLRP10, NOD2) involved in dampening the innate immune response (Modi et al. 2017). These results suggest that PPROM may be caused by infrequent genetic variants that modulate fetal membrane strength leading to weakening of the membranes and ultimately ending in premature rupture.

However, limitations of whole exome sequencing (WES) include missing structural variants. WES cannot detect important intragenic variation including areas of regulatory elements responsible for gene expression and it can miss some exons. The efficacy of the capture of exons depends on the percentage GC nucleotide composition of the targeted sequence (or GC capture bias). For exons with especially high GC content, exome sequencing can fail to produce enough coverage for accurate variant detection and calling. For these purposes whole genome sequencing (WGS) is better as it will also identify the less common mutations such as frame shift mutations (insertion/deletion of nucleotides that is not divisible by 3 and changes the reading frame order) and point mutations (single substitution / insertion / deletion of single nucleotide). Achieving only a low coverage may miss many variants, whilst cost increases for a deeper coverage makes it a prohibitively expensive technique for large cohorts.

2.3 Transcriptomics

The transcriptome is the collection of all RNA (ribonucleic acid) molecules in a cell, tissue or organism. Transcription is the first step in gene expression in which information is taken from the gene for creating proteins or transcripts used to regulate gene function. RNA molecules mirror the sequence of the DNA bases to transcribe the code. This process is controlled separately for each gene in the genome (Figure 2.3). A single strand of the DNA helix is used as a template and RNA creates a copy of the same base information as the non-coding strand, except the base uracil (U) is used instead of thymine (T). The enzyme RNA polymerase is used to link nucleotides to create a chain of nucleotides. RNA is then processed by splicing with a 5' cap and poly-A tail put on either ends of the read and is subsequently called messenger RNA (mRNA). A stretch of three bases (codon) in the mRNA determines the position of an amino acid in a growing protein molecule (Figure 2.3).

By analysing the mRNA it is possible to determine which genes are being transcribed in all available cells present in a sample at that specific time point. However, the measurement of mRNA levels provides an imperfect reflection of protein levels and activity. The concentration of a protein is controlled not only by the level of its mRNA, but also by the rate of mRNA translation into protein and protein degradation. Other protein modifications, such as phosphorylation, are also important determinants of activity. With these limitations in mind, measurement of global mRNA expression gives insight into the overall level of gene and protein expression (Sealfon and Chu. 2010). Similar to genomics there are two common methods to measure expression; hybridisation (microarray) or a seq-based approach.

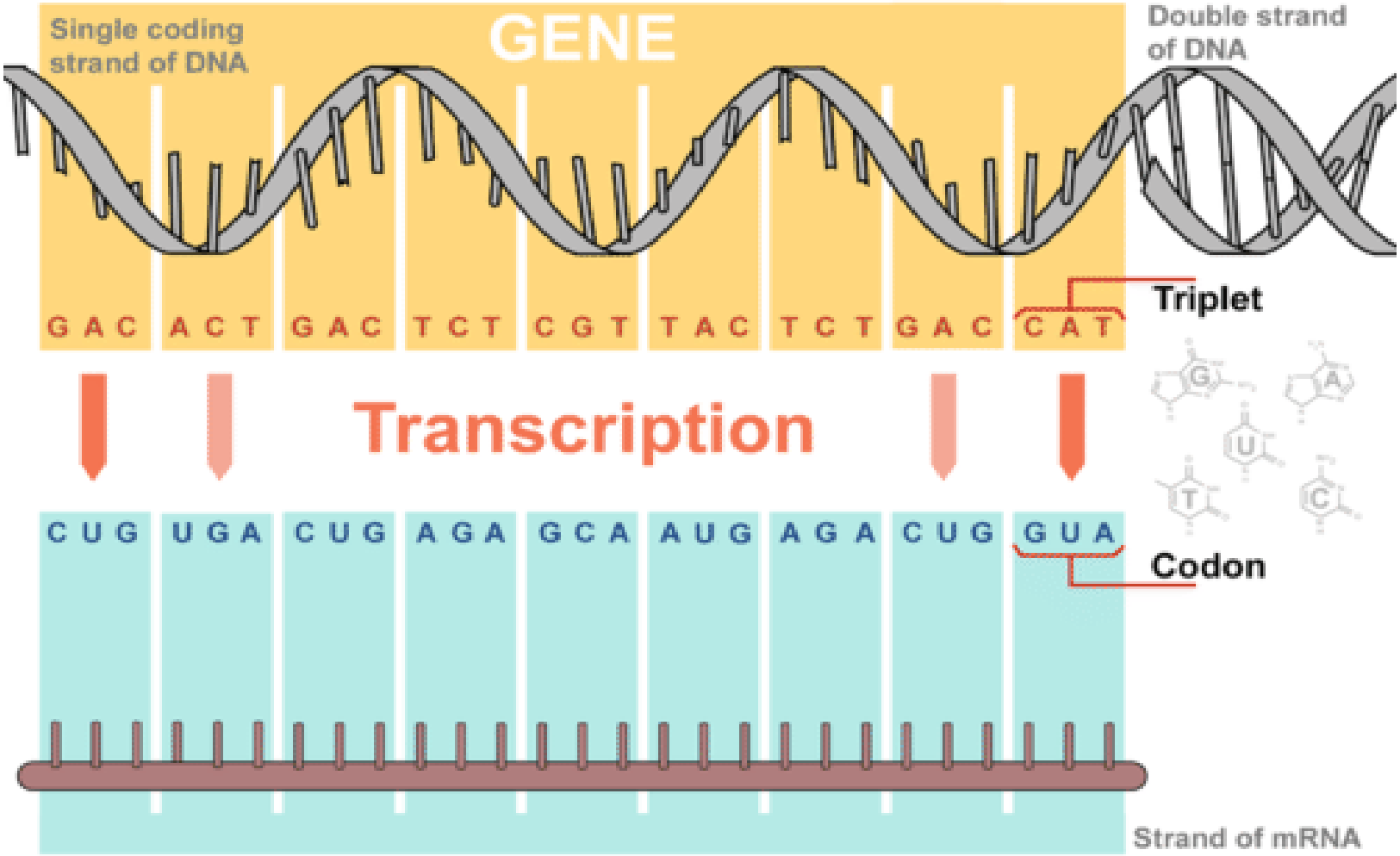


Figure 2.3. Overview of Transcription. Transcription uses the sequence of bases in a strand of DNA to make a complementary strand of mRNA. Triplets are groups of three successive nucleotide bases in DNA. Codons are complementary groups of bases in mRNA. (Image used with permission from https://courses.lumenlearning.com/wmopen-biology1/chapter/translation/)

Microarray

Microarray technology was first introduced in 1995 by Patrick Brown and colleagues (Schena et al., 1995). RNA microarrays have been widely used to identify regulated genes, pathways, or gene networks in a variety of cells and tissues when two or more related biological conditions are compared.

Figure 2.4 shows a schematic of microarray. cDNAs are amplified from individual clones in a library. Each cDNA fragment representing an individual gene of interest is immobilized on a glass slide that has been coated with nucleotide-binding chemicals. These slide arrays can be printed as whole genome microarrays or with a focused selection of genes of interest (Sealfon and Chu. 2011). Because of their shorter turnaround time, ease of analysis and cost-effectiveness microarrays remain the most popular approach in transcriptomic profiling. Microarray also still yields higher throughput than RNA Seq which has significant advantages when working on projects with large numbers of samples. However, as microarray is based on these hybridization probes that are designed from prior sequence knowledge, they cannot detect structural variations or discover novel transcripts. This also limits their sensitivity as they cannot detect differences in very similar sequences such as different isoforms of the same molecule.

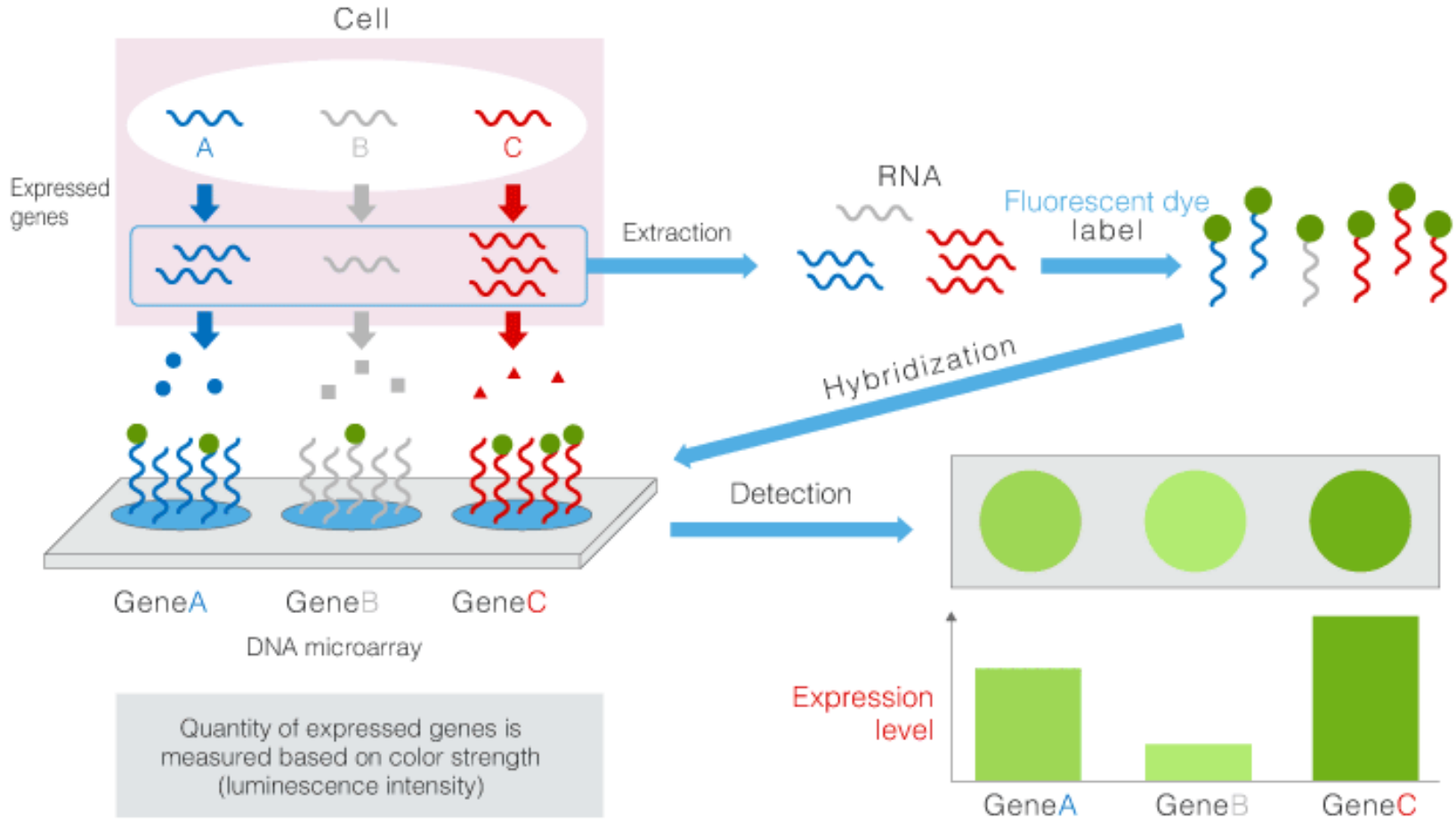


Figure 2.4 Schematic of microarray technique to demonstrate relative expression levels of genes in a sample. Image from http://www.3d-gene.com/en/about/chip/chi_003.html

RNA Seq (RNA Sequencing)

The development of novel high throughput DNA sequencing has also provided a new method for both mapping and quantifying transcriptomes. This method termed RNA sequencing improved on its forerunners of Sanger sequencing and tag-based methods.

A population of RNA is converted into a library of complimentary DNA fragments with adaptors attached to one or both ends (Figure 2.5. Step a4). Each molecule is then sequenced in a high throughput manner to obtain short sequences from one end (single-end sequencing) to both ends (paired-end sequencing). The reads are typically 30-400 bp depending on the sequencing technology used. Following sequencing the resulting reads are aligned to a reference genome or reference transcripts or assembled *de novo* (Wang et al. 2009). One advantage that it offers over hybridisation methods is that it is not limited to detecting transcripts that correspond to existing genomic sequences. This is useful when the reference genome remains to be determined.

Another advantage is that the background signal compared to microarray is very low because sequences can be unambiguously mapped to unique regions of the genome. RNA Seq. is also highly accurate in quantifying levels of expression (Nagalakshmi et al. 2008).

RNA Seq. is the first sequencing-based method that allows the entire transcriptome to be surveyed in a very high throughput and quantitative manner. Unfortunately, it remains expensive for large numbers of samples or complex genomes. The sequence coverage, or the percentage of transcripts surveyed, has implications for cost. In general, the larger the genome, the more complex the

transcriptome, or to detect rare variants a greater sequencing depth is required for adequate coverage and the more expensive the process.

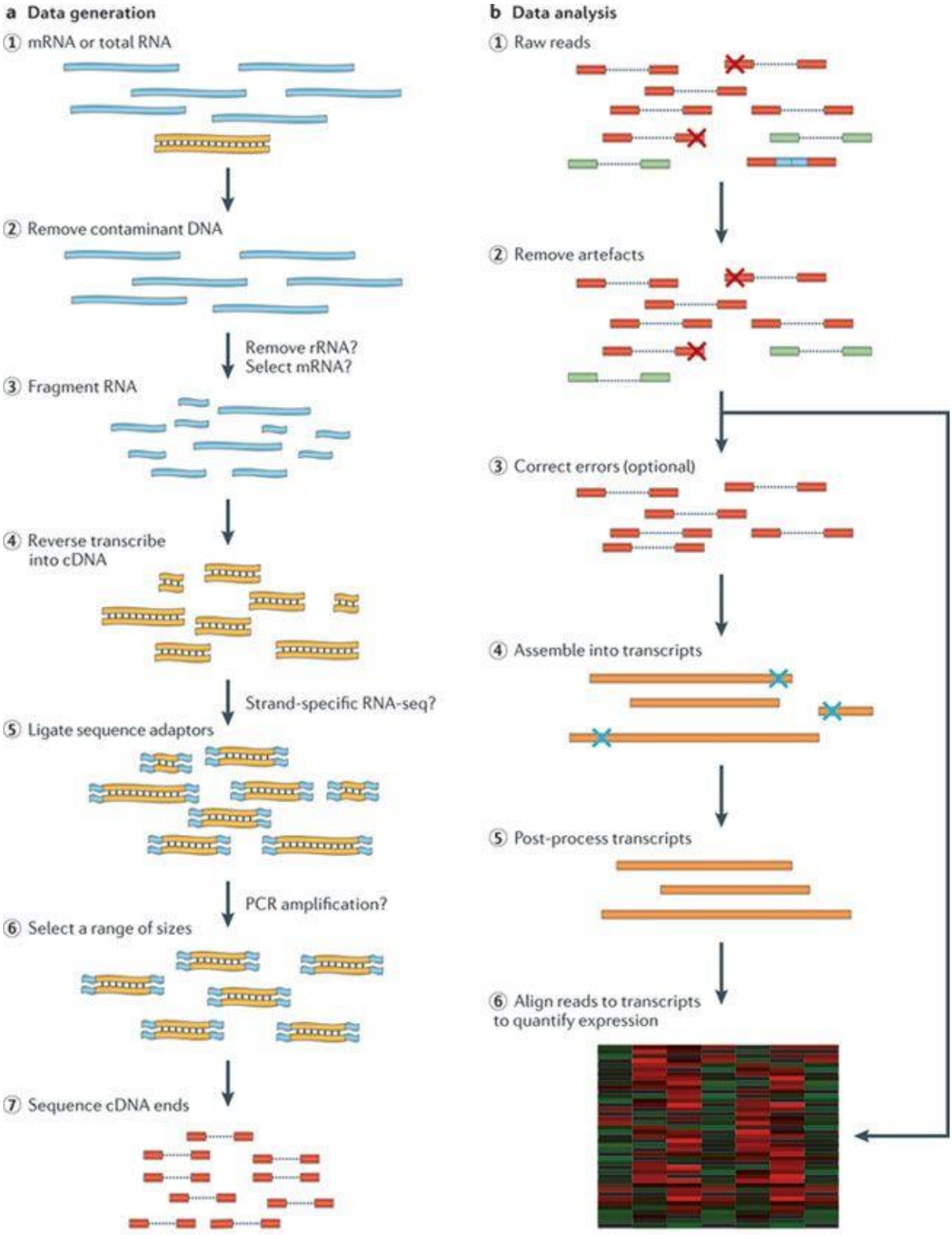


Figure 2.5 The data generation and analysis steps of a typical RNA-seq experiment. Image was obtained with permissions from "next generation transcriptome assembly" Jeffrey A. Martin, Zhong Wang. Nature Reviews Genetics. 2011 Sep 7;12(10):671-82.

Transcriptomics for spontaneous PTB

A systematic review of transcriptomics as applied to preterm birth (Eidem et al. 2015) showed that even though there have been at least 134 genome-wide transcriptomic studies of pregnancy and PTB, spontaneous PTB was only investigated in 7% of all studies and 18% of preterm birth studies. The majority, 76% of studies focussed on medically indicated PTB and specifically, pre-eclampsia.

Placental tissue was analysed in 61% of studies which has limited utility in a predictive setting as it is unavailable for testing without risk to the pregnancy (Eidem et al. 2015). Obtaining a placenta post-partum may give key information about the pathophysiology of preterm and term labour, however, there is a risk that the genes expressed once the placenta is no longer required for its biological function are likely to be different to those expressed in utero and may provide misleading biological information. Additionally, the other most frequent tissues to be studied include myometrium and fetal membranes, which are also usually only accessible at delivery. The advantage of studying these tissues are that they are directly involved with the pregnancy and labour, however, like the placenta, the gene expression at or following delivery may give useful knowledge about the process of labour but limited insight to predictive markers of labour.

The evidence for treating PPROM as a separate clinical entity to sPTL is growing as differences in gene expression are identified between these conditions. The most well replicated genes in the phenotype of PPROM (i.e. at least 2 gene expression studies) are listed in Table 2.2 below adapted from Eidem *et al.* (2015). In the following section, some of the most recent transcriptomic studies using different tissue types are discussed.

Table 2.2 List of genes found in two or more gene expression studies

Entrez Gene ID	Gene	Function	Number of studies
972	CD74 molecule	Regulates antigen presentation for immune response	2
6280	S100A9	Calcium binding protein; cell cycle progression and differentiation	2
3576	CXCL8	C-X-C motif chemokine ligand 8 (IL8); mediator of inflammatory response	2
7805	LAPTM5	Transmembrane receptor associated with lysosomes	2
23574	PRG1	P53 responsive gene 1	2
1117	CH13L2	Chitinase 3 like 2 protein is involved in cartilage biogenesis, various isoforms.	2

Myometrial transcriptomics

The first study to examine RNA sequencing data (Chan et al. 2014) in human myometrium samples obtained at term caesarean section following labour (n=5) and non-labouring term caesarean sections (n=5) demonstrated that transcriptomics separated these groups on differentially expressed genes (DEGs). However, the numbers used in this study were very small. Reassuringly, the identified genes were broadly concordant to those differentially expressed genes identified in previous microarray experiments (Havelock et al. 2005; Bukowski et al. 2006; Mittal et al. 2010). This study by Chan *et al.* (2014) added information on transcript abundance, microRNAs, splice variants, and transcript isoforms. A decrease in progesterone receptor transcripts concomitant with a decrease in FOXO1 mRNA was observed. FOXO1 regulates PGR signalling by altering PGR target genes during transcription in human endometrial stromal cells (Takano et al. 2007). Overall, the data shows that labour is associated with an inflammatory signal and genes or pathways related to immune response, chemotaxis, and cytokine signaling. Migale *et al*. (2016) used this

human sequencing data to compare orthologous genes to the expression of mouse genes from murine models. The three models used were a term gestational model, a sPTB model induced by injection of Escherichia coli LPS serotype O111:B4 (i.e. induced by inflammation) and a model that induced sPTB by the injection of RU486 (induced by progesterone withdrawal). Interestingly, they found that the changes in human myometrial transcriptome at term most closely resembled transcriptome changes of the *preterm* mouse model induced by inflammation, which suggests a dominant role of inflammation in human labour irrespective of gestation.

It must be taken into consideration that the human myometrial samples are taken at the point of artificial trauma to the uterus at CS and may be a transcriptomic reflection of an acute inflammatory reaction in the myometrium at the site of the incision. However, in support of the theory that human labour is an inflammatory state, similar findings of inflammatory signatures are being detected in other studies of term and preterm placenta (Sharp et al. 2016) and in alternative tissue samples such as maternal decidua (Rinaldi et al. 2017).

To identify the core genes and regulatory networks facilitating the transition of the myometrium from a quiescent to active labouring state Stanfield *et al.* (2019) performed an integrated analysis using two existing transcriptomic datasets (Mittal et al. 2010; Chan et al. 2014) and a dataset from RNA Seq analysis of myometrium from CS before and after the onset of labour (NCBI Gene Expression Omnibus; GSE80172). One hundred and twenty-six genes were significant across all databases and machine learning models exhibited high reproducibility between studies. This is demonstrating better classification and characterisation of myometrial activation during labour. Parturition-signalling networks were created using differential

expression data for non-labouring, early labour and active labour again, attesting to the importance of inflammation in the onset of labour. (Stanfield et al. 2019)

Placental transcriptomics

A major issue in the transcriptomic studies of sPTB in humans is the inability to collect healthy control placental tissue sampled at the same gestational age as placental tissue from actual preterm births. Therefore, gene expression differences identified after the standard comparison of sPTB and term placental tissue may reflect differences in both sPTB pathology *and* gestational age, and it is difficult to tease these two factors apart. To try and tackle this problem Eidem *et al.* (2016) matched gestational age of human placental sampling to a closely related species; macaque monkeys. They identified 29 sPTB specific candidate genes not thought to be related to gestational age. Selected genes overlapped with previously identified pregnancy-pathology related genes including serine peptidase, CD163 and VSIG4 that have been characterized as maternal biomarkers of pre-eclampsia. PDE2A is a gene containing a SNP associated with recurrent miscarriage and ADORA3 modulates secretion of matrix metalloproteinases and is important in the PPROM signaling pathway (Kim et al. 2008).

Human decidua transcriptomics

The maternal – fetal interface (decidua basalis and decidua parietalis) may play a key role in the onset of labour both at term or preterm. It is the anatomical site of contact between maternal and fetal tissues.

Rinaldi *et al.* (2017) removed decidua from fetal membranes sampled at birth from four groups; term and preterm, non-labouring and labouring women. The decidual lymphocytes were isolated, RNA bead chip microarray was used to evaluate gene expression change and qRT-PCR was used to validate microarray results. Like

previous findings, term and preterm birth were associated with widespread gene expression changes related to inflammatory signalling pathway. However, no change in lymphocyte subpopulations were seen between the groups and no functional data of cell populations were available. There was elevated expression of CDID in PTL decidua which may suggest activation of invariant natural killer cells (iNKT) in PTL samples.

Bukowski *et al.* (2017) attempted to solve issues of reduced power of small sampling numbers by increasing the number of samples taken from the same individual. They collected maternal and fetal blood, chorion, amnion, decidua, placenta and myometrium and identified expression profiles uniquely identifying four phenotypes. A ten by ten samples cross validation was performed to predict how well their model could correctly classify samples into the four groups; term non-labour, term labour, preterm non-labour and preterm labour. Forty-two percent of samples could be correctly classified, and this was an improvement on 25% of correctly classified samples using a random classifier. The largest differences in gene expression were observed in decidual samples with the proportion of genes with expression differences at least 1 or 2 standard deviations approximately six times greater than those seen in maternal blood. Results showed that regulation of immune pathways and immunological processes occur at the maternal-fetal interface; mainly in the decidua, chorion and amnion. Although the authors agreed to some degree with previous work stating that expression profiles of the 'term not in labour' group show local suppression of chemokines with simultaneous suppression of the NFkB inflammatory pathway; they disagreed that term labour is heavily related to inflammation. Instead they attributed the labour process to immune suppression. They theorise that such chemokine suppression prevents chemotaxis of immune

cells, such as effector T cells, from trafficking into gestational tissues such as the decidua, thus preventing the onset of labour (Bukowski et al. 2017).

Fetal membrane transcriptomics

Although the use of fetal membranes role in prediction of sPTB remains limited, transcriptomic studies of fetal membranes have been used to try and predict the neurocognitive status of the infant at 18-24 months (Pappas et al. 2015). A retrospective case-control study was conducted to examine the chorioamniotic membranes of 66 very preterm neonates (22-32 weeks) with and without neurocognitive impairment using RNA microarray. One hundred and seventeen genes were differentially expressed among neonates with and without subsequent neurocognitive impairment ($p<0.05$ and fold change >1.5). Differentially expressed genes were input into to a multi-gene model, developed to predict 18–24-month neurocognitive impairment (using the ratios of *OSR1/VWF* and *HAND1/VWF* at birth) (sensitivity = 74%, specificity = 83%) and validated on an independent dataset (n=19) (Pappas et al. 2015)

Cervical transcriptomics

The first study to examine the differences between sPTL (n=6) and PPROM (n=5) by collecting cervical biopsy samples up to 30 minutes following PTL showed distinct differences in the microarray gene expression and clear clustering effects between groups (Makieva et al. 2017). Four novel proteins with the potential to cause cervical remodelling leading to ruptured membranes were identified (PRAM1, CEACAM3, FGD3, and NDRG2) and the activity of MMP9 was found to be higher in the PPROM cervix. Prior to this study only 4% of all transcriptomic studies in term and pre-term human pregnancies utilized cervical tissue and did not look at PPROM as a distinct phenotype.

Small non-coding RNA molecules called microRNA's (miRNA) have been differentially expressed in gestational tissue such as cervix at two time points; 20-23 weeks and 24-27 weeks (Elovitz et al. 2014) and placenta collected at delivery from patients with pre-eclampsia, sPTB < 35 weeks and a term elective CS control (Mayor-Lynn et al. 2011), but several researchers have found that these differences do not extend to peripheral maternal whole blood (Elovitz et al. 2015, Knijnenburg et al. 2019).

Maternal blood transcriptomics

Heng *et al.* (2016) collected maternal blood samples at two separate time points 17–23 and 27–33 weeks of gestation in a low risk pregnant population. Interestingly, when they examined differential gene expression at both time points between the women who went on to have sPTB and PPROM (n=51) there were no differentially expressed genes. Therefore data were combined and analysed as one single group and compared to 114 term matched controls. At timepoint 1 (17-23 weeks; n = 51) there was no differentially expressed genes at a false discovery rate (FDR) <0.05 or at FDR<0.10, but at 27-33 weeks (n=47) and a FDR <0.10 there were 26 differentially expressed genes at between women who had SPTB and term delivery. It is important to note that the mean gestational age at delivery in the sPTB group was 33 weeks and 6 days, almost immediately following timepoint 2. Some of these gene changes may reflect early labour which may limit their usefulness as an early predictive biomarker.

In the same study by Heng *et al.* (2016) no significant change in any gene was detected between SPTB and term delivery. Paired data and gene set enrichment analyses provided additional evidence that inflammatory genes were consistently raised at 17–23 and 27–33 weeks of gestation in the blood of asymptomatic women

with sPTBs compared to women with term deliveries. Significantly enriched inflammatory pathways included leukocyte migration, lysosomes, NF-kB activation, pathways involving cytokines and their receptors (e.g. IL1, IL2, IL6, IFN, IL1R, TNFR2, CCR3, CXCR4 and CD40) as well as toll-like and NOD-like receptor signalling. In contrast, women with SPTBs had lower RNA metabolism, RNA processing and T cell activation (including CTLA4 pathway) compared to women who had term deliveries (n = 163 downregulated gene sets at T1, n = 100 at T2; 77 common gene sets). Therefore, despite not observing any significant gene at FDR<0.05, numerous gene sets were significantly associated with sPTB. This team hypothesized that circulating maternal leukocytes respond to 'signals' from gestational tissues and alter their gene expression. The most striking gene set enrichment result was that women who had SPTBs have increased interleukin signalling, mainly driven by IL1 and IL6, and leukocyte migration into gestational tissues as early as 18 weeks compared to women who had term deliveries. This in theory could accelerate cervical ripening by increasing local oxytocin and prostaglandin production, weakening the fetal membranes and causing early contractions. Unfortunately, this study did not obtain cervical length measurements or fetal fibronectin and the prediction models could not be compared to current clinical tools.

This study by Heng *et al.* (2016) was one of three (Bukowski et al. 2017, Heng et al. 2014) included in a recent meta-analysis of transcriptomic studies of gene expression profiling to identify gene expression differences detectable in maternal whole blood (Vora et al. 2018). Spontaneous preterm birth was defined as delivery less than 37 weeks' gestation. Combining the studies provided additional power and revealed 210 differentially expressed genes, and clear enrichment of immune

mediated pathways. Interestingly, 18 of these 210 genes also demonstrated differential expression in the second trimester, suggesting a possibility for early identification of patients who might deliver preterm. *IL-1R1* and *TFPI*, both of which encode immune-related proteins, were found to be differentially expressed and secreted longitudinally. Perhaps the most interesting finding of all was that preterm maternal whole blood showed upregulation of innate immunity and downregulation of adaptive immunity suggesting perhaps a breakdown in maternal-fetal tolerance.

2.4 Metabolomics

The study of metabolomics is the large-scale study of small molecules known collectively as metabolites. Metabolites are typically small molecules that are the intermediate products of metabolic reactions occurring within cells or biological systems, and the end-product of the interaction of the system's genome with its environment. (Rochfort. 2005) By undergoing catabolism in a complex metabolic network, metabolites can also be components of higher order biological structures (DNA, transcripts, proteins) and cell structures energy, or ATP (Dunn et al. 2011).

Although the study of metabolite profiling has a long history that started in the 1950's (Rochfort. 2005), the phrase "metabolomics" was coined by Oliver S.G. *et al.* in 1998 who identified metabolites as part of a functional analysis of the yeast genome (Olivier et al. 1998). The Metabolomics Society defines metabolomics as; '*the comprehensive characterization of the small-molecule metabolites in biological systems, which can provide an overview of the metabolic status and global biochemical events associated with a cellular or biological system'* (Metabolomics Society. 2019). There has been some overlap with the term *'metabonomics'* defined as the "*quantitative measurement of time-related multiparametric metabolic responses of multicellular systems to pathophysiological stimuli or genetic modification*" (Nicholson et al. 1999). Due to similar analysis techniques, 'metabolomics' and 'metabonomics' are frequently used interchangeably, however the all-encompassing term 'metabolomics' remains the most popular in the literature (Rochfort. 2005, Dunn et al. 2011., Smolinska et al. 2012.) and will be used from here onwards in this thesis.

Untargeted metabolomics studies are characterised by the simultaneous measurement of a large number of metabolites from each sample. This strategy is

known as a top-down approach and avoids the need for *a priori* specific hypothesis on a set of metabolites, and instead analyses the global metabolomics profile (Alonso et al. 2015). Recent advances in the technologies used to extract and analyse this type of data have revolutionised its wide range of application including biomarker discovery (Meyer et al. 2013, Armitage and Barbas. 2014, Julia et al. 2014) and the analysis of complex disease. Of note, metabolomics has been successfully used to investigate complex pregnancy conditions such as gestational diabetes, fetal growth restriction and pre-eclampsia (Horgan. 2009, Sulek. 2014, He. 2015).

Metabolites are the downstream products of gene transcription and translation, and metabolomics can give a clearer picture of a phenotype than genomics or transcriptomics. However, this layer of omics data increases in complexity as there are currently 114,100 metabolite entries on the Human Metabolome Database (HDMB) including both water soluble and lipid soluble metabolites (Wishart. 2018). Due to this complexity there is not one method that can capture all the known metabolites and to do this would require multiple techniques and instruments.

The most widely used technologies in metabolomics are nuclear magnetic resonance (NMR) due to its high precision and mass spectrometry (MS) with excellent reproducibility. Fourier transform infrared (FT-IR) is also becoming more popular in diagnostics. It has a rapid high throughput and is relatively inexpensive, but with low chemical specificity.

A chromatographic separation technique is often coupled with mass spectrometry such as liquid chromatography (LC) or gas chromatography (GC) to enhance resolution. An overview of data processing for metabolomics studies is shown in Figure 2.6.

In this section of the chapter I will review metabolomic publications that have attempted to elucidate mechanistic pathways, or predictive metabolites of sPTB using either NMR or MS using different biological fluids for analysis. A summary table is included on page 88 (Table 2.3).

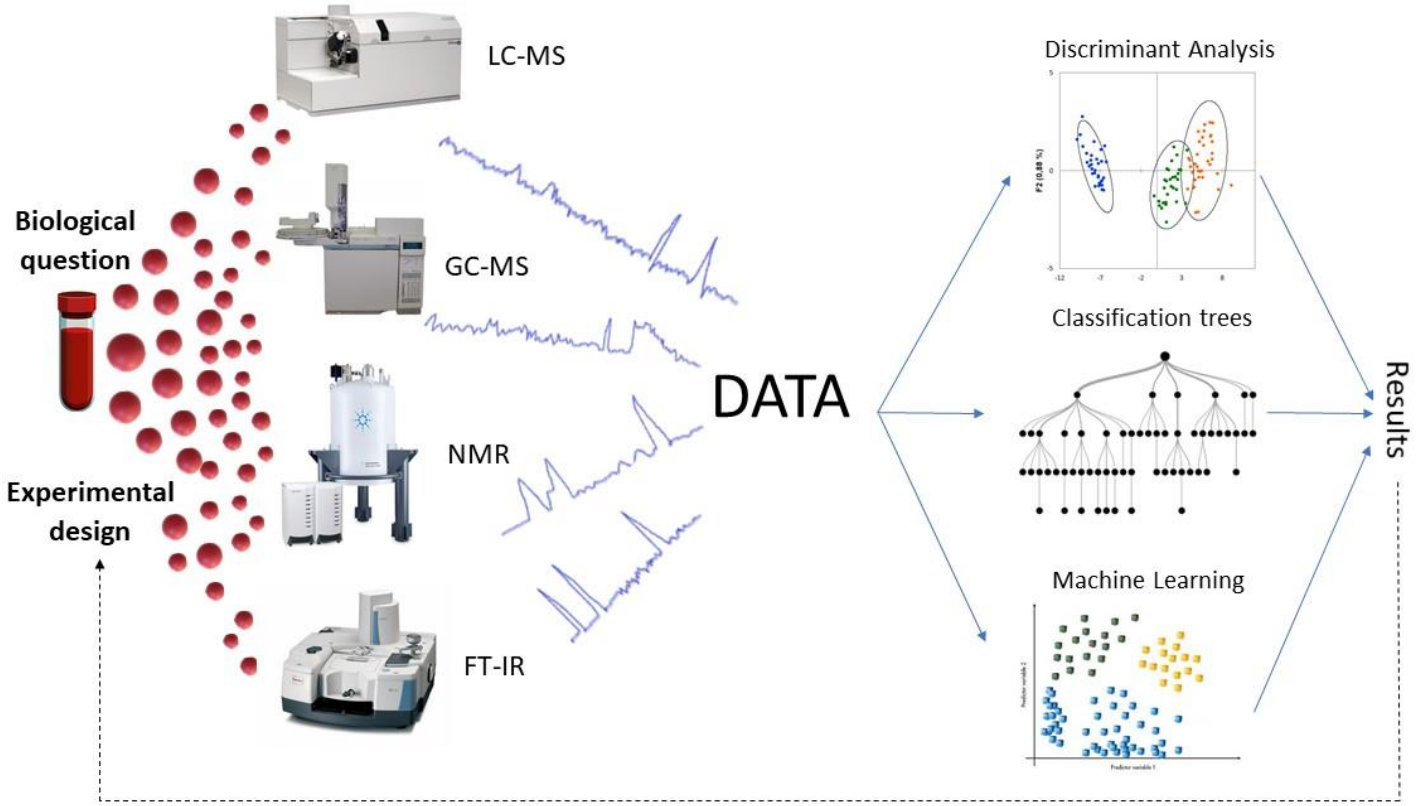


Figure 2.3 Overview of experimental design in metabolomic studies adapted from Gromski et al. 2015 LC-MS - *Liquid chromatography-Mass Spectrometry*; GC-MS - *Gas Chromatography-Mass Spectrometry*; NMR - *Nuclear Magnetic Resonance*; FT-IR - *Fourier Transform-Infra Red Spectroscopy*.

Amniotic Fluid

Two studies have looked exclusively at AF for the prediction of sPTB using two different mass spectroscopy techniques (Menon et al. 2014, Baraldi et al. 2016). One used 24 samples obtained at amniocentesis from an asymptomatic group of women between 21- and 28-weeks pregnancy and compared metabolic signalling to 8 samples obtained at term caesarean section (Baraldi et al. 2016). The other obtained AF samples during labour from 25 women <34 weeks in labour and compared metabolomic signalling to 25 term controls (Menon et al. 2014). These studies faced challenges in their study designs as AF is not an easily accessible biological fluid and gestational matched controls were not available. Samples had to be regrouped based on detection of paracetamol metabolites but both studies revealed different metabolites either raised or lowered in the preterm birth group. This lack of concordance is possibly due to the different analytical techniques used and the differences in metabolites present during labour versus asymptomatic women.

Orczyk-Pawilowicz *et al.* (2016) using NMR and Virgiliou *et al*. (2016) using mass spectrometry combined analysis of AF metabolites with the metabolites in maternal plasma and serum respectively in an untargeted approach. Their aim was to identify if the metabolite signatures of AF were correlated in maternal blood, which is a much easier and more accessible biological fluid to study. In the study conducted by Orczyk-Pawilowicz *et al.(*2016) the metabolomic profiles of plasma and AF samples of healthy women with normal pregnancies in the second trimester and three timepoints in the third trimester were studied. The highest correlation was observed between AF and plasma in the transition from the second to third trimester. However only 8 samples of plasma and 7 samples of AF were available for the second trimester group.

Interestingly, Virgiliou *et al.*(2016) only sampled in the second trimester but observed no correlation between maternal serum and AF based on the most commonly detected metabolites. Maternal serum was recommended for future metabolomic profiling studies due to ease of access and repeatability of sampling. The Virgiliou *et al*. (2016) study included 35 women delivering between 29 weeks and 36 weeks and compared metabolite profiles to 35 women delivering at term. Using partial least square discriminant analysis (PLS-DA) there was good separation between the groups, which was attributed to 13 lipid features. Despite a well conducted study, the author did not attempt to validate the results. The groups were also close in gestational age and there was only one week between the latest preterm delivery and the earliest term delivery; and four weeks between the average gestations of labour (35 weeks cf. 39 weeks). This tells us little about clinically important preterm births (<34 weeks) but suggests that lipids may be an area of interest for further research. Samples were obtained from healthy patients requiring amniocentesis or blood tests to rule out other pathology. Unsurprisingly the resulting populations used in the study are heterogenous, with varying number of samples taken from different trimesters in the study with only a small subset with paired plasma and AF samples from the same woman (n=50; T2=1, T3=15, term=26, prolonged pregnancy=8). The AF was obtained through both transabdominal amniocentesis and transvaginal amniotomy. These two different collection methods would be subject to different contamination risks, particularly transvaginal amniotomy occurring at term delivery.

From these studies different trends were observed in the two different biofluids of serum and plasma. Very early pregnancy elevations in total cholesterol and levels of triglycerides were associated with a 2.8-fold increased risk for preterm

birth before 34 weeks and a 2-fold increase risk for PTB between 34 and 37 weeks. Pyroglutamic acid and tryptophan were found to be in lower levels in the maternal serum of women with preterm labour.

Pyroglutamic acid is derived from glutathione and decreased levels suggest a potential glutathione deficiency. A deficiency could result in a reduced ability to neutralise toxins and defend against oxidative stress.

Tryptophan is an important precursor for production of bioactive metabolites such as serotonin. Serotonin and tryptophan levels have been linked to depression in pregnancy and preterm delivery (Waters K. 2010). Tryptophan may also be a marker of inflammation through activation of the kynurenine pathway. Metabolism is catalysed by the IDO enzyme (indoleamine 2,3-dioxygenase) and acts as an endogenous regulator of T-cell proliferation. Inflammation leads to more IDO activity and expression and increased metabolism of tryptophan resulting in lower levels in maternal serum.

Urine

NMR was the method employed to examine a nested case-control group from the Rhea mother-child cohort (Maitre et al, 2014). Samples of urine were obtained between 10-14 weeks, from 1317 women in Crete, Greece. Eighty-eight women had sPTB and 26 women had medically induced preterm births (IPB), all defined as birth less than 37 weeks gestation. Samples from 288 healthy controls were also collected.

Thirty-four urinary metabolites were identified from spectra and two methods of univariate analysis was performed to select likely candidate metabolites to subject to multivariate regression analysis. The analysis of PTB < 37-week outcomes was conducted both on the combined clinical subtypes (PB) and separately on each subtype (sPTB and IPB). Formate, N-methyl2-pyridone-5-carboxamide (2-Py),

glycine, TMAO, lysine and the singlet at 0.63 ppm (a steroid conjugate) significantly varied between sPTB and control groups ($p<0.05$). However, only three of the six candidate metabolites for sPTB; formate, lysine and the singlet at 0.63 ppm, showed a significant trend ($p<0.05$) when the trend of these metabolites in the proportion of women with PTBs was examined (dataset split in quartiles of metabolite levels to investigate dose-response and sPTB). When a logistic regression model was used to account for other factors such as maternal age, education, parity and smoking habits, high lysine and low formate levels were significantly associated with a higher risk of sPTB. (Maitre et al. 2014)

High levels of tyrosine, acetate, trimethylamine and formate were also significantly associated with a decreased incidence of fetal growth restriction (birthweight below 10^{th} centile) (IORs between 0.27 and 0.14). Whereas, high levels of N-acetyl glycoproteins were associated with an increased risk of iatrogenic PTB.

It is very difficult to compare or validate these findings with other metabolomic studies due to the differences in analytical platform and biofluid chosen. One other group looked at both maternal urine and plasma in different trimesters of healthy pregnancy and compared this to the non-pregnant state (Pinto et al. 2015) but no cases of sPTB were examined. As expected, levels of some amino acids required by the fetus decreased in the first trimester, but novel findings included early changes in citrate, lactate, and dimethyl sulfone levels. This is possibly due to a metabolic energy shift in the first trimester. Alteration in creatine levels was also noted, along with creatinine changes. Plasma high density lipoproteins (HDL) and low and very low-density lipoproteins (LDL + VLDL) levels were confirmed to increase throughout pregnancy, but at different rates and

accompanied by increases in fatty acid chain length with increase in lipolysis in the third trimester.

Metabolites in Maternal Blood

Plasma metabolites were examined using mass spectrometry from women who had preterm labour between 24 and 37 weeks of gestation (n = 57), threatened preterm labour but delivered at term (n=49) and samples collected at normal term delivery between 38 and 41 weeks (n=25) (Lizewska et al. 2018). PLS-DA models differentiated preterm and term births based on metabolomic profiles alone. This may be partly due to differences in gestational age across groups, but this was somewhat mitigated by including a threatened preterm labour group that were sampled at a similar gestation as the preterm labourers but when on to deliver at term. Fatty acids were found to be the group most significantly different but based on the results of the study by Pinto *et al*. (2015) this may not be wholly unexpected in view of the fact that there are significant gestational age changes with this group. In a study that took samples from women at delivery, higher levels of omega-3 fatty acid, docosahexaenoic acid (DHA), were seen in the preterm group when compared with a threatened preterm group (who went on to deliver at term) with no difference between preterm and term groups suggesting that detection of this fatty acid metabolite may occur with initiation of labour. Lower levels of amino acids were seen in women with preterm labour compared to threatened preterm labour possibly associated with a source of energy or a link to oxidative stress. Lower tryptophan in the PTB group had the highest statistical significance of change compared to threatened PTB (delivered at term) and term labour groups and has been associated as an inducer of oxidative stress (Elisia et al. 2011). However, it is difficult for this study to truly assess the differences in metabolites between PTB and gestational age

as the preterm birth group on average were sampled at 30 weeks and the threatened PTB group were sampled at 32 weeks.

Souza et al. (2019) did not have this challenge with their study design as they analysed samples stored from the "SCOPE" cohort, an international pregnancy biobank serving to predict novel biomarkers for complications of pregnancy. Samples taken at 15 weeks and 20 weeks of pregnancy were analysed by GC-MS for 164 nulliparous women from Cork, Ireland (sPTB <37 wks., n=55) and 157 nulliparous women from Auckland, New Zealand. Decane, undecane and dodecane, belonging to a class called alkanes, were significantly associated with sPTB (FDR <0.05) at 20 weeks in the Cork subset but not in the Auckland cohort. The observation of elevated alkanes in the maternal serum of women who had sPTB was also linked by the authors to a possible oxidative stress response. However, this association was not seen in the most clinically important sPTB <34 weeks nor confirmed in the Auckland cohort, therefore their confidence remains limited in this finding. (Souza et al. 2019)

A nested case-control study of samples collected prospectively from 305 women attending with reduced fetal movements (RFM) in the third trimester of pregnancy examined the serum from 40 women with poor pregnancy outcomes using ultra performance liquid chromatography mass spectrometry (UPLC-MS) (Heazell et al. 2012). This was made from a composite of sPTB with normal weight centiles (n=3), preterm SGA (n=8), term SGA (28) and unexpected admissions to SCBU (n=1). The controls were taken from the same cohort and included women that had a normal birthweight infant with no perinatal complications (n=40). Principal component analysis (PCA) could not separate cases and controls. Univariate analysis demonstrated that most classes of metabolites (glycerolipids, glycerophospholipids,

fatty acids/organic acids and vitamin D metabolites) showed a trend towards decreased concentrations in the poor pregnancy outcome individuals. Two progesterone metabolites were significantly downregulated in the poor pregnancy outcome group; 17 hydroxy pregnenolone sulphate and pregnanediol-3-glucuronide, intermediates of progesterone production. It is possible that there is downregulation of the entire progesterone pathway in poor pregnancy outcomes. This study is largely biased by a fetal growth restriction population and it is unclear how to apply these results to sPTB, particularly as samples have been collected in the third trimester.

The previously mentioned study by Virgiliou *et al.* (2016), took plasma samples from 35 women delivering at a mean of 35 (29 − 36 + 5) weeks in labour and compared them to 35 term controls, however all patients underwent amniocentesis. This group used lipid profiling and discovered that 13 lipid features contributed significantly to group separation, but the features could not be identified with their methodology. Using targeted profiling 54 metabolites were detected in maternal serum samples, using Orthogonal projections to latent structures-discriminant analysis (OPLS-DA) the term and preterm groups showed separation based on their metabolomic profiles and 8 metabolites in the serum showed significance following unpaired t tests ($p < 0.05$). Pyroglutamic acid was found to be higher, while hypoxanthine and tryptophan were found to be lower in PTB samples. Using a second set of samples to confirm performance of serum predictors the results were found to be the same with the addition of choline, contributing to the classification of the two study groups in serum. This is the first study of maternal blood biomarkers that have collected blood prospectively between 14 and 23 weeks.

Table 2.3 has summarised all significant metabolites when comparing preterm +/- associated pathologies such as fetal growth restriction to a term control in

a PTB cohort irrespective of whether the metabolite was increased or decreased relative to the control group and irrespective of the methodology; NMR or MS based techniques. The only study that achieves sampling at the same gestation in both the case and control groups is Virigliou *et al* (2016), and many of the top metabolites found in the other studies but not this one may be reflecting gestational age changes.

Table 2.3. Summary of metabolites identified in metabolomic studies of preterm labour. Metabolites that have been identified in four or more studies are in bold.

Author	**Baraldi**	**Heazall**	**Lizewska**	**Maitre**	**Menon**	**Romero**	**Souza**	**Virigliou**
Samples (total)	n=31	n=80	n=131	n=363	n = 50	n=168	n=109	n=70
PTB samples	n=21	n=11	n=57	n=88	n = 25	n=112	n=55	n=35
Mass Spec.	✓	✓	✓		✓		✓	✓
NMR				✓				
Amniotic Fluid	✓				✓	✓		✓
Maternal Blood		✓	✓				✓	✓
Maternal Urine				✓				
Timepoint collected	21-28 wk	RFM; 28-40 wk	Admission; 24-40 wk	T1*	Delivery; 29-40 wk	22-33 wk	T1 = 15 wk T2 = 20 wk (+/- 1wk)	14-23 wk
Alpha sorbopyranose						✓		
Amino acid chain	✓							
Acetate				✓				
Acetaminophen metabolites					✓ (7)			
Alanine				✓				
Beta hydroxyl phenylethylamine						✓		
Biliverdin					✓			
Catechol						✓		
Cholesterol						✓		
Citrate				✓				
Decane							✓	
Dodecane							✓	
Eicosanoic acid						✓		
Fatty acids	✓	✓	✓(11)		✓(18)			
Formate				✓				
Fructose						✓		
Galactose						✓		
Glutamine						✓		✓
Glycerol						✓		
Glycerolipids		✓						
Glycerophospholipids		✓						
Glycine				✓		✓		
Heptanedioic acid						✓		
Hexose cluster						✓(3,5,6)		
Histidine			✓		✓			
Hydropyridine	✓							
Hypoxanthine								✓
Inositol						✓		✓
Isoleucine						✓		
Lactate				✓				
Leucine				✓		✓		
Lysine			✓	✓				
Mannose						✓		
Methionine						✓		
Methyladenine						✓		
Muconic dialdehyde	✓							
N-acetyl glutamine						✓		
N-methyl-2-pyridone-5-carboxamide				✓				
Phenylalanine						✓		
Phenylacetylglutamine				✓				
Phosphatidylcholine	✓							
Prostaglandin		✓						
Progesterone					✓			
Pyroglutamic acid								✓

Pyruvate								✓
Salicylamide						✓		
Succinate						✓		
Steroid conjugate – 0.63				✓				
Theophylline					✓			
Trimethylamine				✓				
Trimethylamine-N-oxide				✓				
Tryptophan		✓	✓			✓		✓
Tyrosine				✓				
Undecane							✓	
Vitamin D		✓						
3-methoxybenzene propanoic acid	✓							
4-hydroxynonenal alkyne	✓							
1-methylurate					✓			

*T1 – end first trimester

2.5 Classification of Births

It is important to be precise and consistent in defining the phenotype for a study population. From the omic studies discussed in this chapter there is wide heterogeneity between classifications of sPTB and the gestational ages of patients included in studies. This is a hurdle for direct comparison of studies and makes meta-analysis of multiple studies arduous.

The complexity of defining clear phenotypes of sPTB are discussed in a series of papers published in 2012 (Goldenberg et al. 2012, Kramer et al. 2012, Villar et al. 2012.). The authors of the articles in this series are recognised experts in PTB research across USA and Canada. They were brought together as a direct result of the Global Alliance to Prevent Prematurity and Stillbirth (GAPPS) meeting to define a prototype classification system for PTB for general consideration.

The issues considered in their published discussions included difficulty defining the lower gestational age of PTB definitions. Frequently the clinical definition of a PTB is based on the potential for a 'livebirth' and excludes births below 23 weeks or other arbitrary gestational week, classifying this group as 'spontaneous miscarriage'. As causes of births between 16-22 weeks do not differ substantially from those after 22 weeks and the authors felt that there was no reason to exclude them from a classification system.

The authors' consensus was that the actual method of delivery (e.g. assisted delivery or caesarean section) should not be included in the phenotype. The focus of the clinical phenotype should derive from features of pregnancy and spontaneous labour, such as short cervix, bleeding and ruptured membranes. Risk factors such as low socioeconomic status and smoking were also considered unhelpful, and although

data were recommended to be collected in a systematic way, it should not be part of a phenotypic classification system.

The issue of potential causes such as stress and assisted reproductive technologies (ART) which are included in some classifications were considered but it was felt unless a condition can be clearly defined and there is a clear pathway to PTB causation it should not be included in a classification system and should also be reported as risk factors.

A controversial area the authors covered was the issue of stillbirths. The preference of the authors was to include all births above a lower gestational age threshold for PTB, whether it was a termination or not. They felt "a system that includes some terminations but not others would likely be confusing for all" (Goldenberg et al. 2012). Although it is important to keep classification methods simple and understandable, I disagree with this rationale. My opinion is that it remains important that any birth included in the classification system, whether the infant was born alive or not, should be subject to the same phenotyping definitions and classifications. Intrapartum and antepartum stillbirths should be included if the signs and symptoms of preterm labour are present, e.g. spontaneous contractions or ruptured membranes. However, if there is no evidence that the parturition process has started (i.e. no fluid leakage, no contractions or bleeding) and there is no likelihood that the baby would have delivered if not for the intervention of the obstetric team, it should be defined as *care giver initiated* or *iatrogenic PTB*. The same criteria would be applied to any livebirth or antepartum stillbirth. The key for the definition, in my view, is that the event of labour itself occurred spontaneously. To illustrate my point, if a baby with severe growth restriction was delivered prematurely due to concerns of fetal health and risk of demise; this would be

classified as *care giver initiated.* If in another pregnancy, very early growth restriction occurred resulting in an intrauterine death before obstetric intervention could occur, the subsequent induction of delivery should also be classed as *care giver initiated*. By combining all stillbirths into a definition of spontaneous PTB, more heterogeneity of pathology is included under the umbrella term of sPTB. An increasingly puristic definition will only enable the scientific community to get closer to identifying true biological pathways and reducing "noise".

Despite wanting to avoid confusion, the authors recommend dividing iatrogenic deliveries into three or four groups of "urgent", "discretionary", "iatrogenic" or "social". This seems to be largely unhelpful as these decisions are based on many factors including human decision making which itself is a very complex process. For the purpose of a sPTB classification, I feel that care giver-initiated birth classifications are outside of this remit and should be identified as simply a single group of "care giver initiated". Despite these areas of disagreement in approach, the authors make one thing clear; no matter what classification is used, it should be well defined with no ambiguity on how to classify difficult cases (e.g. threatened preterm labourers who are ultimately augmented).

The authors subsequently published a prototype classification system based on their considerations and divided the phenotype into maternal, fetal, placental and signs of parturition (Villar et al. 2012). Manuck *et al.* (2015) attempted to use this classification system to retrospectively classify 1025 women delivering <34 weeks for whom data had been collected prospectively during a case-control study. They claimed they had successfully classified women into 9 sPTB phenotypes: 1) infection/inflammation, 2) decidual haemorrhage, 3) maternal stress, 4) cervical insufficiency, 5) uterine distension, 6) placental dysfunction, 7) premature rupture of

the membranes (PPROM), 8) maternal comorbidities, 9) familial factors. Despite a committed effort to apply a strict and well-defined classification system for phenotyping, there are certain limitations that makes this design difficult to translate across research studies. Firstly, 4% of that cohort had no evidence of any described phenotype. They were essentially unclassifiable, and this is a finding that cannot be ignored. The authors could have created a 10th category for this group of women such as 'unexplained PTB'. An additional 78% met the criteria for more than one phenotype which makes women hard to put into mutually exclusive groups for comparison. Unsurprisingly, in this attempt at using the prototype classification system by Manuck *et al*. (2015), more than half of women had evidence for strong, moderate or possible evidence of maternal stress. All clinical data including 'perceived stress' was collected in the month following a PTB <34 weeks and is not reflective of stress levels during pregnancy. As prediction of PTB remains poor, it is hard to record self-reported stress, unless this is done prospectively. Evidence of biological stressors should be examined through omics signatures, rather than trying to include this in a classification system. As there is currently no clear mechanistic pathway to causation, this should be reported as a risk factor rather than being an integral part of the classification system.

Most women in this cohort did not have antenatal cervical screening. The criteria for classifying "cervical insufficiency" was split into strong, moderate and weak. Moderate evidence for cervical insufficiency was "cervical length <1.5cm prior to 28 weeks". Contrary to this assertion, this evidence should be viewed as exceptionally weak evidence for cervical insufficiency, if evidence at all, particularly in a population without previous PTB. To illustrate this point, evidence from a randomised control trial of over 47,000 low risk women who were screened for a

short cervix and randomised to expectant management or cerclage showed that only 24% of women with cervix <15 mm delivered prior to 33 weeks. Therefore, although short cervix is certainly a recognised risk factor, it is insufficient to imply causation. Additionally, true cervical insufficiency has an unclear and often inconsistent definition. Overall, it is thought to apply to a very small group of women in the region of 1% of PTBs and generally presents before 24 weeks of pregnancy (ACOG practice bulletin No. 142. 2014). The American College of Obstetrics and Gynaecology defines cervical insufficiency as "the inability of the cervix to retain a pregnancy in the absence of contractions or labour or both". The criteria defined in this study for cervical insufficiency does not make it exclusive of contractions and accounts for 14.4% of deliveries with a mean gestational age at delivery of 29.5 weeks (+/-SD of 3.3). As cervical shortening and effacement must also occur in labour, this definition only serves to confuse these terms further.

Therefore, an alternative classification system based on the considerations of these proposed classification systems of phenotyping will be used for my study and this is outlined in Chapter 3.

2.6 Methodologies for Integrating Omics

Evaluating each 'omic' data type individually before integrating data is essential, as each data type poses unique challenges. Individual analysis methods, including quality control performed for each omics layer, will be presented in the relevant omics chapter. The methodologies to integrate the data layers will be discussed here.

A review of omics data integration by Ritchie *et al.* (2015) broadly categorized all methods of data integration into two types of approaches; multi-staged analysis and meta-dimensional analysis. (Figure 2.7)

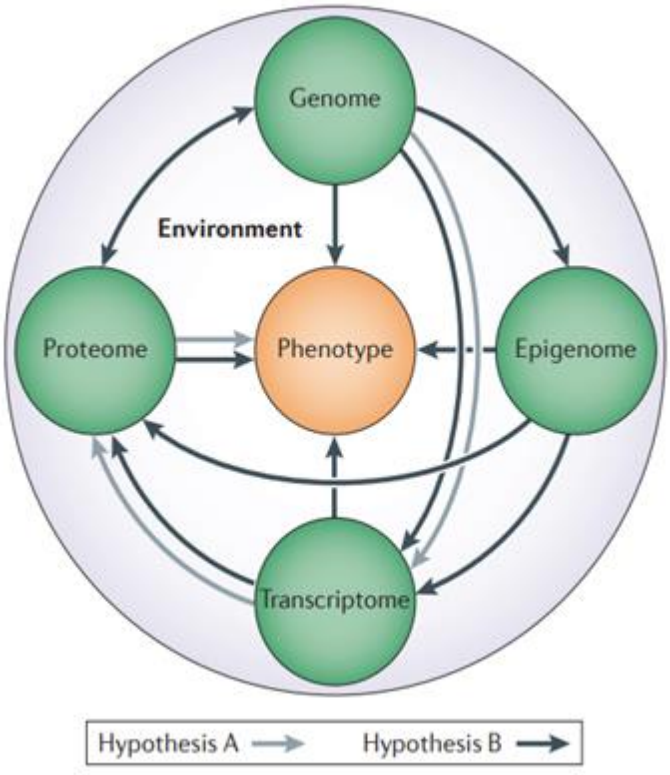


Figure 2.7. Alternative hypothesis of complex trait aetiology. Hypothesis A (grey arrow) is the theory that variation is hierarchical, such that variation in DNA leads to variation in RNA and so on in a linear manner. Hypothesis B (black arrow) is the idea that cross talk across omics layers leads to a phenotype.

Multi-Staged Analysis

In multi-staged analysis, predictive models are constructed in a stepwise or hierarchical manner reflecting hypothesis A of Figure 2.7. The analysis is divided into multiple steps and use genetic variation as the foundation of all the other omic variations. Steps usually include filtering SNPs associated with the disease on a genome-wide significance level, and then testing these SNPs for association with another level of omic data. For example, using transcriptomic data or gene expression levels. These SNPs are known as expression quantitative trait loci (eQTL). Alternatively, methylation QTLs (mQTLs) when SNPs are associated with DNA methylation levels or metabolite QTLs if associated with metabolite levels. These data are then used to test for correlation with the phenotype of interest.

An example of such an integrative analysis is performed by Knijnenburg *et al.* (2019) who integrated molecular and clinical data for 629 families. Whole genome sequencing, mRNA sequencing, miRNA sequencing and DNA methylation profiling were carried out on maternal whole blood from the same individuals. After performing a genome wide statistical test to identify candidate PTL genes, eQTL and mQTL analyses were performed to identify genes and methylation probes that overlapped with the candidate genes. Their study did not replicate the findings of the previously mentioned GWAS by Zhang *et al*. (2017) at a genome wide significance level of 10^{-8}, but at a less stringent level could identify four of the same genes (albeit with different SNPs). Specifically, for the very early PTBs (<28 weeks) various significantly associated variants were uncovered as well as differentially expressed and methylated genes, many of which are involved in growth factor signalling, inflammation and immune related pathways. However, in this study maternal blood samples were taken in the four days following delivery so this may reflect

transcriptomics of preterm labour or involution of the uterus rather than any causative factor for sPTB (Knijnenberg et al., 2019).

However, omic integration can occur on separate populations when samples cannot be obtained from within the same populations. Brubaker *et al.* (2016) disappointed with the poor results of multiple GWAS studies of PTB, examined top scoring SNPs just below genome wide threshold significance from PTB GWAS. They performed an integrative protein-protein interaction (PPI) network analysis including candidate genes associated with the relevant PTB-SNPs and myometrial tissue transcriptome data to try to identify networks of genes and proteins regulating the onset of labour. The genomic and transcriptomic data was obtained from different datasets but examined the same preterm labour phenotype. Six hundred and twenty nine biological or cellular processes were enriched using the PTB-SNP data from a mixed cohort of 3,485 mother-child pairs, but when this was refined through functional mRNA expression data from a total of 22 tissue expression datasets (from either term or preterm, and both labour and non-labour myometrial tissue samples), 38 significant subnetworks associated with preterm labour were found and 22 networks associated with term labour. The authors concluded that TWIST1, MEF2C, PLA2G4C and LGALS2 may be worthy of further investigation as the MEF2C-LGALS2 and MEF2C-PLA2G4C term subnetworks were the only ones noted to be dysregulated in the preterm myometrium. In term labour all these genes were downregulated, whereas in preterm labour they were upregulated by qRT-PCR plus there was evidence of coregulation of these genes. TWIST1 acts to modulate downstream genes and is a repressor of MEF2C, it carries a PTB associated SNP, and is part of the negative feedback loop for the cytokines TNF-a and IL-1B in the

NF-KB signalling pathway. Biological plausibility was offered for PLA2G4C as it is a phospholipase which regulates prostaglandin synthesis independent of oxytocin.

Meta-Dimensional Analysis

In meta-dimensional analysis all omics data are combined simultaneously to create a multivariate model associated with a given outcome, reflecting hypothesis B of Figure 2.7 (Ritchie et al. 2015). This allows for combinations of heterogenous datatypes, including clinical data and assays that cannot be mapped back to a specific gene. The three main approaches for this type of analysis are concatenation-based integration, transformation-based integration and model-based integration (Ritchie et al. 2015).

Concatenation Based Integration

In concatenation-based integration all available datasets are merged into a single matrix for modelling. An advantage of this method is that once a data matrix has been created, then multiple statistical tests or machine learning methods for modelling can be applied to the data for analysis. The challenge for this type of integration is identifying the best way to combine different types of data in a meaningful way, whilst avoiding bias from certain data types and ensuring it is computationally feasible (Ritchie et al. 2015).

Transformation Based Integration

Here data is combined only after the original data has been transformed into an intermediary state, such as a kernel or graph matrix. This is a matrix that represents relative positions of all samples by valid graph or kernel functions. In short, kernels allow transformation of randomly distributed or linearly inseparable data to linearly separable ones by increasing the dimensional space of the data points. This approach

is robust to different measurement scales, but some data-specific properties may be lost when integrating them. As data is transformed independently, it may make it difficult to detect some of the effects between e.g. SNP/gene and metabolite interaction if the transformation removes the ability to detect the original interaction (Ritchie et al. 2015).

Model-Based Integration

In this method each omic data type is used as a training set to develop a model. A final model is then created from these training set models (Figure 2.8) which preserves the data specific properties. However, these methods of model integration are well known for overfitting and is suitable if the data types to combine are extremely heterogenous and the other methods of concatenation and transformation are not suitable.

To my knowledge, Ghaemi *et al.* (2019) have performed the only study that has attempted to do this type of omics combination in pregnancy. They compared multivariate predictive modelling using an Elastic Net (EN) algorithm to predict gestational age. Using stacked generalisation, the individual omics models were then combined into a single model. Three hundred and fifty seven samples taken at 51 separate timepoints from seventeen women with term pregnancy were analysed for seven omic subtypes; cell-free transcriptomics, antibody-based cytokine measurements in plasma and serum, microbiome analyses (of vaginal swabs, stool, saliva and tooth/gum), mass cytometric analyses of whole blood, untargeted metabolomics and targeted proteomics analysis of plasma (Ghaemi et al. 2019). Their multi-variate analysis algorithm (EN model) with cross-validation steps (leave-one-subject-out) were compared to the other machine learning algorithms of Random Forest, Gaussian Process, Support Vector Regression and XGboost. Their model

increased their predictive power by combining all the datasets and revealed novel biological interactions. A strong relationship between pregnanolone sulphate and the behaviour of the NF- KB myeloid dendritic cells and regulatory T cells was revealed by the model. These play a critical role in feto-maternal tolerance. Although modulation of immune cell function by progesterone and its derivatives is not a new concept, the specific immune cell subsets of this action are not well understood. Additionally, a strong interaction between the transcript of protein factor CSH-1 (cell-free RNA) and STAT5activity in CD4+ T cells, may suggest that CSH-1 may directly activate the JAK2/STAT5 signalling pathway in CD4+ and CD8+ subsets in pregnancy. However, the limitations of this study are that the number of subjects in this proof of concept project (n=17) are small compared to the large number of variables generated by the seven omics platforms described increasing the possibility of finding a false positive.

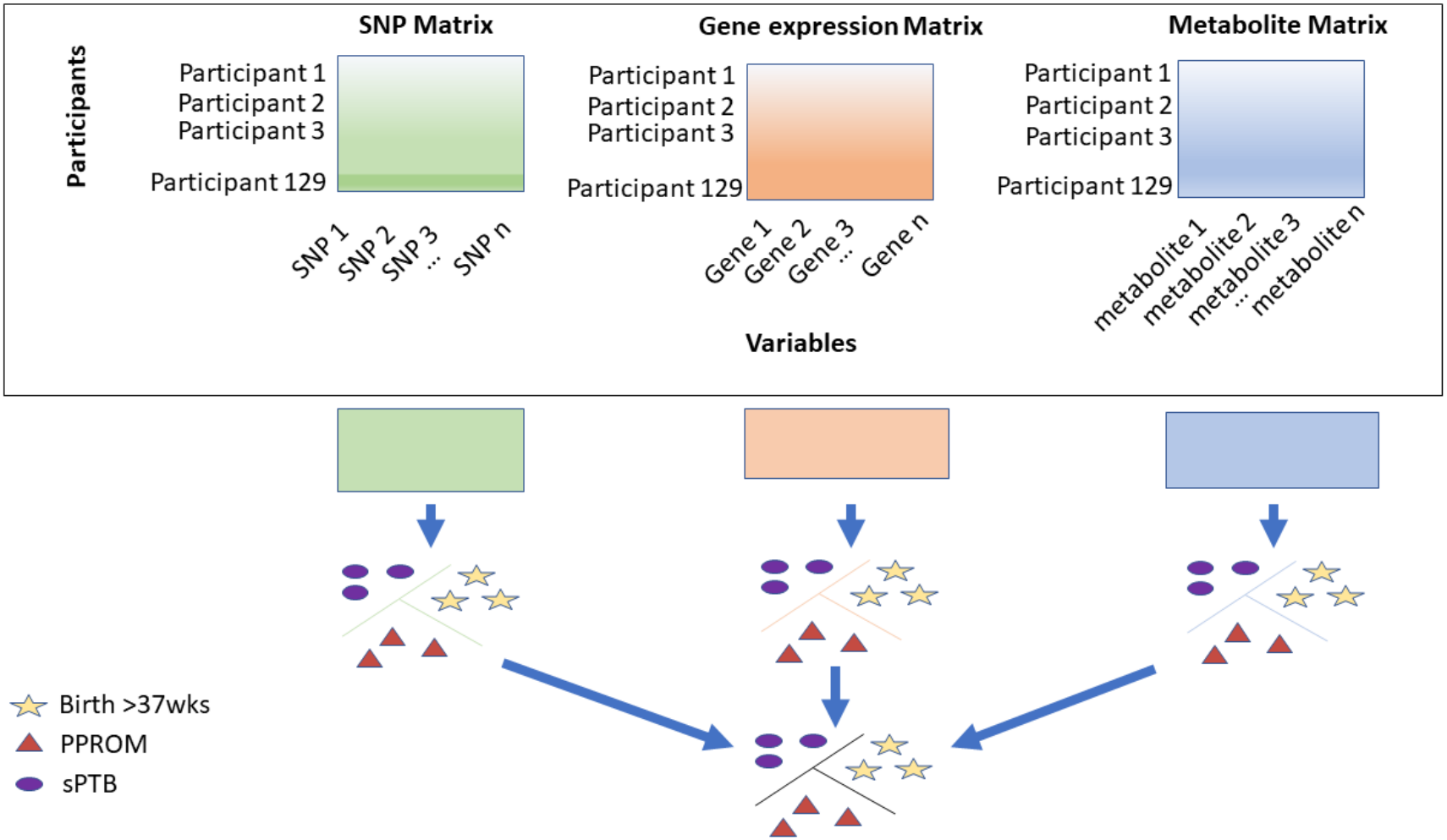


Figure 2.8 Schematic of model-based integration used in data analysis. This figure is adapted from Figure 4 published in Ritchie, M., Holzinger, E., LI, R., Pendergrass, S., Kim, D. 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nature reviews genomics.

Prediction Modelling Bias-Variance Trade Off

An important concept to understand in prediction modelling is the trade-off between a model's ability to minimize bias and variance (Figure 2.9), the prediction errors leading to underfitting and overfitting a model. It is important to understand how closely the model follows the actual patterns of the data.

Variance is the variability of a model prediction for a given data point or a value which tells us the spread of the data. A model with a high variance pays a lot of attention to training data and doesn't generalise on data points that it has never seen. As a result, these models perform well on training data but has a high error rate on test data. This is also known as 'overfitting'.

Bias refers to the gap between the value predicted by your model and the *actual* value of the data. (Theobald. 2017. pp 93). A high bias can cause an algorithm to miss the relationship between features and the target output (i.e. underfitting the model).

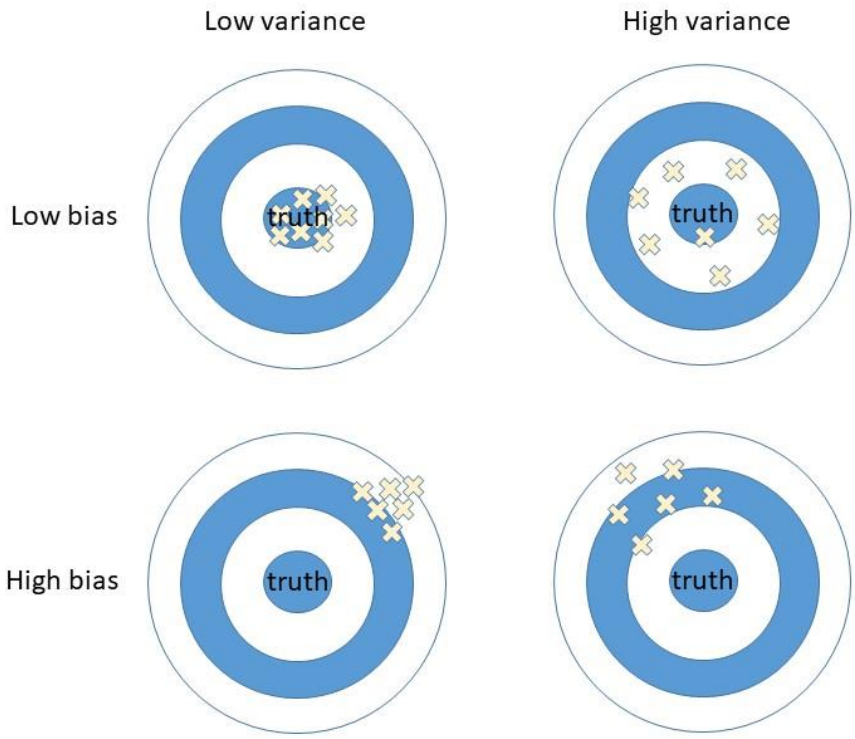


Figure 2.9. Targets used to represent the trade-off between bias and variance. Adapted from Ramchandani, P. "Random Forests and the bias-variance trade-off", *Towards Data Science.* https://towardsdatascience.com/random-forests-and-the-bias-variance-tradeoff-3b77fee339b4

2.7 Limitations of omics approaches and data integration

Multiple technical platforms are usually available for the same type of -*omic*. Various manufactures produce multiple versions of e.g. microarray and sequencing platforms which have different coverage of the genome. Alternatively, different types of technology can be used to investigate the same -*omic*, such as MS and NMR for metabolomics investigation which will ultimately result in the identification of different compounds within the same sample. Advances in technology usually mean that improvements in quality occur with new versions and researchers are keen to take advantage of the latest developments. This technological heterogeneity makes reproducibility, validation and meta-analysis of results challenging.

There are systematic differences in high throughput data between different laboratories, batch-effects and multiple operators which is widely accepted and documented. Efforts are made to reduce "batch-effects" by standardising experimental protocols and applying quality control steps prior to data analysis, but it is impossible to completely eliminate batch effects. They can be responsible for spurious findings unrelated to the outcome of interest. Using methods to account for these variations and applying appropriate statistical models such as mixed-effect models can at least address some issues in technical variation.

Genomics is a unique -*omic* in so far as the genome is fixed per individual cell and tissue type; how the genome is expressed and the subsequent molecular states existing in each tissue will vary dramatically and will also be influenced by environmental factors. Therefore, the selection of tissue type for disease study and the heterogeneity of tissues chosen will play a factor in -*omic* combination and choosing the same fluid or tissue is preferable. Even within a tissue, a sample will involve several cell types with its own unique -*omic* profile leading to heterogeneity

of results depending on proportions of cell types within samples. Using purified cell types can be an answer but can become unrealistic when faced with the cost of performing different omics on multiple cell types.

Recruitment to PTB studies will always remain challenging due to the difficulty in prediction of sPTB, the relatively low incidence in the general population and trying to clean or split data even further to investigate very specific phenotypes. Individual studies are almost certainly going to be underpowered for multiple omics analysis, and one solution is to try and improve power by obtaining more data. Meta-analysis of omics data may be the only way that power will be increased sufficiently to enable identification of candidate pathways following millions of comparisons between so many data points. This in itself poses its own challenges as data is likely to be collected, processed and recorded in very different ways introducing bias between studies for meta-analysis, as has been a challenge historically for comparison of randomised clinical trials.

We are also limited by our current knowledge of gene-transcript-protein-metabolite interactions. In network analysis we rely on pre-existing knowledge to inform pathways of data. The same analysis performed on the same data in the future may reveal a completely different result as our knowledge about the function of the human body in pregnancy at a cellular and sub-cellular level is likely to increase and pathways will become increasingly informed.

Data reduction necessary to create models that are computationally feasible may accidentally filter out important associations. For example, a single functional SNP may associate with sPTB. However, if this SNP is in linkage disequilibrium with another non-functional SNP, the functional SNP may be filtered out in favour of the non-functional SNP. Therefore, it is important to understand the assumptions of

any model and limitations of analysis before making inferences and interpretations of the data (Ritchie et al. 2015).

2.8 Conclusion

The development in “omics” technology has led to exciting breakthroughs and new avenues of investigation for sPTB prediction. Although we are yet to translate these changes to clinical and patient benefit our increasing understanding of the complex pathways underpinning sPTB is increasing and this is a promising and novel area of investigation.

Chapter 3: Study Population

3.1 Introduction

Having argued in the previous chapter the importance of using omics technology for the prediction of sPTB and provided an overview of the various methodologies available; this chapter will describe the design and methodology selected to combine omics datasets for preterm birth prediction and investigate sPTB phenotypes using multiple omics and systems biology approaches. It was important to design a study that minimizes the limitations of the omics approaches described in chapter 2, to remain focused on the research hypothesis, but still to be sufficiently pragmatic to be achievable.

Chapter 1 demonstrated that current ability to predict sPTB remains poor as mechanisms of disease are not understood. Currently used screening methods are insufficient to detect all women at risk. The sequelae of preterm birth can lead to significant mortality and morbidity and there is a clear need for improvements in disease prevention, which cannot be obtained until we have a better understanding of causation or a more robust way of accurately discriminating those at high risk.

Our ultimate goal is to establish clinically useful personalized risk assessment with a combination of clinical and comprehensive molecular phenotyping. This approach will lead to better and safer use of currently available preventative therapies (drug repositioning) and development of novel, more effective therapies both in high and low resource settings.

3.2 Aims

To investigate preterm birth phenotypes by using multiple omics and systems biology approaches in high risk population. My goals were:

- To recruit over a 3-year period a prospective cohort of women with sPTB at <34 weeks gestation (cases) and a cohort of women with spontaneous term delivery (controls). All women should have well-characterized clinical phenotypes with biologic samples suitable for multi-platform systems biology analysis that were collected at a minimum of 2 time points during gestation and at delivery.
- To apply multi-omic high-throughput technologies to generate longitudinal datasets.
- Pilot integration of these datasets into systems biology approaches to allow for interpretation

3.3 Population Identification

The samples used for analysis in this study come from participants recruited as part of "The development of novel biomarkers for prediction of preterm labour in a high-risk population" study" (REC reference: 11/NW/0720) collected between 1st March 2012 and 28th May 2015 (Appendix A and B). This project was co-sponsored by Liverpool Women's Hospital and the University of Liverpool (Appendix C). Women were consented and recruited at the Liverpool Women's Hospital Harris-Wellbeing Preterm Birth Prevention Clinic. In routine clinical practice women are screened for a short cervix in pregnancy if they have:

- a history of sPTB or PPROM (between 16 and 33^{+6} weeks), *or*
- if they have had significant excisional treatment of the cervix (2 x Large Loop Excision of the Transformation Zone (LLETZ) or single knife cone biopsy) from 16 weeks pregnancy onwards.

To try to avoid large variation in aetiology of sPTB in our recruited women we limited this study to women who had only had a history of sPTB or PPROM. Women attending because of a history of excisional cervical surgery but no history of preterm labour were excluded.

Our clinical audit figures from 2010-2013 showed that at the Liverpool Women's Hospital the sPTB rate <34 weeks is 17% for women with a history of sPTB or PPROM. In 2011, 135 new women were referred, 95 (70%) had a history of sPTB or PPROM <34 weeks. Estimating a 50% recruitment rate, we expected to recruit 140 women over 3 years, and approximately 24 cases of sPTB <34 weeks. The advantage of recruiting cases experiencing a recurrent preterm birth is that there may be positive selection of aetiologies of sPTB inherent to the mother rather than an individual pregnancy. For example, women with a genetic predisposition to sPTB

may be more likely to have two pregnancies affected by sPTB than if infection was the cause, which may only affect a single pregnancy.

Women with a history of sPTB or PPROM who had a subsequent term birth *without* treatment for short cervix were used as a control group.

Inclusion Criteria

- Previous PPROM >16 and <34 weeks
- Previous sPTB >16 and <34 weeks
- Singleton pregnancy
- Willing to undergo transvaginal ultrasound scan
- Age >18 years
- Understands English
- Understands study requirements, agrees to participate and written consent obtained.

Exclusion Criteria

- Iatrogenic PTB
- PPROM <16 and >34weeks
- Multiple pregnancy

3.4 Time points and Sampling

Two time points were chosen for sampling; 16 and 20 weeks. Serial sampling was chosen to provide more information than isolated measurements of biomarkers at a single timepoint.

Sixteen weeks is the gestation at which high risk patients commonly attend for their first cervical length screening (Care et al. 2019). This is after the risk of first trimester miscarriage has passed, but prior to the risk of late second trimester miscarriage or preterm birth. At the first visit to the clinic, a member of the clinical staff recruiting team would discuss the project and provide an information leaflet to the potential participant (Appendix D). The woman would have her clinical appointment as normal and if she agreed to participate, a consent form was signed (Appendix E). Bloods were taken at the end of the appointment when the participants data was also collected. Twenty weeks is four weeks prior to the risk of sPTB and when low risk women would routinely attend hospital for anomaly scan, which would facilitate rolling out any successful screening tests.

As recommended by the World Health Organisation (Wilson and Jungner. 1968), a good screening test would detect pathology early to allow for enough time for intervention to decrease this risk. Sampling at 16 and 20 weeks in an asymptomatic population would hopefully allow for this time prior to the onset of labour. Although blood taking is widely accepted in medicine as a necessary test for a variety of conditions, it is invasive and can be painful. More frequent sampling every two weeks was considered. However, it was likely to deter women from participating. Accordingly, 16 and 20 weeks were considered to be the most useful time points in pregnancy for a screening test for sPTB prediction.

Three omics layers per patient were selected to be analysed from blood for predictive biomarkers of preterm birth; genetics, transcriptomics and metabolomics. Venepuncture performed with BD Vacutainer® Eclipse™ blood collection needles allowed for a:

- 6ml BD vacutainer® K_2EDTA for maternal genome (lavender small),
- 2.5ml PAXgene Blood RNA Tube
- 6ml BD vacutainer® tubes containing clot activator for biomarkers (red tube) to allow storage of serum for metabolomics study

All samples were inverted 10 times to allow for mixing between the blood and tube reagents then stored on ice immediately after collection until taken to the laboratory at a convenient time but no more than 1 hour after sampling. Samples were processed according to the standard operating procedure (appendix F).

3.5 Method of Classification

Hospital records were used to ascertain delivery details for all women giving birth at Liverpool Women's Hospital. For participants that delivered elsewhere, the research department at the delivering unit were contacted and asked to provide delivery details. Where this was not possible, as a final resort, participants were contacted directly by telephone.

Figure 3.1 shows the workflow for the classification of births, and Table 3.1 shows how clinical diagnoses were defined. The judgement of the contemporaneously recorded treating clinicians was used to record diagnoses unless further information became available later that refuted this. The clinical notes of participants whose first event (sPTB or PPROM) was $\leq 36^{+6}$ weeks gestation was accessed and reviewed independently by myself and senior clinical lecturer, Dr Andrew Sharp. In the cases where there was a discrepancy between reviewers, the case was reviewed by a third researcher, Professor Zarko Alfirevic, until the team reached a consensus on classification of the birth. Both Professor Alfirevic and Dr Andrew Sharp and have extensive experience both clinically and in research, in the field of preterm birth. They have both worked at the specialist preterm birth prevention clinic in Liverpool Women's Hospital since it opened in 2010 and have published research papers in this field.

For patients with PPROM the gestation at the time of ruptured membranes was taken as the significant gestation for the analysis, rather than the gestation at subsequent birth. For women with spontaneous preterm labour, the gestation at birth was used as the significant gestation for analysis.

Additional risk factors and phenotypic pregnancy features related to preterm birth were systematically recorded but not included in the overall classification

system. Table 3.2 shows a collection of risk factors that have been defined for reporting. An additional step was included for the women who had PPROM to describe whether infection was thought to be present at the time of PPROM. Infection is considered a cause of PPROM and a consequence of ascending genital tract infection without protection from intact membranes. Therefore, a seven-day period was used to distinguish between early chorioamnionitis (likely cause of PPROM) and late chorioamnionitis (likely consequence of PPROM). This workflow is detailed in Figure 3.2.

If an intrauterine death or termination of pregnancy had been preceded by PPROM, this was included as a PPROM case. In cases where there was an induction of labour for suspected PPROM both the method of diagnosis (history and pad check only, history and speculum examination +/- bedside test +/- USS of mean pool depth of AF) and the success of the induction were evaluated. In cases where there was no definitive method of PPROM diagnosis (e.g. history and pad check only) or there was suspicion of a false positive diagnosis (recurrent USS showing normal AF volumes) the case would be excluded if an artificial rupture of membranes (ARM) was required during labour to facilitate progress.

For cases where there was either growth restriction $<5^{th}$ centile or severe pre-eclampsia diagnosed, providing that the onset of preterm labour was spontaneous these cases were included in our definition and these risk factors reported as possible signs of placental dysfunction.

Threatened preterm labour that was later augmented were excluded (i.e. no symptoms of labour could be documented such as shortening or dilation of cervix, or ruptured membranes).

Major classification of births

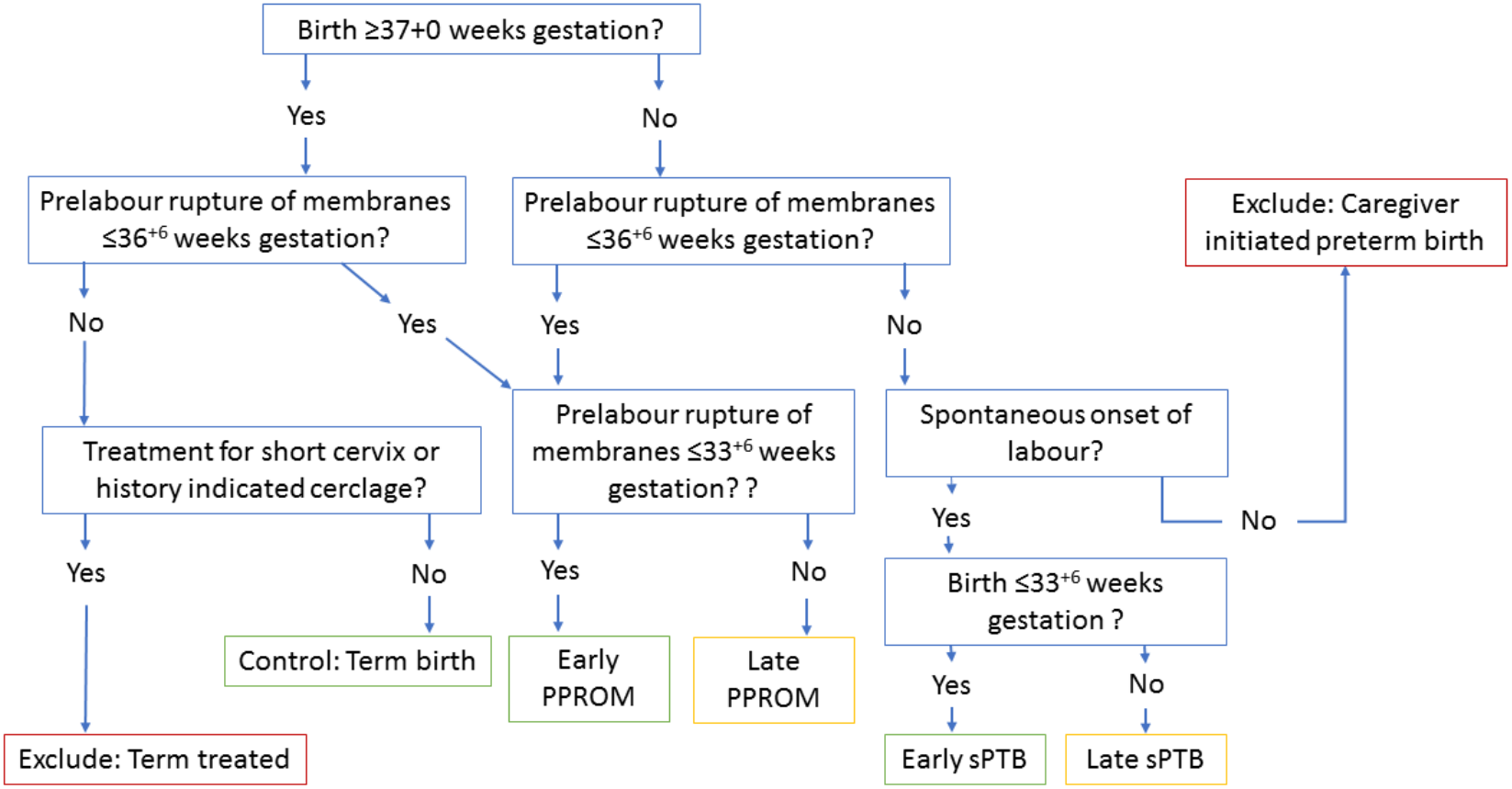


Figure 3.1. Workflow showing classification of cases and controls.

The workflow used the following definitions:

Table 3.1. Definitions used for clinical diagnosis

Classification	**Description**
Spontaneous labour	Regular uterine activity with cervical shortening or dilation such that the treating clinicians judged labour to be present
Prelabour rupture of membranes	Rupture of membranes confirmed either by speculum examination or the AmniSure ROM or ActimPROM bedside test without onset of spontaneous labour in the following 12 hours
Caregiver initiated preterm birth	Induction of labour or Caesarean section $\leq 36^{+6}$ weeks gestation *without* evidence of PPROM or spontaneous preterm labour

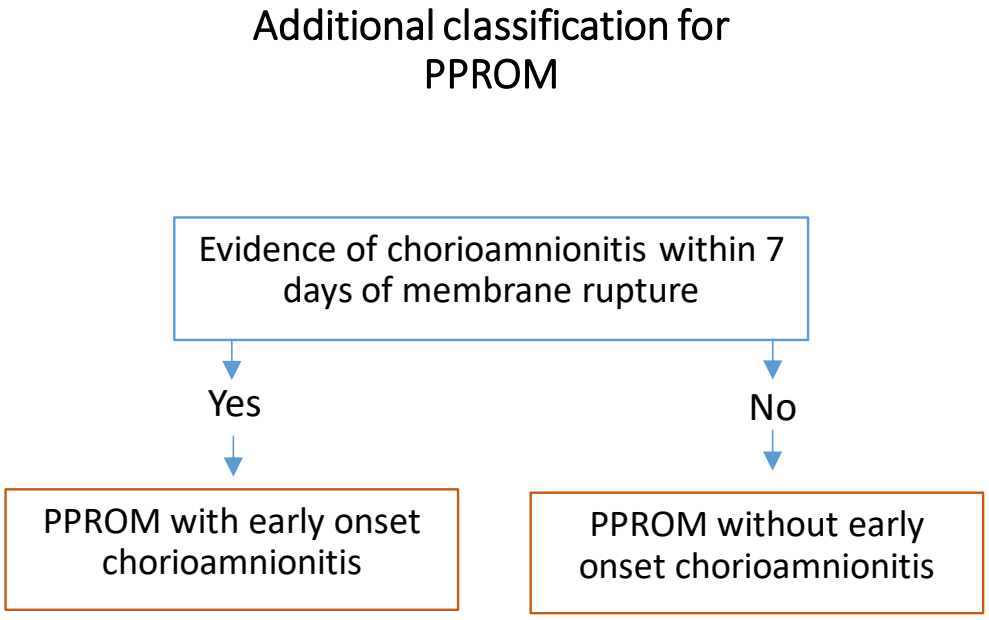


Figure 3.2. Workflow for additional classification for women with PPROM. See Table 3.2 for definition of chorioamnionitis.

Table 3.2. Definitions of contributing factors to sPTB based on the clinical phenotyping tool developed by Villar et al. (2012).

Contributing factors to preterm birth	**Description**
Chorioamnionitis	Contemporaneous clinical notes documenting diagnosis of chorioamnionitis due to one or more of; persistent maternal pyrexia (≥38^{o}C), purulent or foul-smelling amniotic fluid, or the use of broad spectrum antibiotics to treat chorioamnionitis based on the clinical situation and evidence of chorioamnionitis on placental histology or blood culture. *Note: IF histologic evidence of chorioamnionitis only and no clinical features this was classified as subclinical chorioamnionitis.*
Placental dysfunction	Evidence of placental dysfunction contributing to preterm labour. Defined as evidence of placental abruption at time of delivery or on placental histology report, birthweight under 5th customised centile, or severe preeclampsia
Extra amniotic infection	Contemporaneous clinical notes documenting concern about major systemic infection due to one or more of; raised white cell count (>15 x 10^{9}/L), raised C-reactive protein (CRP >30), persistent maternal pyrexia (≥38^{o}C), microbiological culture of pathological organism from a normally sterile site, or the use of broad-spectrum antibiotics for presumed extra amniotic infection (e.g. UTI).
Polyhydramnios	Maximum pool depth ≥ 10cm on ultrasound assessment
Uterine anomaly	Documented uterine anomaly
Maternal comorbidities	Maternal medical condition that affects a major organ system or is associated with preterm birth (for example antiphospholipid syndrome, chronic hypertension, chronic renal failure, Elhers-Danlos syndrome, epilepsy, pre-existing and gestational diabetes)
Cervical shortening	Received treatment for short cervical length <28 weeks gestation. Short cervix is defined as ≤25 mm or <3rd centile for gestational age on a validated cervical reference chart for pregnancy.

All participants were classified in the following mutually exclusive categories:

Table 3.3. Final classification system. Term births highlighted in green were included as controls. Early PPROM (yellow) and sPTB (blue) were included as two separate groups of cases.

Participant population	**Major classification**	**Description of birth**
High risk *Previous PPROM or sPTB between 16^{+0} -33^{+6} weeks gestation*	Term	Birth ≥37+0 weeks gestation; no Rx for short cervix
		Birth ≥37+0 weeks gestation; Rx for short cervix. EXCLUDED
	Caregiver initiated preterm birth	Caregiver initiated preterm birth
	Early PPROM	PPROM ≤33^{+6} weeks gestation with early chorioamnionitis
		PPROM ≤33^{+6} weeks gestation with polyhydramnios
		PPROM ≤33^{+6} weeks gestation without early chorioamnionitis
	Late PPROM	PPROM 34^{+0}-≤36^{+6} weeks gestation with early chorioamnionitis
		PPROM 34^{+0}-≤36^{+6} weeks gestation without early chorioamnionitis
	Early sPTB	sPTB ≤33^{+6} with evidence of infection
		sPTB ≤33^{+6} without evidence of infection
	Late sPTB	sPTB 34^{+0}-≤36^{+6}
	Unknown	Unable to ascertain birth details

3.6 Sample Size and Power Calculation

A challenge for multiomic studies is obtaining enough samples to power results. The power of a study is the probability of successfully detecting a given effect size. The importance of this remains paramount, as reflected by the American National Institute of Health's funding application *non-optional component* of sample size and power analysis. Yet statisticians face challenges when asked to estimate a sample size required for a multiomics study. Classical methods for calculating statistical power to reject a null hypothesis at a specified level of significance are not applicable to this problem (McKeigue. 2019). In "omics" the high dimensionality is a source of difficulties since the number of variables vastly exceeds the number of samples or participants in the studies. Some "omics" studies focus on reducing the risk of making a false assertion that an observed difference is true when it isn't (type 1 error). In GWAS studies, allele frequencies and effect size are taken into consideration and then a correction (e.g. Bonferroni correction) is applied for multiple comparisons. Most analyses of GWAS data sets consider genetic variants on a microarray chip comparing hundreds of thousands of SNPs. This leads to a stringent statistical cut-off level that defines a true variant for common and complex diseases (p value threshold of $<5 \times 10^{-8}$). Many sub significant hits may also be mapped to genes involved in disease that are true risk loci but differentiating them from false positives is difficult.

Power analysis for high throughput sequencing based experiments (not used in this thesis) are even more complex. Firstly, due to the unique parameters for sequencing read depths that directly affect the ability to detect variants or gene expression and therefore need to be considered in the power analysis (Li et al. 2018). Secondly, the number of possible applications for sequencing greatly exceeds

microarray, introducing a variation of unique statistical scenarios (Li et al. 2018). The data platforms used for omics studies assay thousands of SNPs, gene transcripts and metabolites to try and account for individual differences in disease susceptibility. Methods described to date for estimating the sample size required for classification using high dimensional biomarkers require more understanding of how changes in one variable are associated with changes in a second variable, or covariance of the biomarkers (Dobbin and Song. 2013).

Unfortunately, most of the published methodologies for microarray summarised by Lin *et al.* (2010) and Jung *et al.* (2012) and high dimensional data (Li et al. 2018) are restricted to two group (case-control) comparisons and require a user-defined effect size. This does not allow for comparison of three groups (PPROM, sPTB and control). Additionally, power is calculated individually on each type of omics analysis, to date there has been no described method to perform a power calculation for multiomics data from the same patient and published studies that have attempted omics data combination in omics have not discussed a power calculation (Knijnenberg et al. 2019).

It might be tempting to increase sample sizes in groups by retaining a broad classification for sPTB, e.g. including all patients who deliver <37 weeks or combining cases of sPTB and PPROM. Unfortunately, this is likely to have the opposite effect, as increasing the heterogeneity within study groups makes it very difficult to separate true signals from noise and importantly it will become harder to validate the results. Clear or stricter phenotyping of sPTB will reduce the potential number of samples included in the analysis, but it may serve to increase biological homogeneity and thus increase power by increasing effect size differences. For sPTB

the issue of heterogeneity is almost an inherent problem until we better understand the causes.

Women experience labour at different gestational ages, experience different symptoms of labour, if at all. They rupture membranes and/or experience contractions and dilate the cervix at different relative time points of pregnancy and experience differences in predictive symptoms such as bleeding or a short cervix. However, there is sufficient evidence that, as a minimum, sPTB and PPROM behave differently to each other (Capece et al. 2014).

A strength of omics data combination is that by combining multiple data types, this can provide increased power. Data integration can compensate for missing data in any single data type and multiple sources of evidence all pointing to the same biological pathway or gene reduce the likelihood of a false positive result.

Our study sample is not therefore based on an *a priori* power calculation. This is a pilot analysis of the samples available from this project collected between 1st March 2012 and 28th May 2015 to assess:

- How many samples could be collected
- How much 'omics' data would be available following quality control of samples due to technical requirements of the laboratory,
- Financial limitations
- Patient drop out or failure to follow up rate.

3.7 Results

One hundred and twenty-nine women were recruited to the study. Written consent was obtained at the time of the first blood collection. One participant was subsequently excluded as she had previously only had an episode of *threatened* preterm labour. According to clinician signatures on the consent forms, I personally recruited 80% of all 128 participants. All women were followed up and gestational ages at delivery were obtained. For five participants birthweights were not available.

Pregnancy Outcomes and Classifications

Table 3.4 shows the final numbers recruited to each category of cases and controls. Twenty-four exclusions are detailed. For sPTB, PPROM and term control cases (highlighted in colour in Table 3.4) their demographics are shown in Table 3.5.

Excluded cases

Three cases were excluded as they were caregiver-initiated (iatrogenic) preterm births. These women did not require treatment for a short cervix and were delivered at 34^{+1}, 35^{+3} and 36^{+1} for maternal cancer treatment, poorly controlled diabetes and fetal growth restriction respectively.

All women who had received preterm birth prevention treatment and had a delivery after 37 weeks were excluded. We are unable to tell if the treatment they received affected the natural history of biological events and presume that intervention prevented a sPTB. However, it is entirely possible that some of these women may have achieved a term delivery without prevention treatment. As this group may have heterogenous biomarkers of risk, they have been excluded completely from analysis. One pregnancy where birth was induced following

PPROM at 17^{+1} was found to have a supernumerary ring chromosome in 17/20 cell lines and was excluded.

Table 3.4 Pregnancy outcomes following delivery of 128 participants. Control cases highlighted in green. Preterm prelabour rupture of the membranes (PPROM) cases highlighted in yellow and spontaneous preterm birth (sPTB) cases highlighted in blue.

Participant population	**Major classification**	**Description of birth**	**Number Recruited**
High risk *Previous PPROM or sPTB between 16^{+0}-33^{+6} weeks gestation*	Term	Birth ≥37+0 weeks gestation; no Rx for short cervix	60
		Birth ≥37+0 weeks gestation; Rx for short cervix. **EXCLUDED**	21
	Caregiver initiated preterm birth	Caregiver initiated preterm birth - **EXCLUDED**	3
	Chromosomal abnormality	Chromosomal abnormalities discovered on post-mortem that may have influenced miscarriage risk - **EXCLUDED**	1
	Early PPROM	PPROM ≤33^{+6} weeks gestation with early chorioamnionitis	2
		PPROM ≤33^{+6} weeks gestations with polyhydramnios	1
		PPROM ≤33^{+6} weeks gestation without early chorioamnionitis or known polyhydramnios	10
	Late PPROM	PPROM 34^{+0}-≤36^{+6} weeks gestation with early chorioamnionitis	1
		PPROM 34^{+0}-≤36^{+6} weeks gestation without early chorioamnionitis	0
	Early sPTB	sPTB ≤33^{+6} with evidence of infection	2
		sPTB ≤33^{+6} without evidence of infection	12
	Late sPTB	sPTB 34^{+0}-≤36^{+6}	15

Demographics

Table 3.5. Demographics for participants in biomarker study split into women delivering after 37 weeks, Preterm prelabour rupture of membranes (PPROM) <34 weeks and spontaneous preterm labour (sPTB) <34 weeks.

	Early sPTB N=14	Early PPROM N=13	Term Birth N=60	P value
Participant Demographics				
Maternal age, mean years +/- SD	31 (6.7)	29 (5.3)	31 (5.2)	.538[a]
Booking BMI, mean +/- SD	27 (5.9)	25 (3.4)	25 (4.2)	.289[a]
Ethnicity				
Caucasian, n(%)	14 (100)	12 (92)	54 (89)	.396[b]
Non-Caucasian, n (%)	0	1 (8)	7 (11)	
Smoking during pregnancy				
Yes, n(%)	1 (7)	5 (39)	16 (26)	.158[b]
No, n(%)	13 (93)	8 (61)	45 (74)	
Clinical Characteristics				
Gravidity, mean+/-SD	3.21 (1.3)	2.9 (1.2)	3.4 (1.6)	.503[c]
Parity	1.21 (0.8)	1.36 (0.9)	1.75 (1.6)	.422[c]
History of previous PTB				
Previous sPTB, n (%)	6 (43)	5 (39)	27 (44)	.187[b]
Previous PPROM, n (%)	6 (43)	6 (46)	33 (54)	
Both previous sPTB and PPROM (%)	2 (14)	2 (15)	1 (2)	
Previous twin sPTB, n (%)	0	0	0	N/A
Cervical surgery, n (%)	3 (21)	1 (8)	1 (2)	.002[b]
Gestational age, visit 1, mean +/- SD days	16+3 (6)	16+3(4)	16+3 (5)	.908[a]
Cervical length visit 1, mean +/- SD	33.3 (7)	35.5 (9)	36.5 (6)	.352[a]
Gestational age, visit 2	143 (7)	20+3 (4)	20+2 (6)	.860[a]
Cervical length visit 2, mean +/- SD	21.8 (12)	27 (12)	36.4 (7)	**.000**[a]
Antiphospholipid Syndrome				
Yes, n (%)	0	0	0	
No, n (%)	14 (100)	13 (100)	61 (100)	N/A
Other chronic medical conditions				
Yes, n (%)	9 (64)	10 (77)	24 (39)	.022[b]
No, n (%)	5 (36)	3 (23)	37 (61)	
Polyhydramnios				
Present, n (%)	0	1 (8)	0	.054 [b]
Absent, n (%)	14 (100)	12 (92)	61 (100)	
Labour and Delivery Characteristics				
Onset of labour				
Spontaneous	12 (86)	3 (21)	33 (54)	.011[b]
Induction	2 (14)	10 (71)	19 (31)	
No labour	0	1 (7)	9 (15)	
Gestational age at PPROM, mean +/- SD	N/A	27^{+5} (42)	N/A	N/A

Gestational age at delivery, mean +/- SD days	29^{+0} (33)	29^{+2} (40)	39^{+1} (9)	**.000[a]**
Birthweight, mean grams +/- SD	1395 (779)	1448 (747)	3290 (453)	**.000[a]**
Neonatal gender				
Female, n (%)	5 (36)	8 (62)	23 (38)	.471[b]
Male, n (%)	8 (57)	5 (38)	35 (57)	
Not recorded (%)	1 (7)	0	3 (5)	
Placental Abruption				
Yes, n (%)	0	0	0	N/A
No, n (%)	14 (100)	14 (100)	61 (100)	
Chorioamnionitis / Infection				
Yes, n (%)	2 (14)	2 (15)	0	.009[b]
No, n (%)	12 (86)	11 (85)	61 (100)	
Growth $<10^{th}$ centile				
Yes, n (%)	2 (14)	3 (21)	11 (18)	.922[b]
No, n (%)	11 (79)	11 (79)	48 (79)	

[a]One way analysis of variance (ANOVA). [b]Chi squared [c]Kruskal Wallis

3.8 Recruitment Feasibility

Monthly recruitment rates are shown below in Table 3.6 with the average monthly recruitment figures in Figure 3.3.

Table 3.6. Monthly recruitment figures.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
2012	-	-	1	5	3	2	2	2	1	1	4	0
2013	3	4	6	4	7	1	10	2	4	2	1	3
2014	2	1	3	2	3	2	4	3	1	4	7	5
2015.	6	5	4	6	9	-	-	-	-	-	-	-

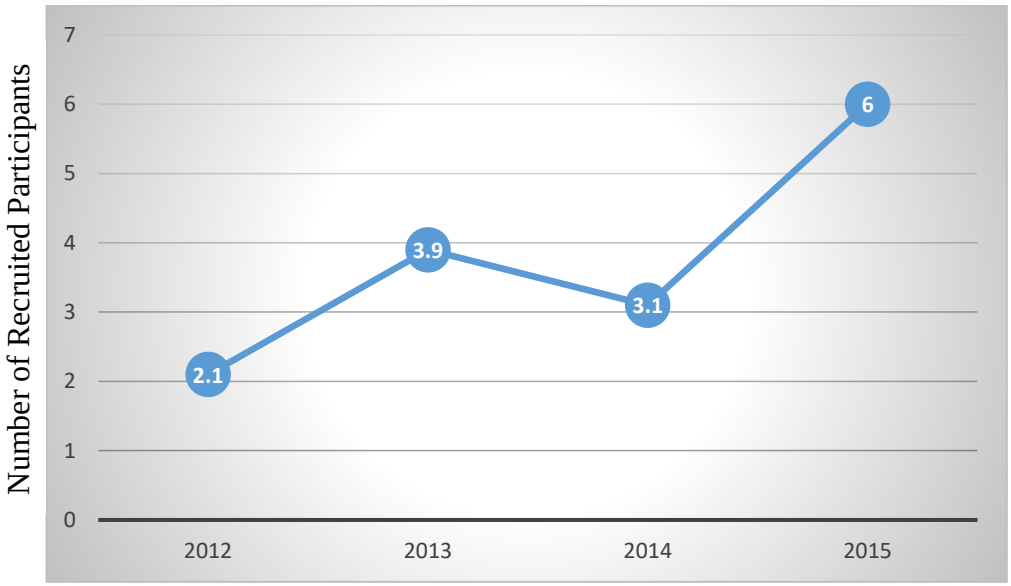


Figure 3.3. Average monthly recruitment figures per year of study from LWH preterm birth prevention clinic. The X axis shows year of recruitment and the Y axis shows average number of recruited women per month.

Overall there was a gradual increase in recruits per month. There are several possible reasons for this; firstly, the preterm labour clinic increased the overall number of new patients from 2012 to 2015, increasing the number of eligible participants. Secondly, once more funding was available laboratory staff were recruited to assist in processing the samples to allow clinicians to remain in clinic

and recruit. Thirdly, to prevent eligible patients attending clinic at the same time, and therefore reducing the likelihood of only one patient being recruited eligible patients were offered appointments approximately 30 mins apart when possible. Lastly with time, the running of the study became more efficient with increasing help from the other clinical and auxiliary staff.

In the laboratory, no RNA was extracted following an extraction run of 24 samples (12 patients) due to an error with the reagents from a single defective RNA extraction kit. This was identified and rectified with the kit supplier, but these samples were lost. This is likely to be an isolated incident but for future studies, performing a test run with an individual sample when opening a new kit would be advised to avoid the loss of samples. The total number of patients from 56 that ended the study with a complete set of omic data at both time points was 25 (45%), although this could have been as high as 37 (66%) if RNA had not failed to be extracted.

For full omics integration, late sPTB and PPROM were excluded to try creating distinct biological groups to identify biomarkers for clinically important preterm birth. However, for independent omics analysis such as GWAS (where meta-analysis with other studies would be possible), a cut-off of 37 weeks has been used and the data for 'late sPTB' has been included. This is discussed in the relative chapters.

3.9 Study Samples for Omic Integration

At several stages of the recruitment process it was possible to reduce the number of samples contributing to the final omics analysis. Figure 3.4 illustrates at what stage a sample could be excluded from the dataset and the reason.

Table 3.7 illustrates how many individual participants had datasets available for each layer of omics analysis and Table 3.8 shows how many participants had complete “omics” sets available for integration.

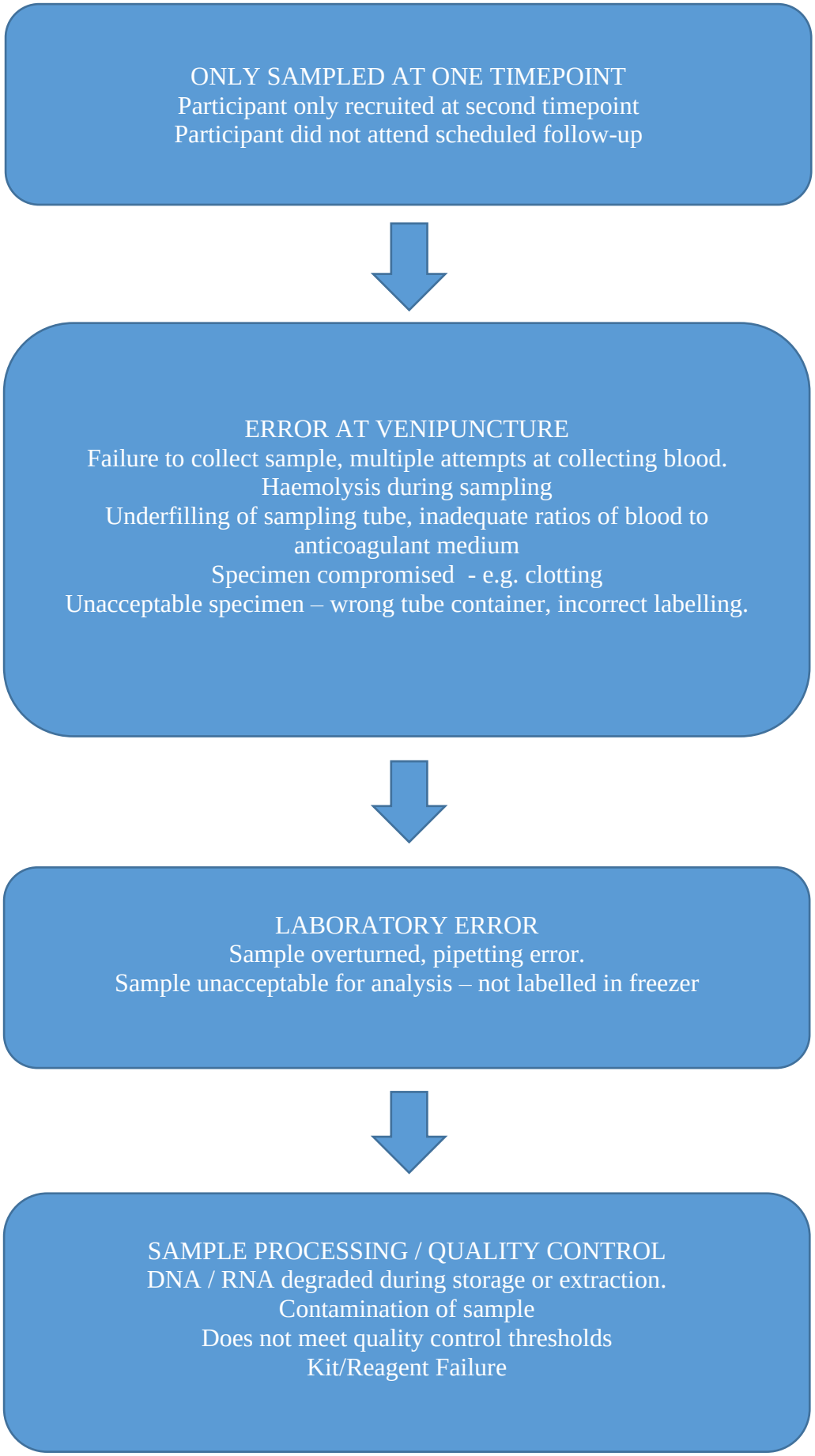


Figure 3.4. Possible areas of error in sampling process

Table 3.7. Omics data collected for all 103 included pregnant participants (exclusions detailed in Table 3.4 removed)

Participant Number	Phenotype	GWAS (DNA)	RNA 16 weeks	RNA 20 weeks	Metabolome 16weeks	Metabolome 20 weeks	Full Omic Integration 16w	Full Omic Integration 20w
1	**TERM CONTROL**	Yes	No	No	No	Yes	No	No
2	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
3	**TERM CONTROL**	Yes	No	No	No	Yes	No	No
4	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
5	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
6	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
7	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
8	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
10	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
12	**TERM CONTROL**	Yes	No	No	Yes	No	No	No
13	**TERM CONTROL**	Yes	No	No	Yes	No	No	No
14	**LATE sPTB**	Yes	No	No	Yes	Yes	No	No
15	**LATE sPTB**	Yes	No	No	Yes	Yes	No	No
16	**PPROM**	Yes	No	No	Yes	Yes	No	No
17	**PPROM**	Yes	No	No	Yes	Yes	No	No
18	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
19	**LATE sPTB**	Yes	No	No	Yes	Yes	No	No
20	**sPTB**	Yes	No	No	Yes	No	No	No
21	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
22	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
23	**TERM CONTROL**	Yes	No	No	Yes	No	No	No
24	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
25	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
27	**TERM CONTROL**	Yes	No	No	No	Yes	No	No
28	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
29	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
31	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
32	**PPROM**	Yes	No	No	Yes	Yes	No	No
33	**PPROM**	Yes	No	No	Yes	Yes	No	No
34	**sPTB**	Yes	No	No	Yes	Yes	No	No
35	**sPTB – infection**	Yes	No	No	No	Yes	No	No
37	**PPROM**	Yes	No	No	Yes	Yes	No	No
38	**PPROM**	Yes	No	No	Yes	Yes	No	No

39	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
41	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
43	**LATE sPTB**	Yes	No	No	Yes	Yes	No	No
44	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
45	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
46	**TERM CONTROL**	Yes	No	No	Yes	No	No	No
47	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
49	**sPTB**	Yes	No	No	Yes	Yes	No	No
52	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
53	**sPTB**	Yes	No	No	No	Yes	No	No
54	**PPROM + chorio**	Yes	No	No	Yes	Yes	No	No
55	**sPTB**	Yes	No	No	Yes	Yes	No	No
57	**LATE sPTB**	Yes	No	No	Yes	Yes	No	No
59	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
60	**PPROM**	Yes	No	No	Yes	Yes	No	No
61	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
62	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
63	**sPTB**	Yes	No	No	Yes	Yes	No	No
65	**sPTB – infection**	Yes	No	No	Yes	Yes	No	No
66	**sPTB**	Yes	No	No	Yes	Yes	No	No
67	**LATE sPTB**	Yes	No	No	Yes	No	No	No
68	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
73	**PPROM**	Yes	Yes	No	Yes	No	Yes	No
74	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
75	**sPTB**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
76	**TERM CONTROL**	Yes	Yes	Yes	Yes	No	Yes	No
77	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
78	**PPROM**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
79	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
81	**PPROM + chorio**	Yes	Yes	No	Yes	No	Yes	No
82	**PPROM**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
83	**sPTB**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
84	**LATE sPTB**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
85	**sPTB**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
86	**TERM CONTROL**	Yes	No	Yes	No	Yes	No	Yes
87	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
88	**sPTB**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
89	**TERM CONTROL**	Yes	Yes	No	Yes	Yes	Yes	No
90	**LATE sPTB**	Yes	Yes	Yes	Yes	Yes	Yes	Yes

91	**TERM CONTROL**	Yes	No	Yes	No	Yes	No	Yes
92	**LATE sPTB**	No	Yes	Yes	Yes	Yes	No	No
95	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
96	**TERM CONTROL**	Yes	No	Yes	Yes	Yes	No	Yes
97	**LATE sPTB**	Yes	No	Yes	No	Yes	No	Yes
98	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
101	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
102	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
103	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
105	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
107	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
108	**LATE sPTB**	Yes	No	No	Yes	Yes	No	No
109	**TERM CONTROL**	Yes	No	No	Yes	Yes	No	No
110	**LATE sPTB**	Yes	No	No	Yes	Yes	No	No
111	**TERM CONTROL**	Yes	No	Yes	Yes	Yes	No	Yes
112	**TERM CONTROL**	Yes	No	Yes	Yes	Yes	No	Yes
113	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
114	**LATE sPTB**	Yes	No	Yes	Yes	Yes	No	Yes
115	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
116	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
117	**TERM CONTROL**	Yes	No	Yes	Yes	Yes	No	Yes
118	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
119	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
120	**LATE sPTB**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
121	**LATE sPTB**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
122	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
123	**sPTB**	Yes	Yes	No	Yes	Yes	Yes	No
124	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
126	**PPROM – poly**	Yes	Yes	Yes	Yes	Yes	Yes	Yes
128	**PPROM – genetic**	Yes	Yes	No	Yes	No	Yes	No
129	**TERM CONTROL**	Yes	Yes	Yes	Yes	Yes	Yes	Yes

		GWAS	RNA 16W	RNA 20W	Metabolomics 16W	Metabolomics 20W	Full Omics for Integration 16W	Full Omics for Integration 20W
Total No. Participants	**103**	**102**	**32**	**35**	**97**	**95**	**31**	**33**
No. TERM CONTROL	60	60	17	21	55	53	**15**	**20**
No. LATE sPTB/PPROM	15	14	5	7	14 (Excluded)	14 (Excluded)	4 (Excluded)	6 (Excluded)
No. sPTB	14	14	5	4	11	11	**5**	**4**
No. PPROM	14	14	5	3	10	9	**6**	**3**

Table 3.8. Data available for each omic analysis and total omic integration after participants grouped based on their obstetric outcomes.

3.10 Conclusion

Between 1st March 2012 and 28th May 2015 from a single large preterm birth prevention clinic (approx. 140 new patients per year) it was possible to recruit 128 patients meeting our specific inclusion criteria. From these patients; 61 (48%) term controls, 14 (11%) sPTB and 14 (11%) PPROM cases were obtained.

RNA was only collected for a subset of 56 patients in this cohort (29 controls, 5 sPTB, 6 PPROM). However, following sample storage, extraction and laboratory quality control checks using our methodologies the available data for three set omic integration (genomics, transcriptomics and metabolomics) at 16 weeks was 15 (51%) term controls, 5 (100%) sPTB and 6 (100%) PPROM cases.

For larger cohort studies it is important to consider all the potential areas that samples may be lost and collect enough samples to allow for error. Recruiting from a larger pool of participants (multi-site studies) depending on timescales and a priori recruitment targets will improve the speed of sample attainment.

Chapter 4: Assessing genetic predisposition to preterm birth in women with recurrent spontaneous preterm birth

4.1 Introduction

In this chapter the quality control and genomic analysis is discussed prior to data integration. As described in chapter 1, sPTB is a complex disease with a complex pattern of inheritance. After consideration of different methods of genetic analysis in chapter 2, a *genome wide association study* (GWAS) was chosen as the analytical tool of choice. A case-control design was selected to integrate with the other *'-omic'* data used in this thesis.

The underlying hypothesis of a GWAS is that there are likely to be several susceptibility variants for common but complex diseases. This results in minor allele frequencies that are high in the population rather than a single gene disorder. This is better known as the "*common disease/common variant*" hypothesis (Reich and Lander. 2001). The focus of a GWAS is the study of single nucleotide polymorphism (SNP) frequencies within populations of interest. These single base pair changes act as genetic markers in a region of the genome that may be involved in disease presentation. This can be directly, as a functional SNP, or indirectly as a tag SNP that is in linkage disequilibrium with an influential SNP. Therefore, significant SNP associations detected from GWAS cannot be assumed to be causal variants. As these genetic variants are so common in the population the effect size of a single variant must be small, therefore, inheritance of a single SNP may only confer a small change in risk for the woman in pregnancy. This study will examine different SNPs that may confer risk.

4.2 Methods

Population

From the 128 women included in this study, 127 women had whole blood available for GWAS analysis. As a case-control design was used for GWAS we were unable to perform a three-way comparison for sPTB, PPROM and our control group of women delivering >37 weeks (TERM). Therefore, cases of sPTB and PPROM were combined to create a single sPTB group for analysis of GWAS data. The gestational cut-off of <37 weeks was used for "cases" and therefore early and late sPTB and PPROM were combined (see definitions Table 3.3).

Allele frequencies can differ between groups of people with different ethnic backgrounds, and multiple 'subpopulations' within a dataset can lead to false positive associations and/or conceal true associations. This is known as population stratification and is an important source of bias in GWAS studies (Marees et al. 2018). Approximately 90% of our recruited population report as Caucasian (Table 3.5). Hapmap3 Caucasians (CEU), Han Chinese (CHB), Japanese (JPT) and Yoruba (YRI) were included as reference populations and participants that anchored against the Caucasian population were retained in the study to increase homogeneity of our small sample size. All (genotyped) non-Caucasians were excluded.

DNA extraction and genotyping

DNA was extracted from the whole blood samples using the Chemagic Magnetic Separation Module I (Perkin Elmer) machine in runs of 12 at the Wolfson Centre for Personalised Medicine, University of Liverpool. Quantification of DNA per sample was subsequently performed under the supervision of Dr Laurence McEvoy using PicoGreen® fluorometric methods and normalised according to Oxford Genomics Centre laboratory specifications with at least 0.5 ug of DNA as a

concentration of 10 ng/ul. (Appendix G) Samples were shipped to Oxford Genomics Centre, Oxford University for genome-wide genotyping using the UK Biobank Axiom Array (Affymetrix).

The UK Biobank Axiom array chip raw data was saved in PLINK file formats, a tool for handling SNP data. Originally these were stored in .ped and .map PLINK files, and transferred to binary PLINK file format (.bed, .bim and .fam files). Quality control steps and analysis were performed using PLINK software version 1.07 (http://pngu.mgh.harvard.edu/purcell/plink/) on the bioinf1 cluster (www.bioinf.liv.ac.uk) accessible on the University of Liverpool campus (Appendix H). The quality control steps described below involved the identification and removal of DNA samples and markers that introduce bias to the study. These critical steps are paramount to the success of GWAS case-control study are well described in the literature, and are necessary before statistically testing for association (Anderson et al. 2010). These steps were performed with the help and guidance from Dr Eunice Zhang at the Wolfson Centre for Personalised Medicine.

Quality Control of Samples

The quality control (QC) filtering of single nucleotide polymorphisms (SNPs) is an important step as part of GWAS studies. This ensures minimisation of potential false positives. SNP QC uses expert guided filters based on QC variables such as Hardy-Weinberg equilibrium (HWE) for controls (HWE $p \leq 1x10^{-6}$) and mean allele frequency (MAF) <0.01 for cases and controls to remove SNPs with insufficient genotyping quality.

Identification of individuals with discordant sex information

Genetic gender was compared to clinically reported gender. As males only have a single X chromosome, genotypes of SNPs on chromosome X can be estimated to be homozygous. Females are expected to have an X chromosome homozygosity estimate (XHE) of <0.2 and males >0.8-1. As all participants in this study are female, any XHE >0.2 was considered to be a contaminant from another sample or a mix up and marked for exclusion.

Low DNA quality - sample genotyping call rates

Samples of low DNA quality or concentration often have below average genotyping call rates and genotype accuracy. The genotype failure rate is a measure of DNA sample quality. Files were created to identify missing genotypes per sample (.imiss file) and per SNP (.lmiss). A pre-specified threshold of > 95% individual call rate was required for inclusion in the analysis.

Heterozygosity Assessment

An excessive individual heterozygosity rate can indicate sample contamination with DNA from another individual, and excessively reduced heterozygosity can signify inbreeding, both of which require removal to ensure data quality. The heterozygosity rate was calculated for each individual. The threshold used was 5 ± standard deviations (SD) to include as many samples as possible.

Identity by Descent (IBD)

Related individuals will share more alleles than what is expected by chance and it is important to exclude related individuals to avoid over representation from a particular genome. The degree of recent shared ancestry for a pair of individuals (IBD) was estimated. The expectation is that IBD = 1 for monozygotic twins, IBD =

0.5 for first degree relatives, IBD = 0.25 for second degree relatives and IBD = 0.125 for third degree relatives. Due to genotyping error, population structure and linkage disequilibrium there is often some variation around these theoretical values and it is typical to remove one individual from each pair with IBD > 0.1875 (Anderson et al. 2010).

Pooled Data Analysis

To try to ensure good quality data in our GWAS we chose a cut off for sPTB of <37 weeks instead of <34 weeks. Given the relatively small sample size for this group we compared our results to the largest published GWAS for preterm birth to assess whether our results showed similar findings, albeit without reaching a significance threshold. A pooled data-analysis with published GWAS data (Zhang et al. 2017) was then performed to test whether our data shows similar findings for the SNPs that have previously reached genome wide significance (GWAS) in a mostly low risk population. The R script used for this analysis has been included in the appendix (Appendix J). This work was performed with Dr. Till Andlauer at the Max-Planck Institute, Munich.

Statistical Analysis

Descriptive statistics were performed for this case-control cohort. SPSS v.22 was used to examine the clinical variables of the cohort for effects on sPTB outcome. Continuous variables; age, BMI and cervical length, were analysed using analysis of variance test (ANOVA). For binary data, cervical surgery and smoking, a chi-squared test was used. Gestational age at delivery is reported for interest as median and range and compared with Mann Whitney U test.

Plink v.107 was used to perform the QC steps and add the binary file sets and generate the files necessary for imputation. The imputation was performed by PhD

student Juhi Gupta using the free online Michigan Imputation Server (Das et al, 2016), available at https://imputationserver.sph.umich.edu/index .html using Minimac3 (Howie et al. 2012). Minimac is a low memory, computationally efficient implementation of the MaCH algorithm for genotype imputation and it can handle very large reference panels. The genotype data were imputed against the Haplotype Reference Consortium (HRC) (Version r1.1 2016). The HRC is the largest and most relevant available panel consisting of 64,976 haplotypes of predominantly European ancestry. Additional phasing tools Eagle (v2.3.2) and ShapeIT (v2) allow pre-phasing of input haplotypes for improved imputation accuracy (Delaneau et al. 2011). Initially an extensive QC process is performed. This is to ensure the SNPs have the correct ID (n = 625518), same position in the genome (n=634721), SNPs are removed if there is a mean allele frequency difference of greater than 20% (n=336) and palindromic SNPs are removed (n=2713).

Manhattan plots were generated using RStudio v.3.1.1. (R, 2017). The threshold for genome-wide significance is 5×10^{-8} and a suggested significance threshold of $p= 1x10^{-5}$ was applied. LocusZoom (v0.4.8), a web based plotting tool (Pruim et al. 2010) was used to focus on the regions of potential genomic interest. Statistical software R (http://cran.r-project.org/) was used for all graphical representation of the results.

Pooled data analysis was performed using RStudio v 3.1.1. (R, 2017) and summary statistical outcomes of the top 10,000 SNPs from Zhang *et al.* (2017) were obtained from the GeneStation repository (www.genestation.org/analysis/gwas/ Zhang_2017/discovery).

4.3 Results

From a pool of 127 participants, 82 individuals remained for analysis (35 cases and 47 controls) following all QC steps.

Low DNA quality

In the total dataset there were 830,115 SNPs present originally. After frequency and genotyping pruning 644,287 SNPs were remaining.

Discordant sex information

One sample failed gender check. This sample failed to meet the XHE of <0.2 but did not score >0.8 to test male. Given the nature of the study it is certain the sample that we collected and processed in Liverpool is female, and inclusion of the sample was considered. However not all processing of the sample occurred in Liverpool and sample contamination or mix up could not be excluded. Additionally, the quality of the DNA in the sample may be insufficient for reliable results and therefore the sample was excluded.

Heterozygosity Assessment

No samples were excluded after applying an individual genotyping call rate and heterozygosity rate QC step (Figure 4.1).

Identity by Descent (IBD)

Four participants samples were excluded due to cryptic relatedness with an identity by descent (PI_HAT) threshold score >0.1875 when paired with another sample (Figure 4.2). The individual removed from the pairs identified are shown in Table 4.1.

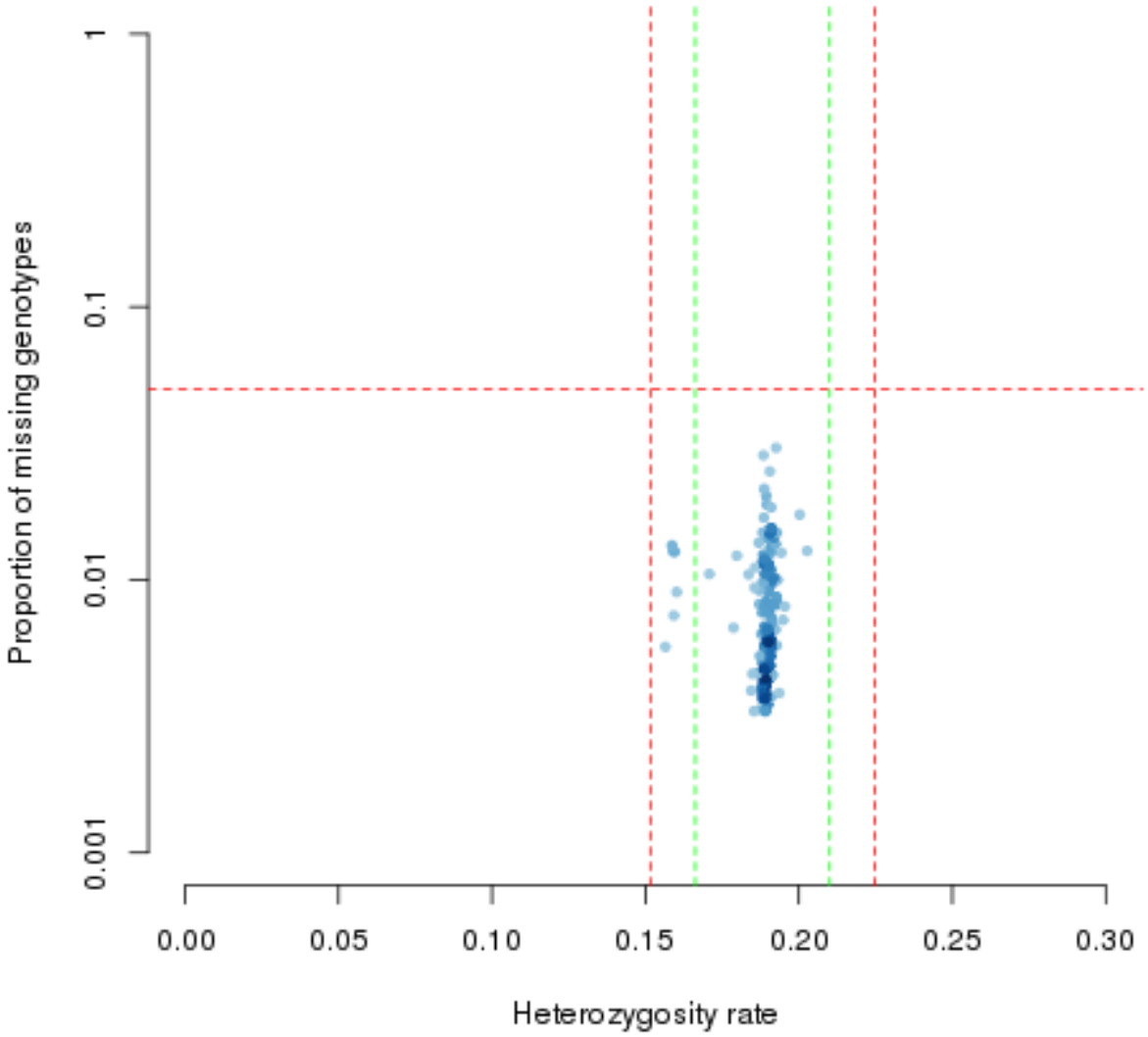


Figure 4.1 Individual genotyping call rate and heterozygosity rate. *The horizontal red dashed line indicates the 95% genotyping call rate applied, the vertical red dashed lines indicate 5 ± SD and the green dashed lines indicate 3 ± SD of heterozygosity rate.*

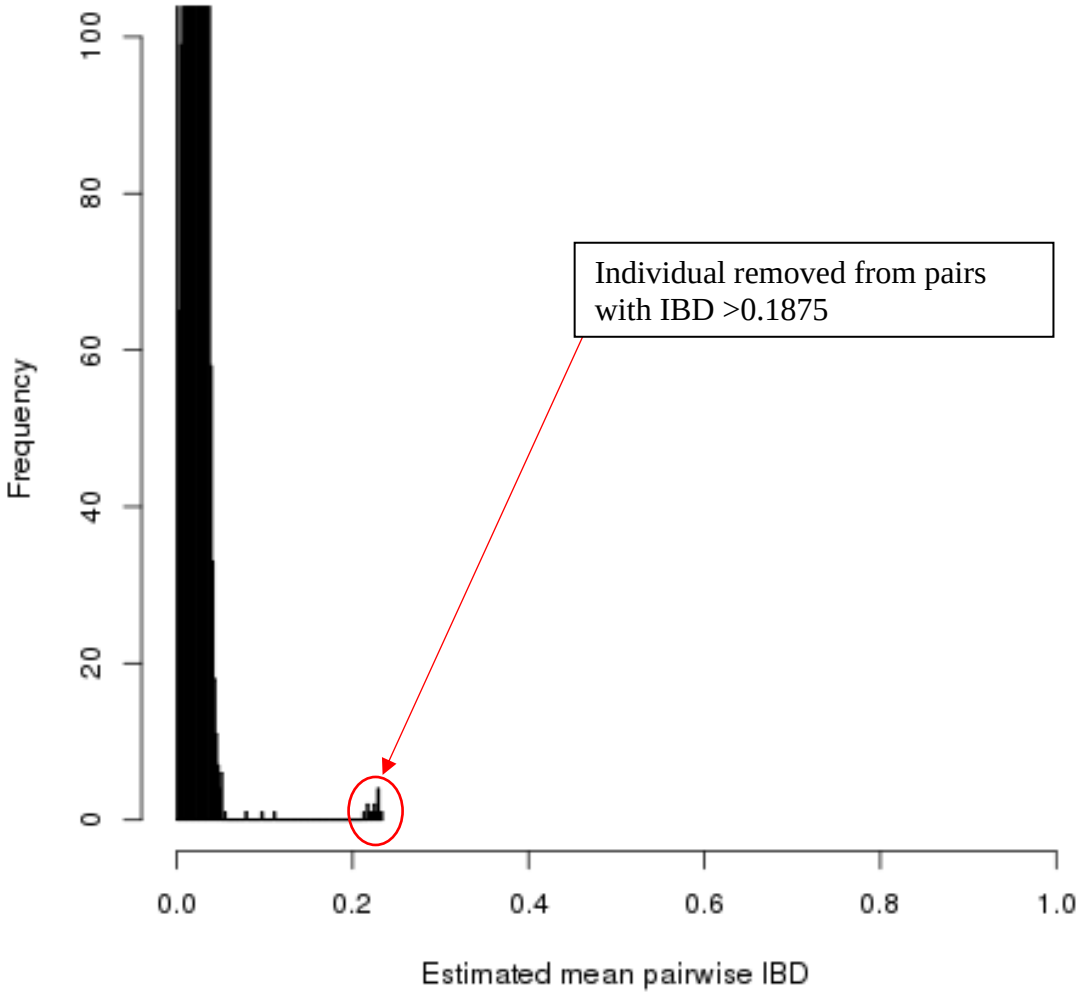


Figure 4.2 Estimated mean pairwise Identity by descent (IBD) of each participant.

Table 4.1 Pairwise comparison of individuals with shared ancestry (PI_HAT >0.1875) F_MISS indicates the missing call rate. PI_HAT indicates the identity by descent threshold. The shaded participants were recorded for removal at the end of the quality control pro

Individual ID 1	F_MISS	Individual ID 2	F_MISS	PI_HAT
PTB_11A	0.01272	PTB_107A	0.009018	0.2213
PTB_11A	0.01272	PTB_23A	0.007393	0.2192
PTB_11A	0.01272	PTB_31A	0.01267	0.2164
PTB_11A	0.01272	PTB_60B	0.01336	0.2306
PTB_11A	0.01272	PTB_62B	0.005675	0.2288
PTB_107A	0.009018	PTB_23A	0.007393	0.2139
PTB_107A	0.009018	PTB_31A	0.01267	0.2164
PTB_107A	0.009018	PTB_60B	0.01336	0.2231
PTB_107A	0.009018	PTB_62B	0.005675	0.2253
PTB_23A	0.007393	PTB_31A	0.01267	0.2266
PTB_23A	0.007393	PTB_60B	0.01336	0.2286
PTB_23A	0.007393	PTB_62B	0.005675	0.2249
PTB_31A	0.01267	PTB_60B	0.01336	0.2281

Ethnicity

Twenty-three additional participants were excluded after being identified as ethnic outliers according to their projection onto the reference populations (Figure 4.3). One hundred participants that clustered within the Caucasian population were retained in the study (n=100). This was ten less than expected from self-reported ethnicity (Table 4.2).

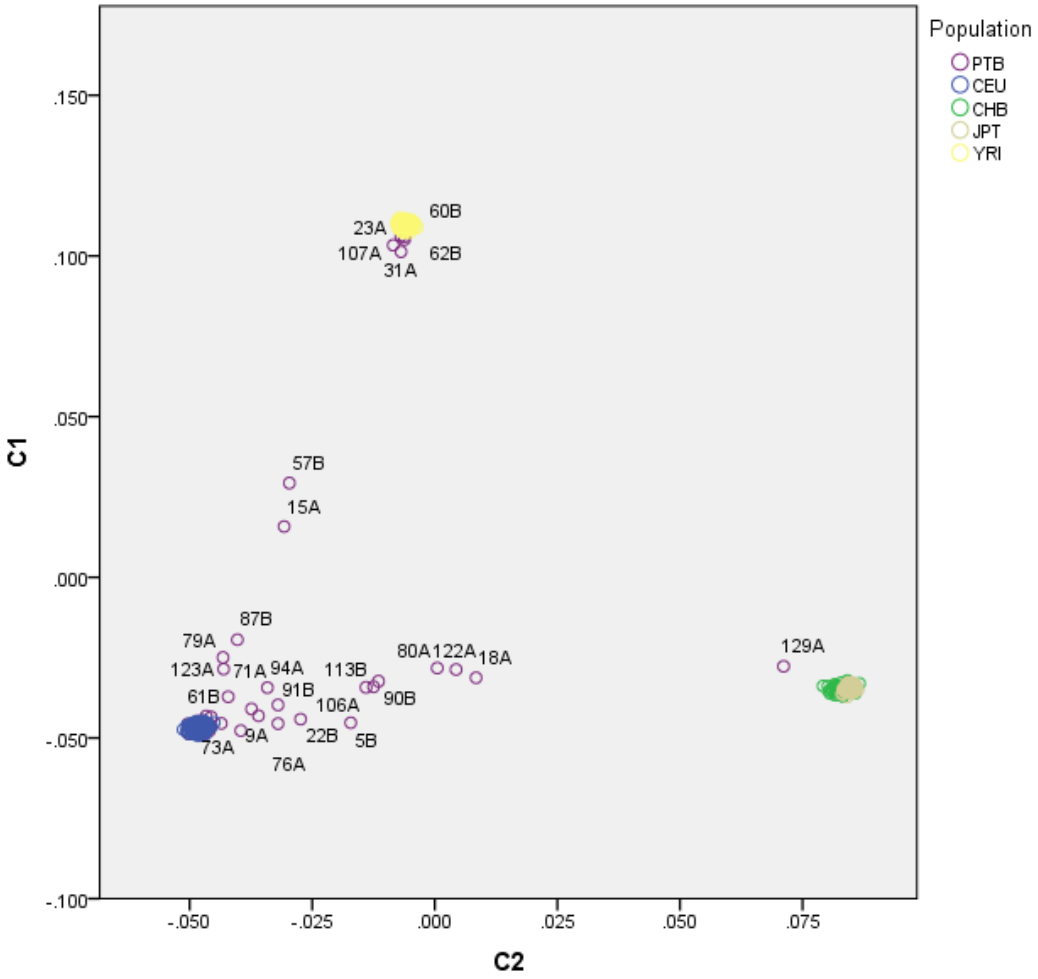


Figure 4.3. Principal component analysis (PCA) of genetic ethnicities of PTB pilot study participants. (C1 = Principal component 1, C2 = Principal component 2). Study samples are labelled in purple, most overlap with blue circles (Caucasian Hapmap population) 27 ethnic outliers were identified for removal. *PTB - Liverpool PTB Clinic Cohort; CEU* - Utah residents with Northern and Western European ancestry from the CEPH collection; *CHB* - Han Chinese in Beijing; China *JPT* Japanese in Tokyo, Japan; *YRI* - Yoruba in Ibadan, Nigeria

Table 4.2. Self-reported ethnicities compared to genetic ethnicities

	Self-Reported Ethnicities	**Genetic Ethnicities**
Caucasian (White British)	110	100
Caucasian (other)	2	Excluded
African Origin / YRI (Black British)	6	Excluded
Chinese	1	Excluded
Asian-Bangladeshi/Indian/Sri Lankan/Other	4	Excluded
Not reported	2	Excluded
Mixed ethnicity	2	Excluded

Population

The seventeen cases treated for short cervix were removed from the final quality control file before imputation for reasons discussed in chapter three.

Participant Characteristics

Demographic and pregnancy related outcome for all preterm birth participants are described in chapter 3. Table 4.3 describes just the participant characteristics for the 35 cases and 47 controls whose data was available for this GWAS analysis.

The groups show no demographic differences in age, BMI, parity and smoking rates. Of note are six women who have had cervical surgery in addition to a previous pregnancy loss (4 single LLETZ, 2 multiple LLETZ and 1 knife cone biopsy) who are ‘cases’ due to a subsequent preterm birth. This is compared to zero

Table 4.3. Demographic and clinical characteristics for preterm birth biomarker study participants included in genome wide association analysis

	sPTB <37 weeks (n=35)	Term delivery ≥37 weeks (n=47)	P value
Maternal Demographics			
Age (SD)	29.0 (5.0)	30.4 (4.8)	.216
BMI (SD)	25.4 (4.9)	24.2 (4.2)	.226
Parity (SD)	1.4 (1.0)	1.7 (1.6)	.296
Cervical surgery	6 (17%)	0	.016
Smoking in Pregnancy	9 (25.7%)	13 (28%)	.503
History of sPTB	21 (60%)	20 (43%)	.205
History of PPROM	12 (34%)	24 (51%)	.130
Pregnancy Features			
Cervical length at 16 weeks (SD)	32.4 (8.0)	36.83 (6.5)	*.009*
Cervical length at 20 weeks (SD)	26.32 (10.7)	37.28 (6.3)	*.000*
Delivery Outcomes			
Gestational Age at Delivery (median, range)	33^{+5} (17^{+2} – 36^{+6})	39^{+3} (37^{+0} – 41^{+5})	.000

women with a history of cervical surgery in the term pregnancy group, however this difference is not shown to be statistically significant (p 0.16).

There is an expected statistically significant difference between cervical length at 16 and 20 weeks between the sPTB group and the term cohort. Cervical length is a known risk factor for sPTB and women with a cervical length below the 3rd centile

receive preventative treatment. Participants who required treatment for a short cervix, and reached term, possibly suggesting successful treatment, have been removed from the term/control cohort. Thus, artificially increasing the difference in cervical lengths between these groups. Interestingly the mean cervical length difference between 16 and 20 weeks is approximately a change of 6 mm shorter in the high-risk group compared to no change in the term delivery group.

Genome Wide Association Analysis

SNP frequencies of women with recurrent sPTB <37 weeks were compared to women with only a history of sPTB <34 weeks. No SNPs reached genome wide significance (Red line on Figure 4.4). However top SNPs on chromosome 3 and 16 fell just below genome wide significance and some interesting SNPs and SNP stacks can be seen in chromosome 11, 12, 21 and 22 which were investigated further.

On the Manhattan plot (Figure 4.4) the x -axis shows each SNP position organised by chromosome, shown in black and grey blocks. On the Y-axis is the negative logarithm of the P-value for association of sPTB in cases compared to controls for each single SNP. Each block is composed of thousands of dots, each dot representing one SNP. The strongest associations with the trait have the smallest p values, therefore their negative logarithms will be the largest number on the Y axis.

The areas of interest include SNPs that fall just below genome wide significance but also the appearance of stacked towers of SNPs suggesting that potentially multiple SNPs from the same chromosome region or even gene are associated with the sPTB trait.

These regions of interest on the Manhattan plot were then examined using LocusZoom, to identify the SNPs of interest above the suggestive threshold. (Figures 4.5-4.9) The SNPs were mapped back to their genes to identify if the biological

function of the gene made the gene target a possible candidate for a risk profile of sPTB in pregnancy.

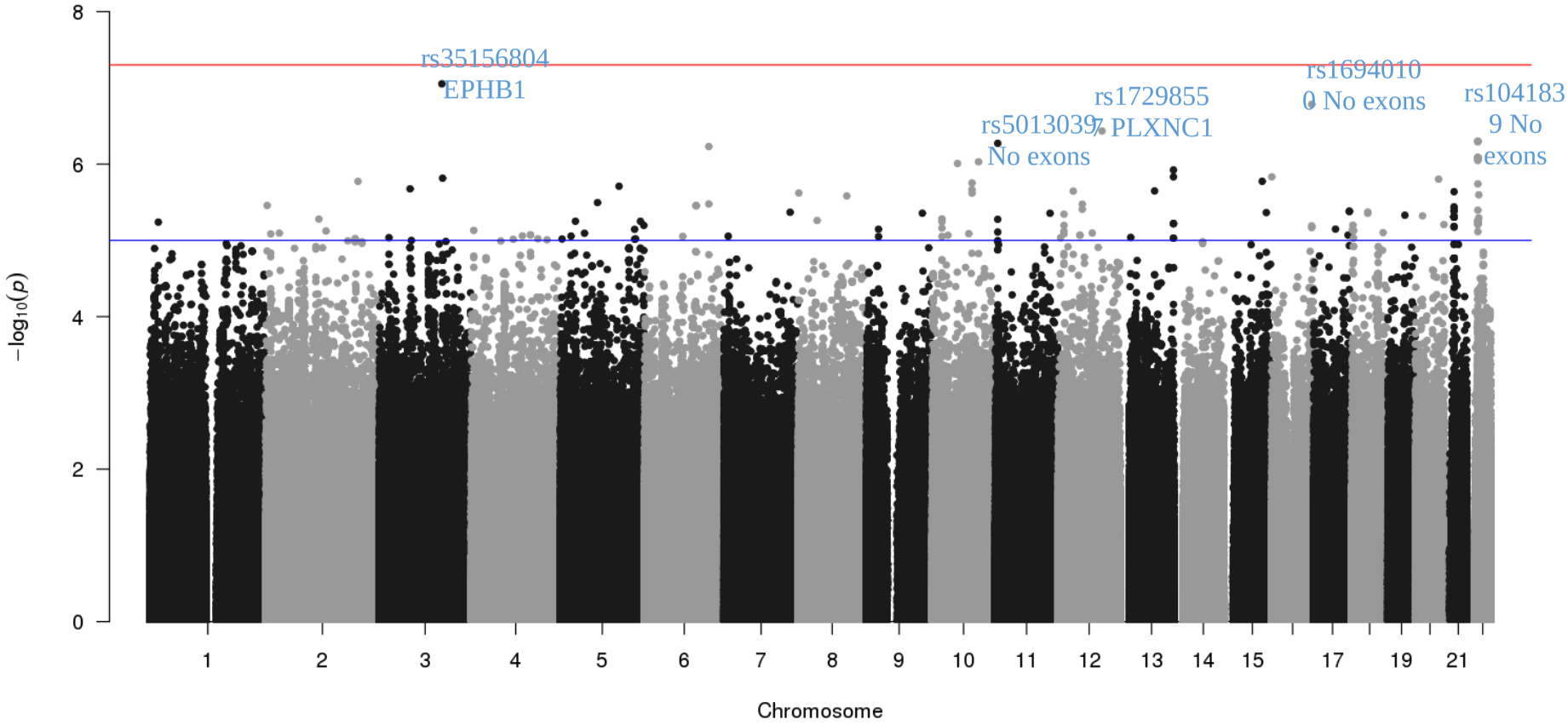


Figure 4.4. Manhattan plot of genome wide distribution. Each dot represents a SNP associated with sPTB. There are no SNPs reaching the genomewide significance threshold level (red line). The blue line shows a suggestive threshold level to investigate SNPs of potential interest. The top SNPs are highlighted along with their mapped genes.

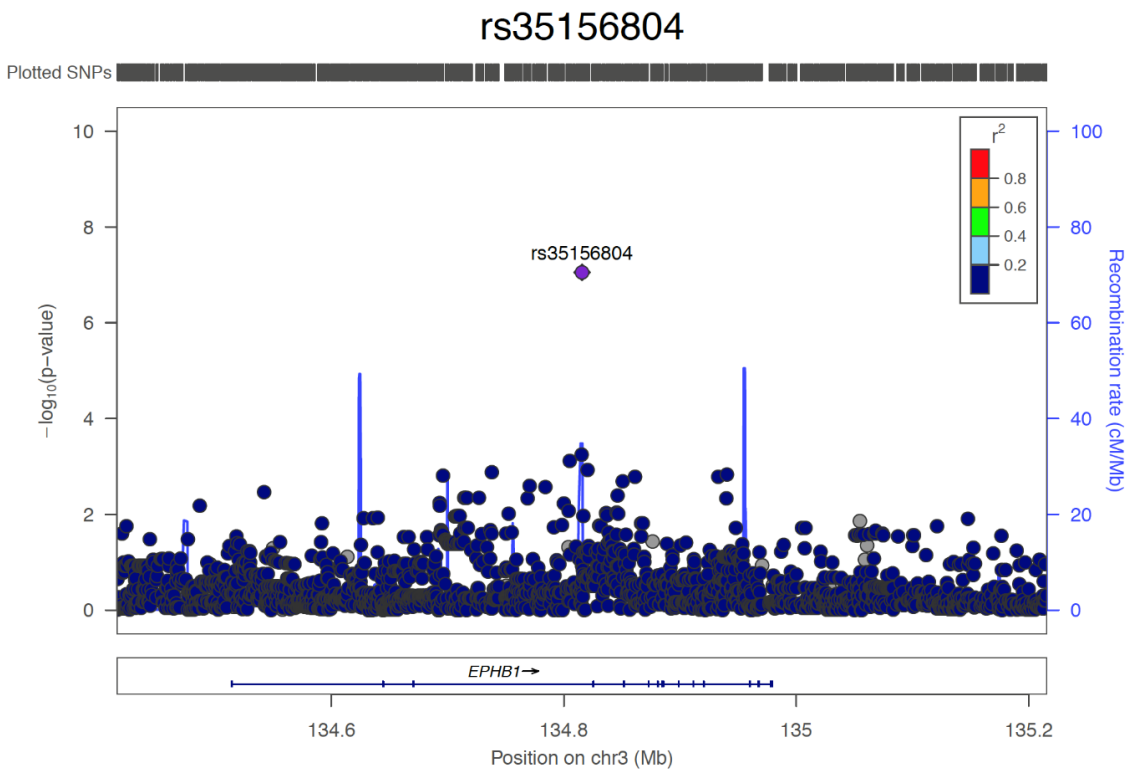


Figure 4.5 Regional LocusZoom plot of top hit from the Manhattan plot in Figure 4.4 on chromosome 3

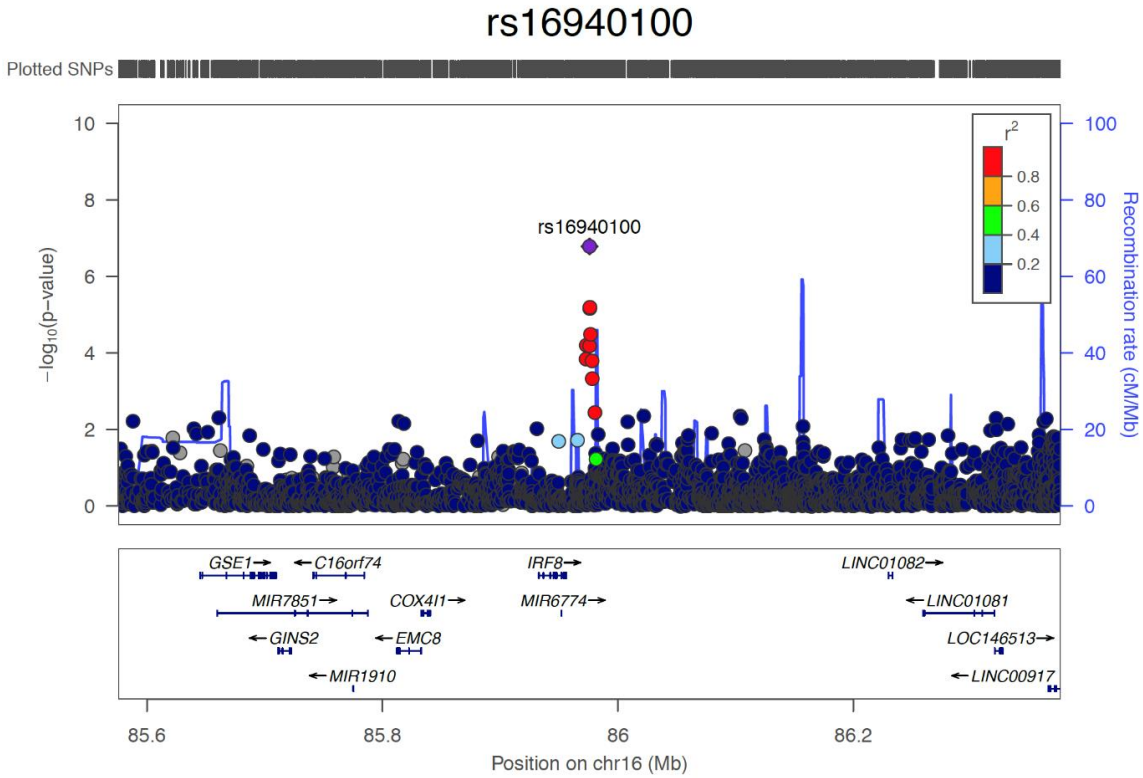


Figure 4.6 Regional LocusZoom plot of top hit from the Manhattan plot in Figure 4.4 on chromosome 16

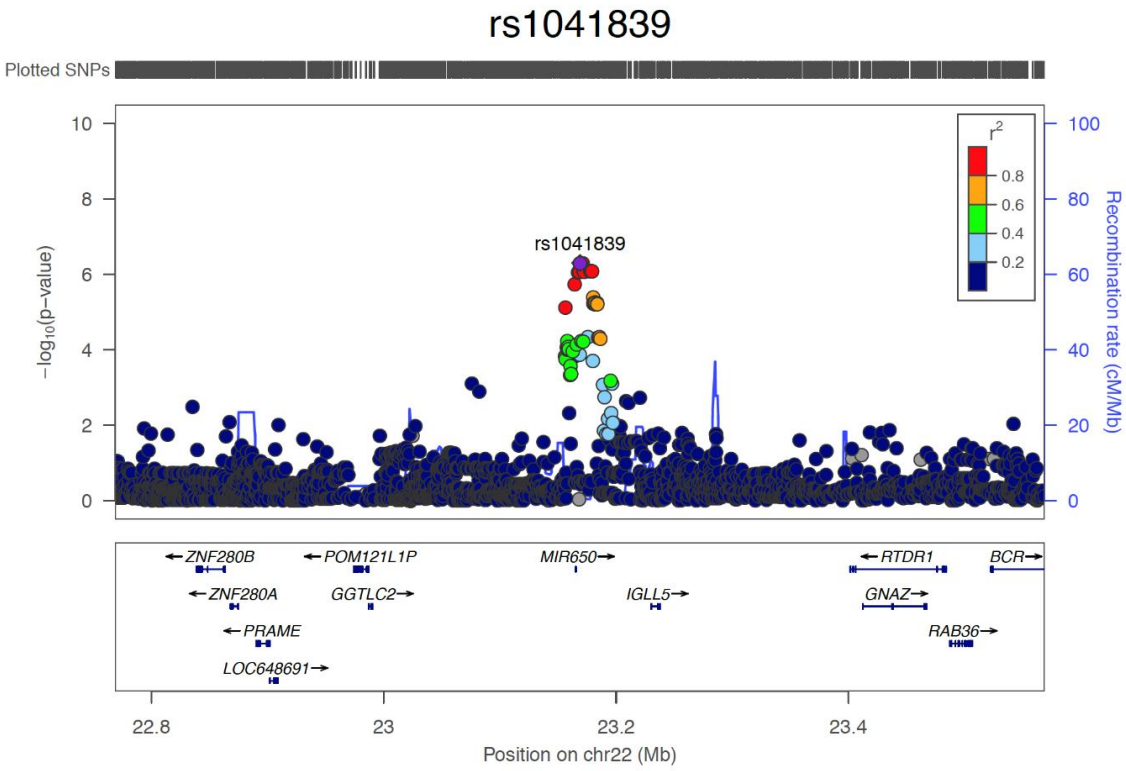


Figure 4.7 Regional LocusZoom plot of top hits from the Manhattan plot in Figure 4.4 on chromosome 22

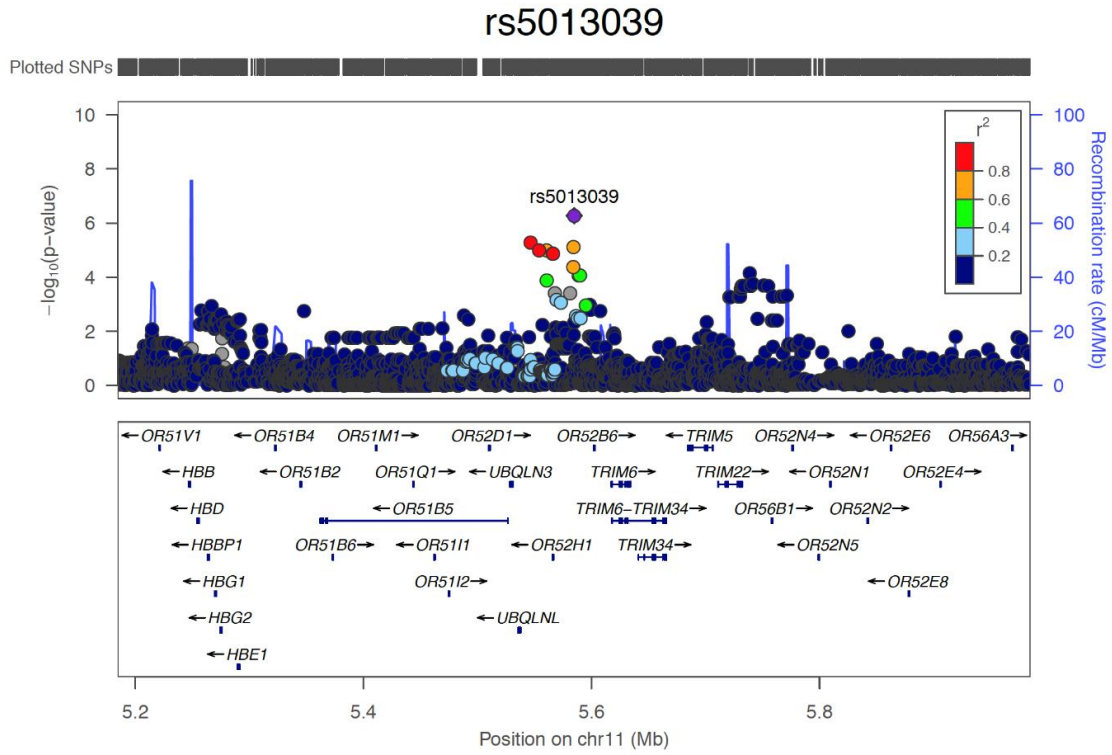


Figure 4.8. Regional LocusZoom plot of top hits from the Manhattan plot in Figure 4.4 on chromosome 11

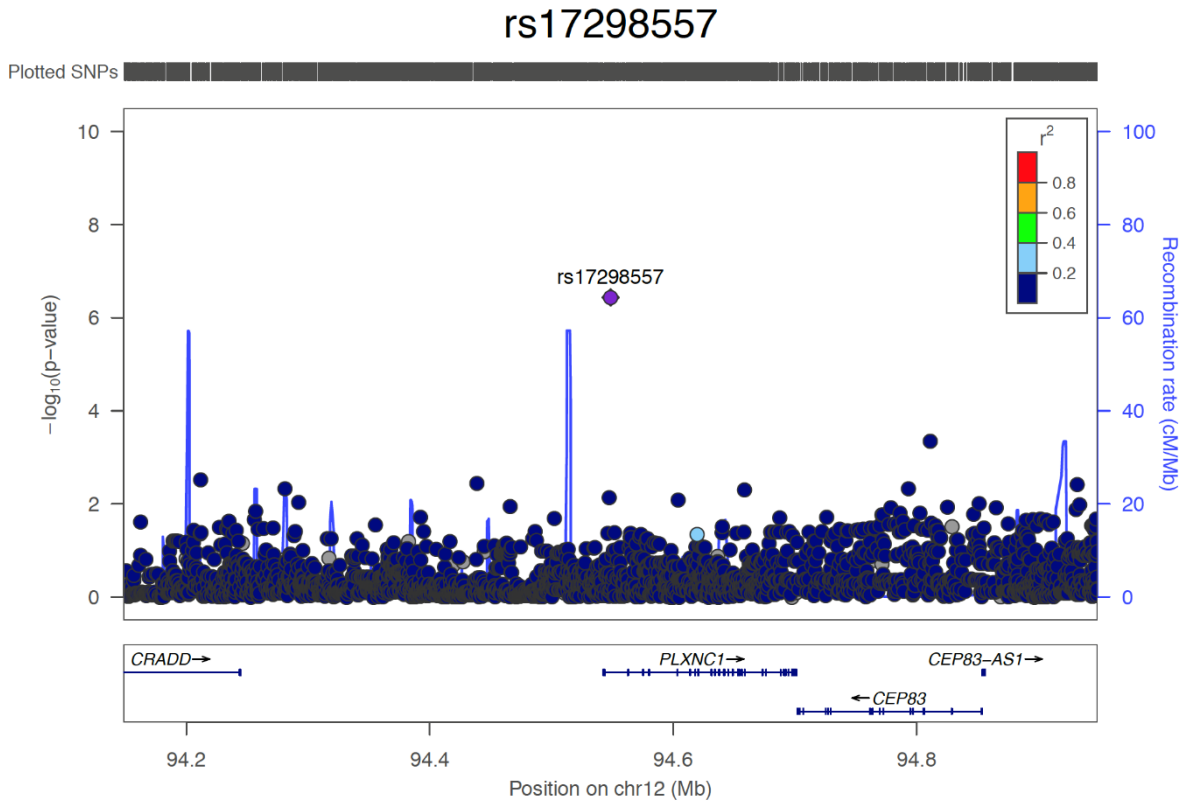


Figure 4.9 Regional LocusZoom plot of top hits from the Manhattan plot in Figure 4.4 on chromosome 12

Pooled Data Analysis

Figure 4.10 is taken from the GWAS publication by Zheng *et al*. (2017) who also used <37 weeks as their cut off for sPTB. They identified two genes, EEFSEC and EBF1 (shown in Figure 4.10) associated with sPTB above the genome-wide threshold level. Pooled data analysis with the same SNPs used to tag these genes using both published data (Zhang et al. 2017) and Liverpool data and found that only EEFSEC shows similar odds ratios (OR) in our population. Our data remains inconclusive for EBF1 due to wide confidence intervals. (Table 4.4; Figures 4.11).

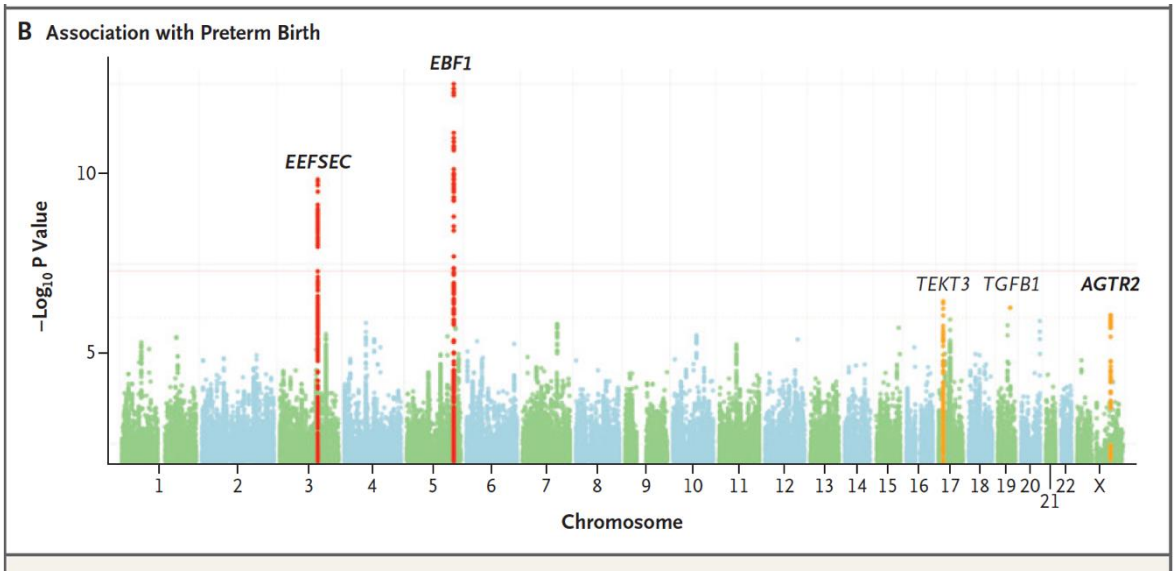


Figure 4.10. Figure from Zhang et al. "Genetic Associations with Gestational Duration and Spontaneous Preterm Birth", N Engl J Med 2017;377:1159 demonstrating SNPS that reach genome wide significance in red and those genes meeting a suggestive threshold in orange.

Table 4.4. Comparison of odds ratios for sPTB SNPs associated with EEFSEC and EBF1 from Zhang et al. (published data) and Liverpool data from this thesis.

	Alleles	**Zhang et al. (2017) Odds Ratio (95% CI)**	**Liverpool Odds Ratio (95% CI)**
EEFSEC			
rs2955117	A/G	1.20 (1.14 – 1.26)	2.01 (1.28 – 2.74)
EBF1			
rs2963463	C/T	1.23 (1.18-1.28)	0.93 (0.19-1.66)

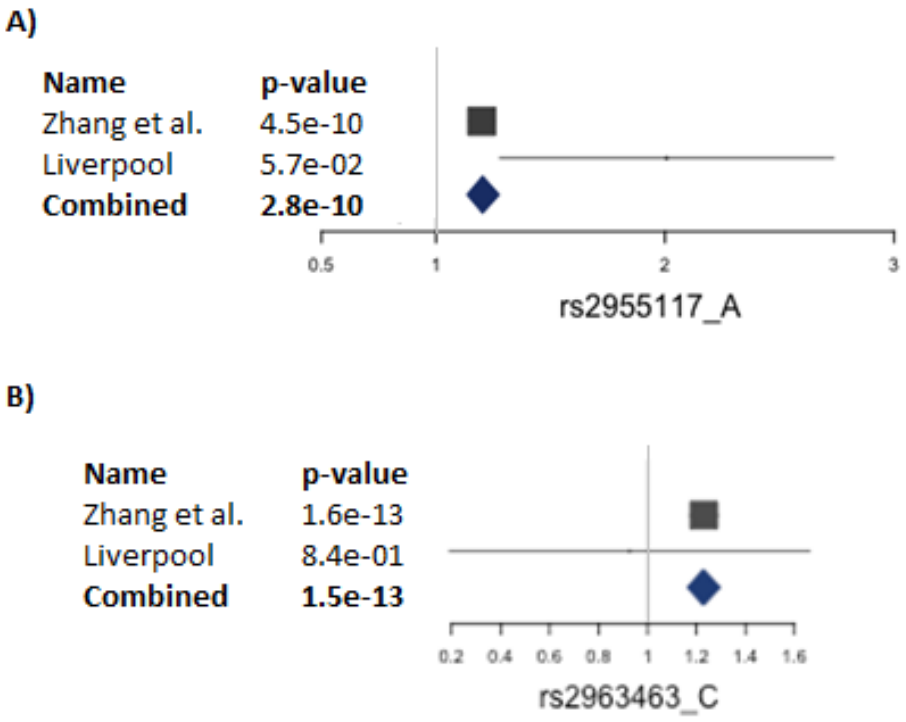


Figure 4.11. Forest plot demonstrating comparison of odds ratio (OR) band confidence intervals (CI) between Liverpool and Zhang et al. 2017 data for A) SNP rs2955117_A mapped to the EEFSEC gene and B) SNP 2963463_C mapped to the EBF1 gene.

4.4 Discussion

Our study did not reveal any genes with genome wide significance, but several SNPs were identified above a pre-specified threshold level for significant association. The top SNPs and SNP towers were investigated on chromosomes 3, 11, 12, 16 and 22. All but rs35156804 on chromosome 3 and rs17298557 on chromosome 12 were identified as intronic variants and did not map directly to a known gene.

SNP rs35156804 mapped just under the genome wide significance level was located to chromosome 3 on EPHB1 gene. EPH receptors are the largest family of receptor tyrosine kinases (RTKs) and are divided into two subclasses, EPH A and EPH B (Figure 4.12). Originally, they were identified as mediators of axon guidance but now EPH receptors are implicated in many processes, including injury and inflammation (Ivanov and Romanovsky. 2006).

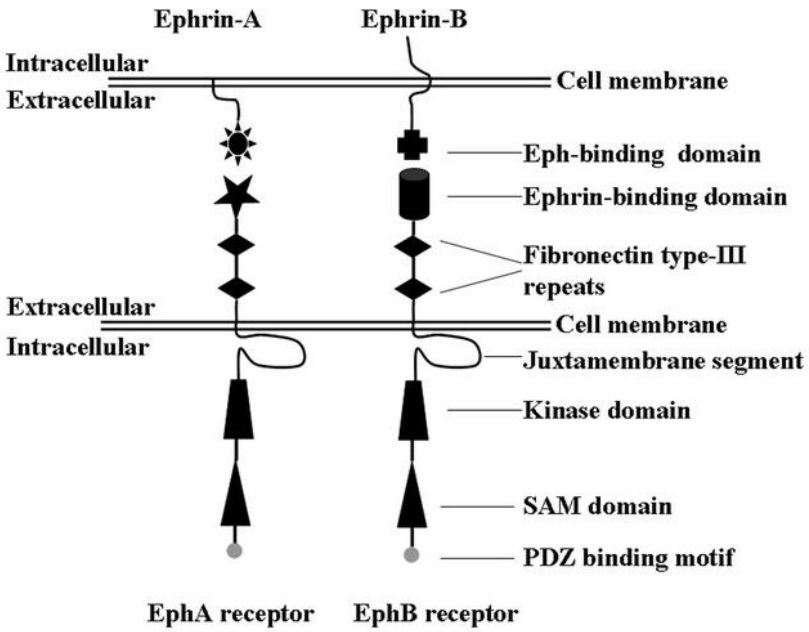


Figure 4.12 Domain structures of EPH receptors and ephrins. Figure taken from Wei, W., Wang, H., & Ji, S. 2017. Paradoxes of the EphB1 receptor in malignant brain tumours. Cancer Cell International, 17, 21.

Their role in inflammation and the immune system remains unclear. A study found that ephrin-B1 is highly expressed in peripheral blood lymphocytes (PBLs) obtained from patients with rheumatoid arthritis (Kitamura T. 2008). Ephrin-B1 ligand and EphB1 receptor are thought to play an important role in this inflammatory condition through influencing function of T cells through stimulation of the production of TNF-alpha in PBLs and IL-6 in synovial cells (Kitamura T. 2008). The function of EphB1/ephrin signalling in the development of immune organs and the corresponding mechanism of immune regulation are an area for future study and may also be implicated in pregnancy.

The EPHB1 gene has been previously linked to sPTB but only, at present, in animal models investigating mechanisms of uterine stretch and over-distension. One study using a non-human primate model of pigtail macaques demonstrated that EPHB1 was significantly downregulated in uterine myometrium that had been excessively stretched using balloon catheterisation to trigger preterm labour compared to controls when measuring mRNA levels (Waldorf et al. 2015). Therefore, this provides a rationale to explore the mechanistic role of EPHB1 further in reproductive tissues.

SNP rs17298557 mapped to gene PLXNC1 on chromosome 12 (Figure 4.9). However, there was no evidence of a tower of SNPs to suggest that other SNPs in this area or on this gene showed an association between cases and controls making it less likely that this gene is truly associated with preterm birth. PLXNC1 has been linked to involvement in inflammatory response which suggests potential biological plausibility behind this finding. This gene encodes a member of the plexin family. Plexins are transmembrane receptors for semaphorins, a large family of proteins that regulate axon guidance, cell motility and migration, and immune response (NCBI

2018. Accessed https://www.ncbi.nlm.nih.gov/ gene/10154). There are, to date, no published associations of this gene in pregnancy in either human or non-human models, therefore further validation of this finding would be required before planned investigation of its role in sPTB is taken further.

Pooled data analysis of Liverpool data compared to Zhang *et al.* (2017) investigated the performance of just a couple of SNPs on known genes of interest rather than all SNPs. In the EEFSEC gene; the Liverpool data shows strong agreement with the directionality of this gene with an odds ratio and 95% CI of 2.01 (1.28 – 2.74) for sPTB <37 weeks for SNP rs2955117 (Figure 4.11). This gene encodes selenocysteine tRNA-specific eukaryotic elongation factor which is aides in the incorporation of selenocysteine into selenoproteins. The physiologic functions of selenium have been linked both to the parturition process and preterm birth (Rayman et al. 2011, Zhang et al. 2017). Additionally, in a population of preterm very low birth weights neonates they have been found to be deficient in selenium at birth and supplementation at 10 μg/day reduces their risk of late-onset sepsis (Aggarwal et al. 2016). Therefore, we would agree that further evaluation of the role of maternal selenium micronutrient status on prematurity risk should be investigated, both in the general obstetric population and specifically within the Liverpool population.

Interestingly, our data was not able to support the association of EBF1 involvement with sPTB within our population with an OR (95%CI) of 0.93 (0.19-1.66) compared to Zhang *et al.* OR (95%CI) 1.23 (1.18-1.28). There is a large confidence interval around the Liverpool OR and these data are largely inconclusive with relation to this gene and its association with sPTB in the Liverpool population. This is most likely to be due to weaknesses in our study design with very low numbers in this case control study, which only makes the strong agreement of the

directionality of the EEFSEC gene more interesting. Alternatively, this difference may be related to the comparison of a high-risk population (Liverpool) with a low risk population (Zhang et al. 2017).

EBF1 encodes for early B-cell factor and is important for B-cell development. It has been associated with control of blood pressure and metabolic risk (Zhang et al. 2017) and therefore may contribute to preterm birth more generally through these pathways that influence gestation, rather than pregnancy or sPTB specific pathways. A high-risk population with recurrent preterm births such as ours may not select for this gene as strongly as a low risk or general pregnancy population.

There are two primary platforms in chip-based microarray technology for assaying upwards of one million SNPs. These two competing technologies, Affymetrix (Santa Clara, CA) and Illumina (San Diego, CA) will choose different SNPs for their assays and use slightly different technologies. The UK Biobank Axiom Array (Affymetrix) was chosen for this project due to the overall genomic coverage which comprises of 820,967 genetic markers. This is also the array used by the UK Biobank (www.ukbiobank.ac.uk), a large research biobank of 500,000 participants from the UK which could provide data for meta-analysis. As we have demonstrated results of multiple GWAS can be pooled together to perform meta-analysis, developed to examine and refine significant effect sizes of published GWAS investigating the same disease. Meta-analysis becomes increasingly difficult if studies use different genotyping platforms and different SNP marker sets, therefore we have tried to stay in line with current major health resources.

The small sample size of our study is a significant limitation that increases the likelihood of finding a false positive result with the hundreds and thousands of

multiple comparisons on only a relatively small number of cases and controls. Nevertheless, despite the small number the quality of the clinical phenotype is excellent, and this cohort is particularly unusual as the women represented by the 'cases' have had recurrent sPTB not just a single preterm birth. To obtain this type and quality of data on the scale necessary to be confident in the findings and validate SNPs of interest will take years of work and the collaboration of many groups, therefore meta-analysis is likely to become a predominant method of genomic and -*omic* research in the field of sPTB research.

The ability to extrapolate our findings into other non-Caucasian populations is another potential weakness, as we excluded non-Caucasians to reduce spurious findings. Disease associating alleles can have different frequencies in different populations as a result of demographic events, such as migration.

The most difficult decision for this –omics analysis was deciding on a preterm birth cut-off for our groups. A <34 week cut-off would be more in keeping with the other omics analysis layers (where this has been used as a cut-off), however making the number of cases smaller would have negatively affected the quality of our results and increased the likelihood of false positive findings resulting in poor data quality for the –omics combination. Consequently, a <37 week cut-off was chosen for this omic layer only. Encouragingly the SNP rs2955117_A mapped to the EEFSEC gene showed the same directionality as the largest GWAS published in this field of research, which supports our findings that this SNP falls just below the genome wide significance threshold in this high-risk population.

4.5 Conclusion

There is recent evidence from a GWAS study that supports the genetic predisposition of sPTB. Our data reinforces the finding that EEFSEC, a gene involved in the creation of selenoproteins, is associated with sPTB. Inconclusive results were found for EBF1 which may be attributed to the small sample size or due to our rare population of women with recurrent sPTB, as opposed to women with just one sPTB. We found no SNPs in our cohort that were of genome wide significance, but several SNPs fell above a suggestive threshold. A SNP from the EPHB1 gene fell just below genome wide significance and further investigation of this gene in the role of sPTB should be considered in other omic layers from this cohort; particularly as this gene has previously been shown to be downregulated in mammal models of sPTB (Kin et al. 2016).

Chapter 5: Longitudinal Transcriptomic Analysis for the Prediction of Spontaneous Preterm Birth

5.1 Introduction

Transcriptomics is the study of all RNA molecules in a cell, otherwise known as ‘the transcriptome’ (Wang et al. 2009). The biomarker potential for disease prediction of RNA was identified over two decades ago (Kusec et al. 1994, Seal et al. 1995, Gillis et al. 1995) and since then biomarker discovery in transcriptomics has occurred for many diseases, but most prominently in the field of oncology (Xi X. 2017). Most transcription studies focus on the measurement of cell messenger RNA (mRNA) and microarray technology allows examination of relative levels of gene transcripts to establish which genes are being up or down regulated at a given moment in time.

As addressed in the literature review in Chapter 2 there have been several studies looking at transcriptomics in both the threatened and asymptomatic high-risk sPTB population. A metanalysis of three studies (n=339 maternal whole blood samples; n=134 preterm, n=205 term) identified by the Gene Expression Omnibus (GEO) database found a set of 210 significant differentially expressed genes in maternal blood (Vora et al. 2018). Many of these were downregulated immune related genes. Based on the maternal data from this study, genes and cell types associated with innate immunity were upregulated in sPTB, while those relevant to adaptive immunity were downregulated. However, several limitations still exist with the data, including the small number of studies with publicly available data that can be aggregated. Samples lack demographic information as well as detailed clinical annotations and are heterogeneous, making comparison difficult. Examples of heterogeneity include the type of study (cohort or case-control), definitions and populations or phenotypes of preterm birth (i.e. asymptomatic and threatened preterm labour, early and late sPTB). Despite these limitations, a link to regulation of

immune function is a recurring theme in sPTB research and study of the transcriptome demonstrates potential for biomarker discovery.

Due to the limitations of cost only a subset of the women in this study provided RNA samples. I examined gene expression across the transcriptome, at 16 and 20 weeks of pregnancy between women with sPTB, PPROM and term deliveries (>37 weeks) in our cohort to identify the leading pathways demonstrating differentiation of gene expression profiles.

5.2 Methods

Population

Samples of whole blood were obtained from 58 participants (94 samples). 2.5ml of whole blood was stored in PAXgene Blood RNA Tube and frozen at -80^{o}C until RNA extraction. Following thawing to room temperature, total RNA was extracted using the PAX gene blood RNA Kit (PreAnalytix/QIAGEN) adhering to the manufacturer's protocol (Appendix I). Purification occurred with a centrifugation step to pellet nucleic acids in the PAXgene Blood RNA Tube. The pellet was washed and resuspended, followed by manual purification.

Quality Control of RNA prior to Hybridization

RNA quantity and purity was established using a NanoDrop (ND) spectrophotometer. The A260 value was used for RNA quantification. RNA has a maximum absorption at 260 nm and RNA concentration is determined by the following conversion; an A260 of 1.0 is equivalent to 40 µg/mL of RNA.

In addition, measurements were also taken at 280nm. The A260/A280 ratio is an indication of the level of contamination of protein, DNA, phenol, ethanol and salts in the sample. A high-quality RNA sample is free of these and contamination affects how efficiently RNA is amplified prior to hybridisation to the array chip. Pure RNA has an A260/A280 ratio of 2.1, however values between 1.7-2.2 were considered acceptable for our protocol. The ND sample values were checked prior to freezing at -80 in preparation for transfer to the Centre for Genomic Research (CGR), University of Liverpool.

The RNA integrity was established using the Agilent 2100 Bioanalyser (Agilent Technologies, Santa Clara, CA). Reverse transcribing partially degraded

mRNA can generate cDNA that lacks parts of the coding region. Therefore, only samples with an RNA integrity number (RIN) >7 were hybridised to Clariom™ D Assay, human (Affymetrix/Thermo Fischer Scientific).

RNA Amplification, Purification, Quantitation & Hybridization

For RNA amplification and gene chip cartridge array hybridization the GeneChip™ WT PLUS Reagent Kit was used and the manufacturers protocol was followed (Figure 5.1). Initially RNA controls were prepared and diluted and labelled together with the total sample RNA. The hybridization intensities of the controls help to monitor the labelling process independently from the quality of the starting RNA samples. Then first strand complimentary DNA (cDNA) was synthesised in a reverse transcription procedure. Total RNA was primed with primers containing a T7 promoter sequence. The reaction synthesized single-stranded cDNA with T7 promoter sequence at the 5' end. Then second strand cDNA was synthesised. Single stranded cDNA was converted to double stranded cDNA. This step used RNase H to degrade RNA whilst DNA polymerase synthesised the second strand to act as a template for synthesising and amplifying antisense RNA (complimentary RNA). This method of RNA sample preparation is based on the original T7 RNA polymerase *in vitro* transcription (IVT) technology known as the Eberwine or RT-IVT method (Van Gelder et al. 1990). A purification step then removed unincorporated nucleotides, salts, enzymes and inorganic phosphates in preparation for the next single stranded cDNA synthesis step. Second cycle primers including dUTP at a fixed ratio relative to

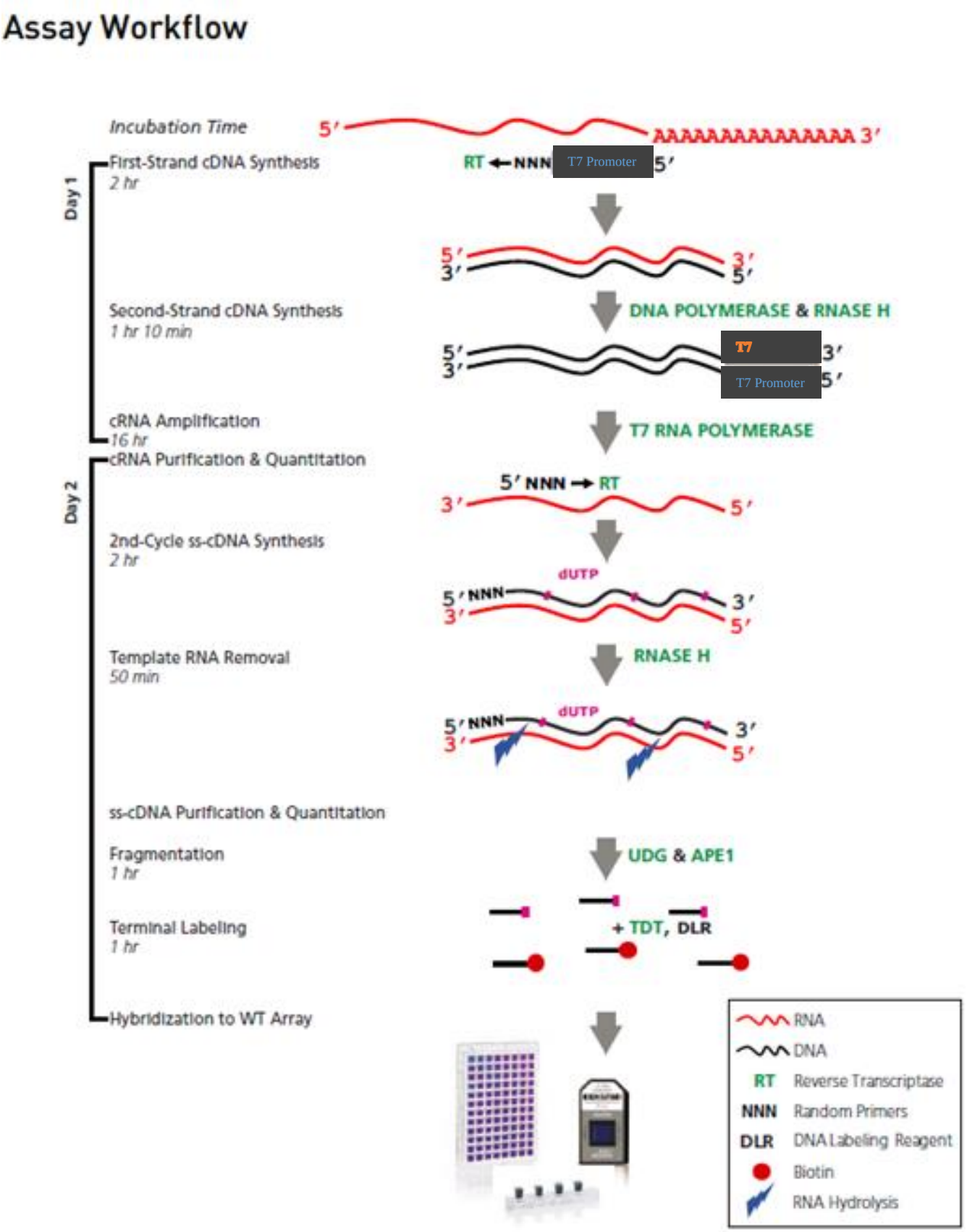


Figure 5.1. WT PLUS Amplification and Labelling Process. *Image taken from GeneChip™ WT PLUS Reagent Kit Manual Target Preparation for GeneChip™Whole Transcript (WT) Expression Arrays User Guide*

dTTP were then used in the next cDNA synthesis step. This was to allow incorporation of uracil into the cDNA strand to allow for easier fragmentation. Template RNA was then hydrolysed by RNase H leaving single-stranded cDNA. To prepare the cDNA for fragmentation and labelling another purification step removed excess salts and unincorporated nucleotides. Fragmentation then occurred by uracil-DNA glycosylase (UDG) and apurinic/apyrimidinic endonuclease 1 (APE 1) at the unnatural dUTP sites in the cDNA. The fragmented cDNA is then labelled by terminal deoxynucleotidyl transferase (TdT) and a proprietary labelling reagent covalently linked to biotin ready for chip hybridization. The chip was loaded onto the Affymetrix GeneChip™ Scanner 3000 7G for scanning.

The Clariom™ D chip was selected, as at the time of the study it had the most comprehensive cover of the human genome. Clariom D can detect greater than 540,000 transcripts, which was the most comprehensive coverage of the human transcriptome at the time of the study. This was to maximise discovery of any actionable biomarkers and ensure any rare or low expressing transcripts not detected by other methodologies were not missed.

Affymetrix® Expression Console™ software was required to perform QC analysis of transcriptome data and was downloaded from the Affymetrix website as part of the Transcriptome Analysis Control (TAC) Software. (https://www.thermofisher.com/uk/en/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/affymetrix-expression-console-software.html). Probes were annotated by this software using NetAffx information associated with this particular probe set. Outliers were identified if they had two standard deviations away from the mean of the metric values for this experiment. Hybridizations that consistently had metric

values at the tails of distribution were removed to prevent problematic downstream analysis. An overview of the quality control steps performed by CGR are shown in Figure 5.2.

Once the metrics were run, a link was prepared by the Centre for Genomics Research to access the data. Further bioinformatic analysis was performed by Juhi Gupta (University of Liverpool) and Professor Bertram Müller-Myhsok (Max-Planck Institute, Munich, Germany and University of Liverpool). Microarray CHP files were normalised using Robust Multi-array Average and scaled (Bioconductor, R).

Once the QC steps were completed, clinical data and phenotypic classification was used to remove cases that did not qualify as sPTB <37 weeks, PPROM <37 weeks or TERM delivery. Only samples that had both timepoints were included in the analysis.

Differential gene expression was initially performed between sPTB and PPROM at Timepoint 1 (16 weeks) and Timepoint 2 (20 weeks) to determine if any gene was differentially expressed between these two subtypes of sPTL.

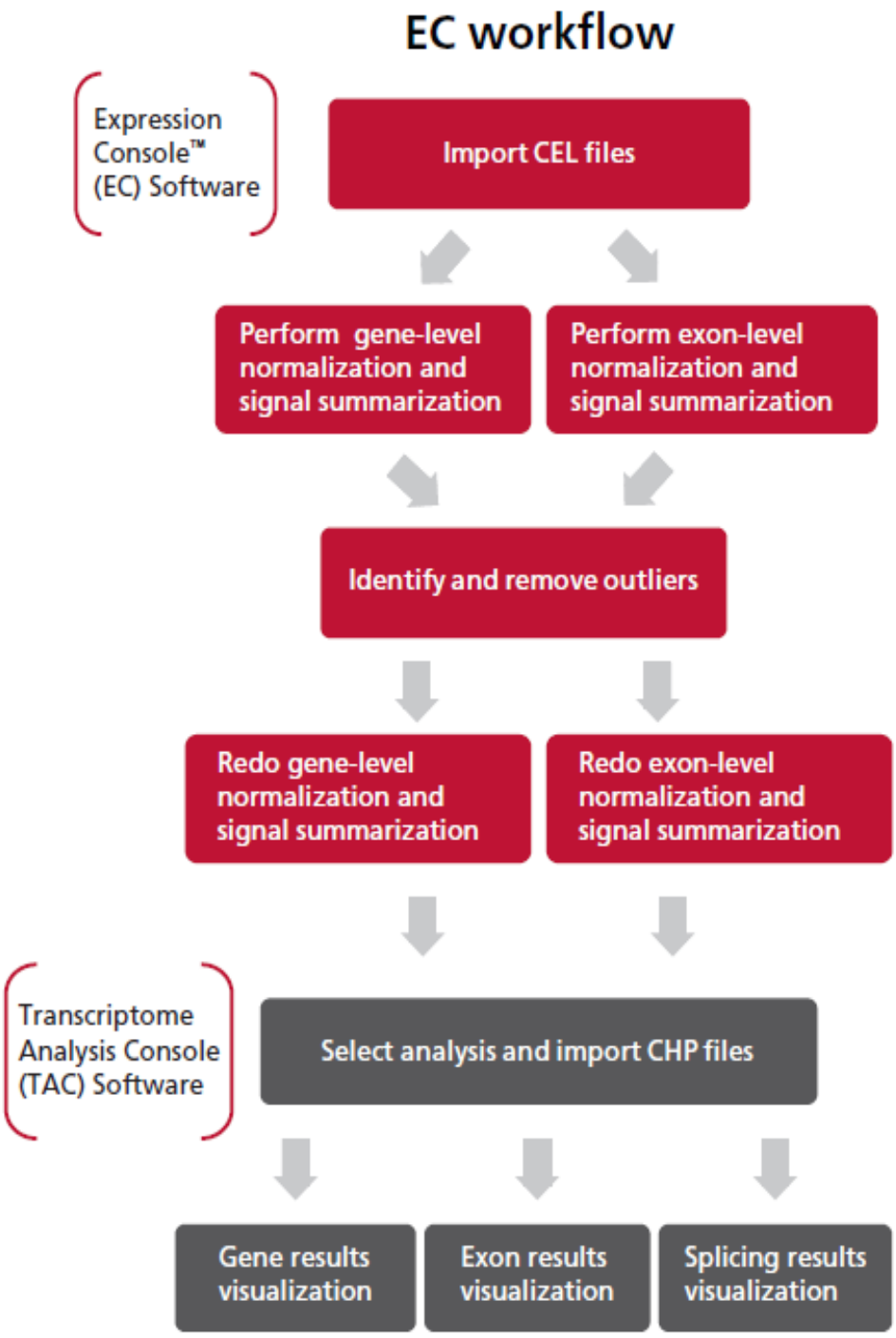


Figure 5.2. Expression Console Software Workflow. The steps in red were performed by the Centre for Genomic Research prior to exportation of the data files for further analysis.

Statistical Analysis

The machine learning approach used was random forest methodology (Breiman. 2001) performed with the R-package "rager". The basis of the machine learning method is the 'decision tree' which asks questions of the data and plots all the possible answers (Figure 5.3). The random forest combines many decision trees together into a single model. The target is what the model should predict, those women with sPTB, PPROM or term delivery. The features are the columns of data in the data matrix used as input data (i.e. transcript levels per individual). We expect there to be some relationship between the features and the target value and the model will 'learn' these relationships during training. To do this, the model splits the data into a training set and a test set. On the test set, the model will 'see' the preterm birth classifications of the target and formulate 'questions' of the features to create a predictive model. Then on the test set, the model does not know the target classifications and tries to classify them using just the data features. The answers can be compared to our classifications ('true' values) to judge how accurate the model performs. This is performed in an out-of-bag (OOB) fashion so model building and testing are not confounded.

The name "*random*" forest gets its name because each decision tree in the model considers a random subset of the data features when forming its 'questions'. The random collection of features is called a 'node' and these input co-ordinates are what the model subsequently splits on. Individually decision trees may have a wide variance in prediction of the outcome, but on average many thousands of decision trees will get closer to the correct answer. This leads to more overall robust predictions as it increases diversity in the model.

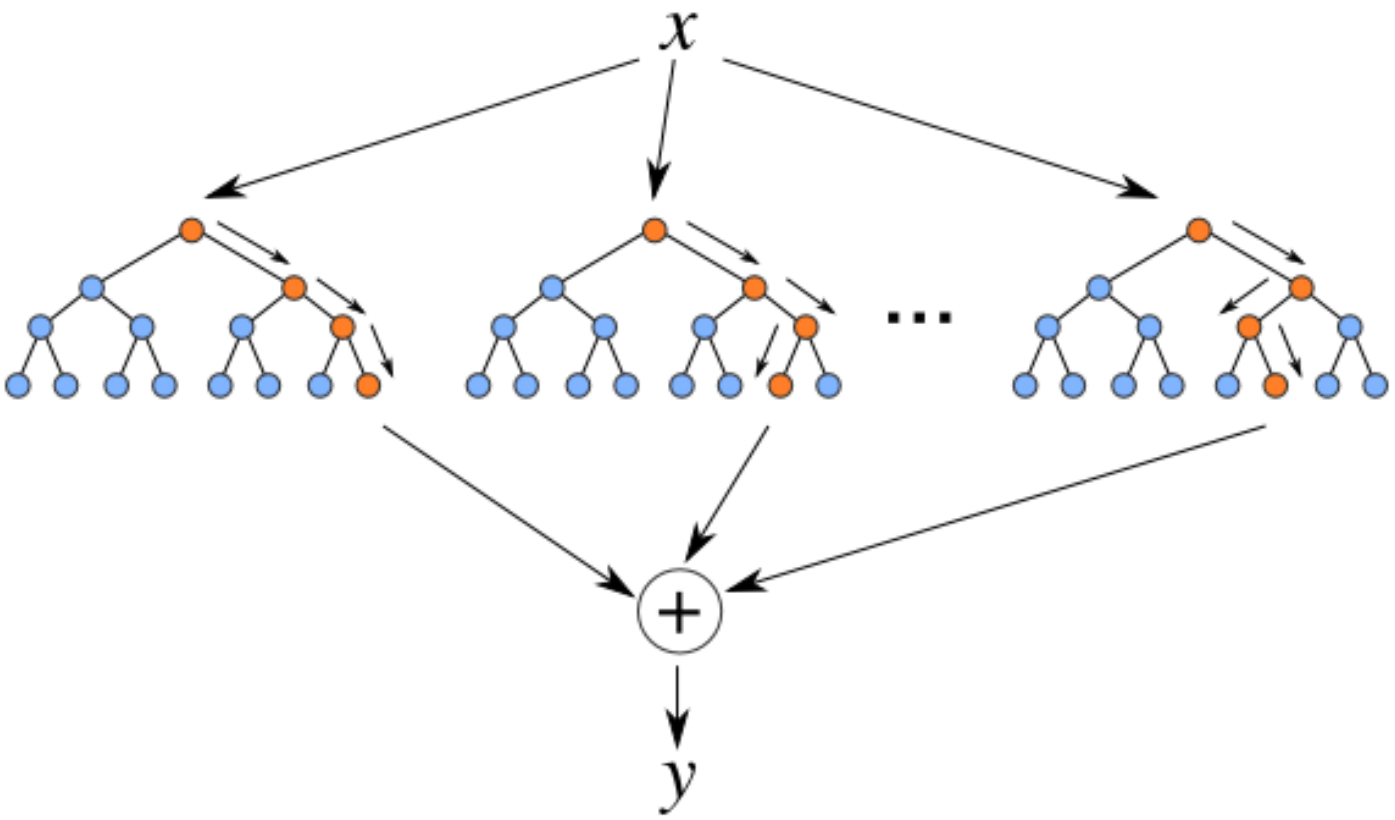


Figure 5.3. Random Forest Overview. Image obtained from Tiereney, B. 2018. *Random Forest Machine Learning in R, Python and SQL – Part 1*

Of note, for this analysis we did not use the random forest for prediction in its original usage, but rather focussed on the possibility of obtaining variable importance's in a multivariate setting. In this analysis, 10,000 trees were generated aiming to predict the phenotype of interest. Historically random forest models have been biased in such a way that categorical variables with a large number of categories are preferred (Altmann et al. 2010). Therefore, we used a normalising feature importance measure to correct feature importance bias called permutation importance score (Altmann et al. 2010). From this, the permutation importance was taken and the scale function in R was executed expressing the permuted importance in terms of a pearsonised variate. Variable importance values of five standard deviations above the mean variable importance was used for further Gene Set Enrichment Analysis (GSEA). In addition, we also performed a confirmatory analysis to verify the variable importance's identified using the alternative method of Janitza *et al.* (2016) and found very good agreement between the two approaches.

Gene Set Enrichment Analysis (GSEA)

Pre-ranked GSEA (Subramanian et al. 2005) was used to determine significantly enriched gene sets/pathways. To achieve our aim, Functional Mapping and Annotation (FUMA) tool (Watanabe et al. 2017) was used to identify the enriched genes and associated systems. As described above, transcripts exhibiting more than five standard deviations above the mean variable importance were used. GENE2FUNC (gene to function) option was used to explore the significance of identified pathways. Identified pathways were further explored using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database.

Hierarchical Clustering of Expression Profiles

To measure the non-linear statistical dependence between random variables a *Randomised Dependence Coefficient* (RDC) was used (Lopez-Paz et al. 2013) and extended (Jia et al. 2019, in preparation) to define the distance metric used in hierarchical clustering.

5.3 Results

From 58 women with whole blood RNA samples, 27 women were included in the final analysis. (Figure 5.4/Appendix K) RNA was available for analysis on 16 term controls (NORM), 9 sPTB <37 weeks (SPON), 2 PPROM < 37 weeks (PPROM). No predictive clinical variables were significantly associated with sPTB (Table 5.1).

Quality Control RIN values can be seen in *Appendix J*. Classical differential expression analysis did not reveal significantly different expression values for any of the transcripts measured, likely owing to small sample size.

Following random forest analysis, 178 transcripts had ENSEMBLE ID's suitable for further analysis and were subject to GSEA. The R code used for the random forest analysis is available to view in Appendix K and an example of one of the 10,000 trees is shown in Figure 5.5. Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) demonstrated significant enrichment of the selenoamino acid metabolism pathway in the high-risk preterm birth population (Figure 5.6). Three significant genes highlighted out of 26 genes from the selenium metabolism pathway (CTH, LCMT1, TRMT11) (Table 5.2). The p value is $3.84e^{-3}$ after adjusting for multiple testing across the entirety of KEGG pathways.

Clusters of women delivering at term (NORM), sPTB (SPON) and PPROM (PPROM) are represented in Figure 5.7. The hierarchical clustering indicates there are differences between these phenotypes across the selenoamino acid metabolism pathway.

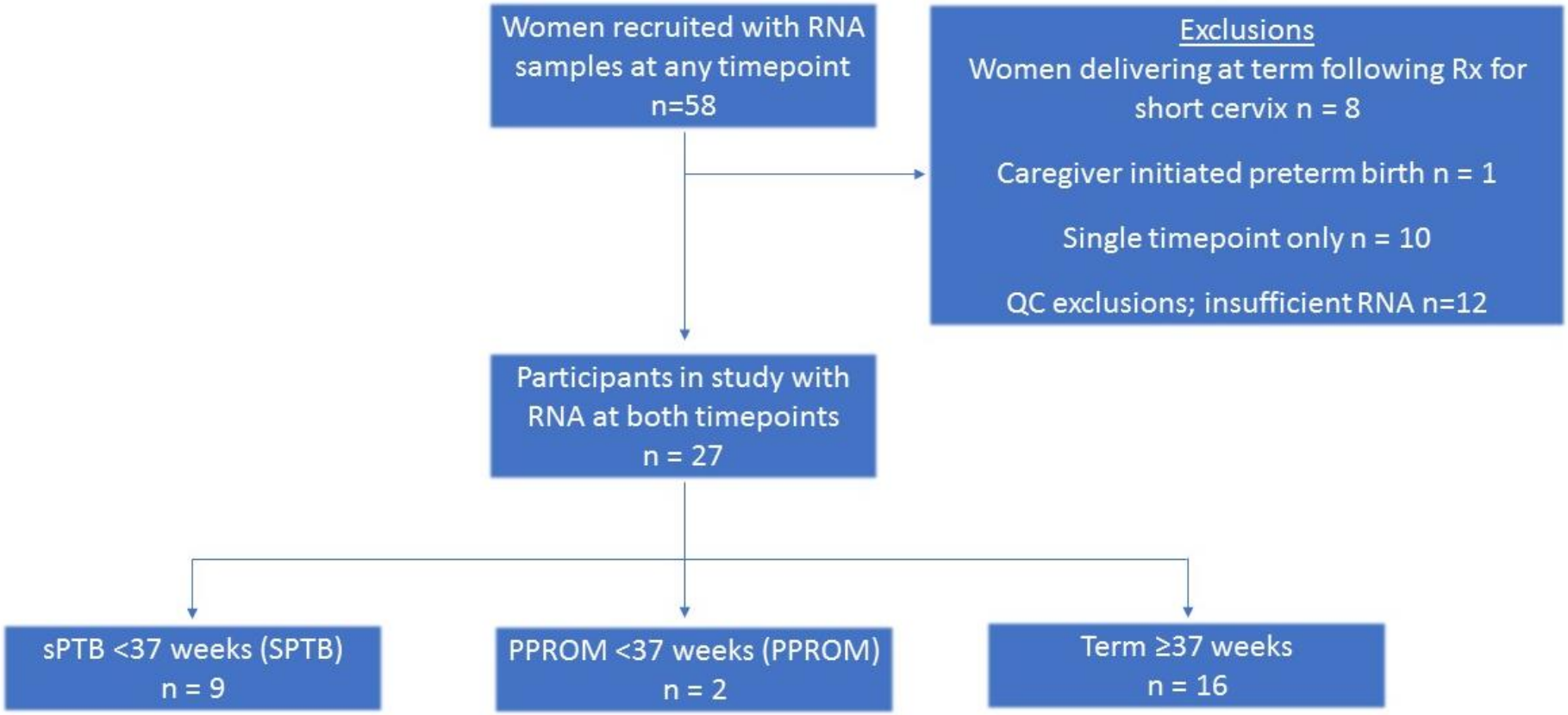


Figure 5.4. Flowchart demonstrating final number of analysed samples

Table 5.1. Demographics of participants in transcriptomic study

	sPTB N=9	Early PPROM N=2	Term Birth N=16	P value
Participant Demographics				
Maternal age, mean years +/- SD	31.3 (5.8)	32.5 (0.7)	31.3(5.6)	.959 [a]
Booking BMI, mean +/- SD	27.9 (6.7)	24.5 (1.8)	24.6 (3.0	.250[a]
Ethnicity				
Caucasian, n(%)	8 (89)	2 (100)	14 (87.5)	.492 [b]
Non-Caucasian, n (%)	1 (11)	0	2 (12.5)	
Smoking during pregnancy				
Yes, n(%)	2 (22)	0	5 (31)	.606[b]
No, n(%)	7 (88)	2 (0)	11 (69)	
Clinical Characteristics				
Gravidity, mean+/-SD	4.1 (1.6)	3 (0)	3.75 (2.1)	.507 [c]
Parity	1.4 (1.0)	1.5 (.71)	1.7 (2.3)	.771 [c]
History of previous PTB				
Previous sPTB, n (%)	5 (56)	1 (50)	8 (50)	.964 [b]
Previous PPROM, n (%)	4 (44)	1 (50)	8 (50)	
Both previous sPTB and PPROM (%)	0	0	0	N/A
Previous twin sPTB, n (%)	0	0	0	N/A
Cervical surgery (single LLETZ), n (%)	0	0	1 (6.3)	.700 [b]
Gestational age, visit 1, mean +-/- SD days	114 (4)	111.5 (.7)	115 (4)	.574 [a]
Cervical length visit 1, mean +/- SD	33.3 (5)	38 (4)	34.6 (5)	.545 [a]
Gestational age, visit 2	141 (4)	140 (.7)	142 (6)	.669[a]
Cervical length visit 2, mean +/- SD	30 (8)	23 (4)	36.6 (7)	**.**0.16 [a]
Other chronic medical conditions				
Yes, n (%)	6 (67)	2 (100)	5 (31)	.074 [b]
No, n (%)	3 (33)	0	11 (69)	
Polyhydramnios				
Present, n (%)	0	0	0	N/A
Absent, n (%)	9 (100)	2 (100)	16 (100)	
Labour and Delivery Characteristics				
Onset of labour				
Spontaneous	7 (78)	0	5 (31)	.037
Induction	0	1 (50)	9 (56)	
No labour	2 (22)	1 (50)	2 (13)	
Gestational age at delivery, mean +/- SD days	239 (18)	221.5 (25)	271 (8)	.000
Birthweight, mean grams +/- SD	2338 (567)	1847 (584)	3209 (522)	.000
Neonatal gender				
Female, n (%)	4 (44)	2 (100)	6 (38)	.245
Male, n (%)	5 (56)	0	10 (62)	

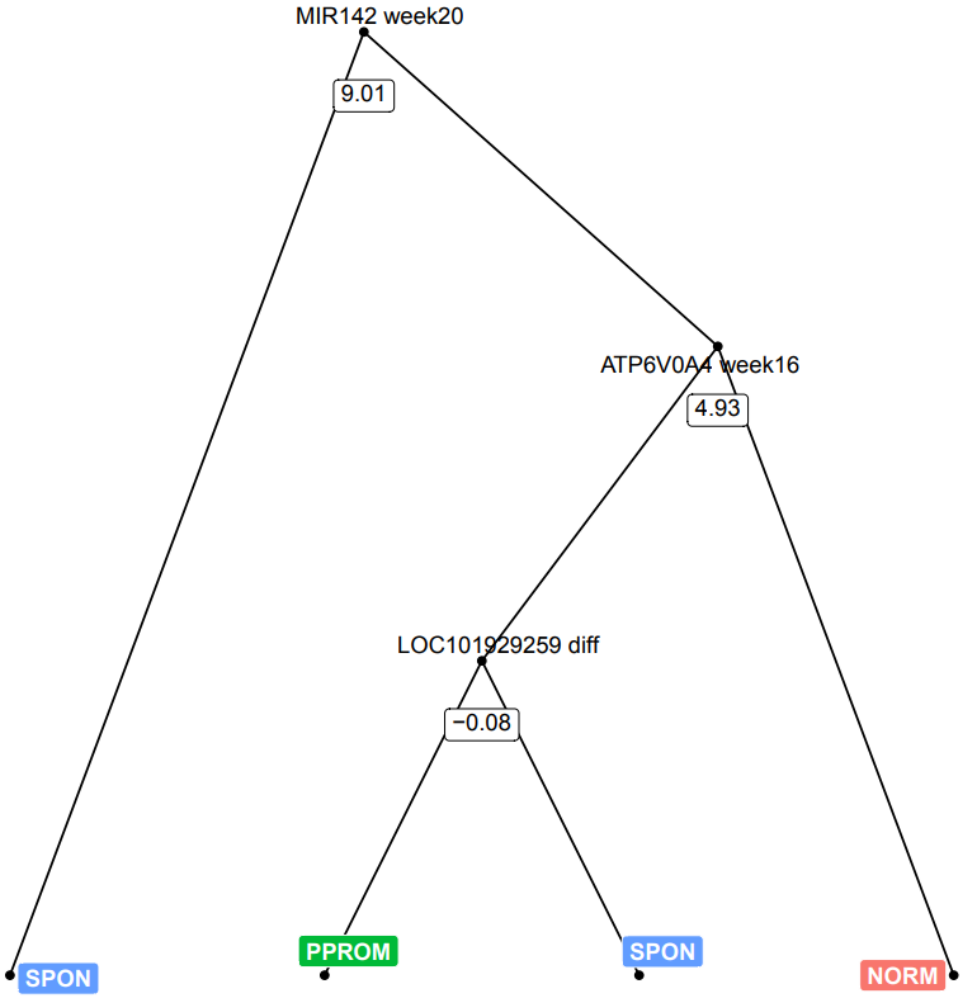


Figure 5.5. One of 10,000 trees from random forest analysis. Nodes are using data from week 16, week 20 and the difference between both weeks to differentiate the data into groups. Gene IDs are shown.

Table 5.2. Functional Mapping and Annotation (FUMA) summary of enriched genes. N = total no genes in the gene-set, n = number of genes enriched in the gene set.

GeneSet	Total genes (N)	Enriched genes (n	P-value	adjusted P	Genes
KEGG selenoamino acid metabolism	26	3	2.07e-5	3.84e-3	CTH, LCMT1, TRMT11

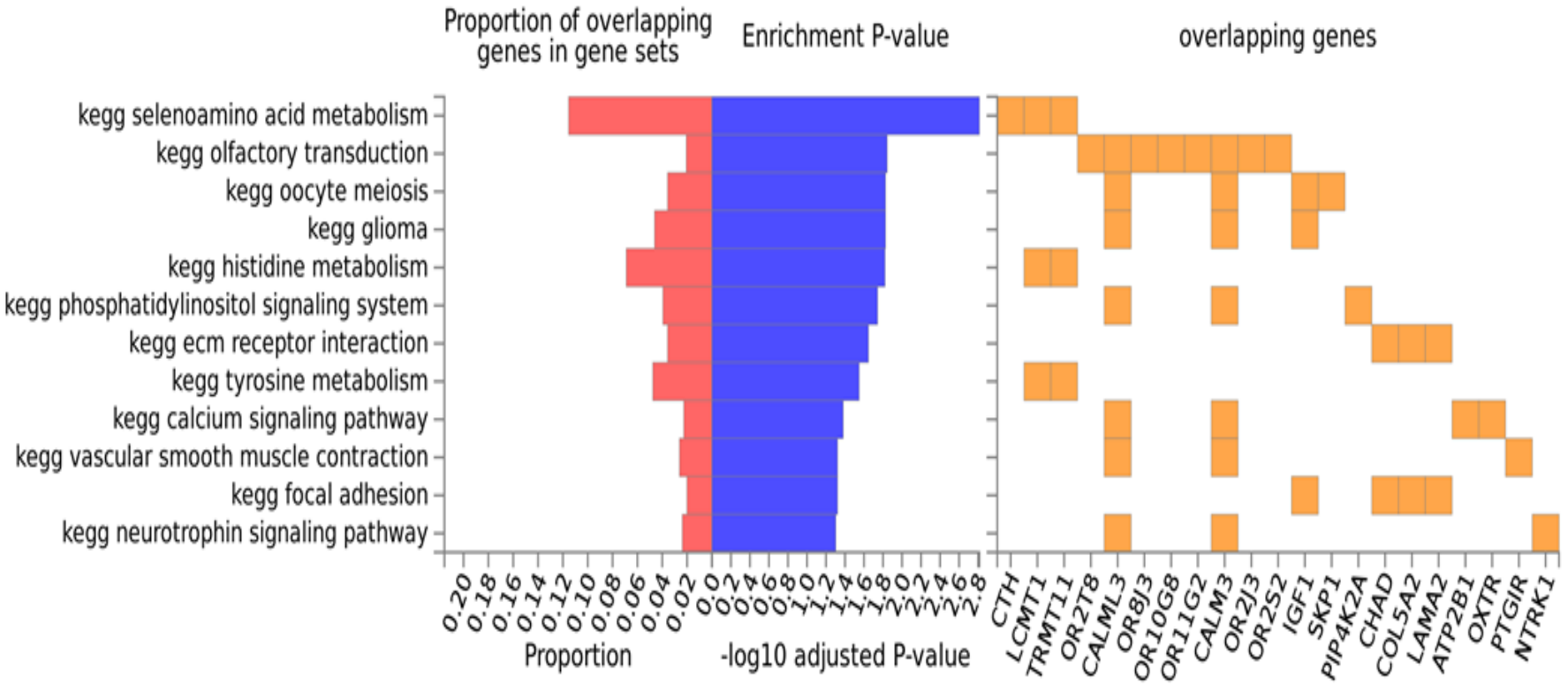


Figure 4. Significant enrichment of selenoamino acid metabolism pathway was identified in a high-risk preterm birth population. FUMA graphical representation of overlapping genes in the identified pathways

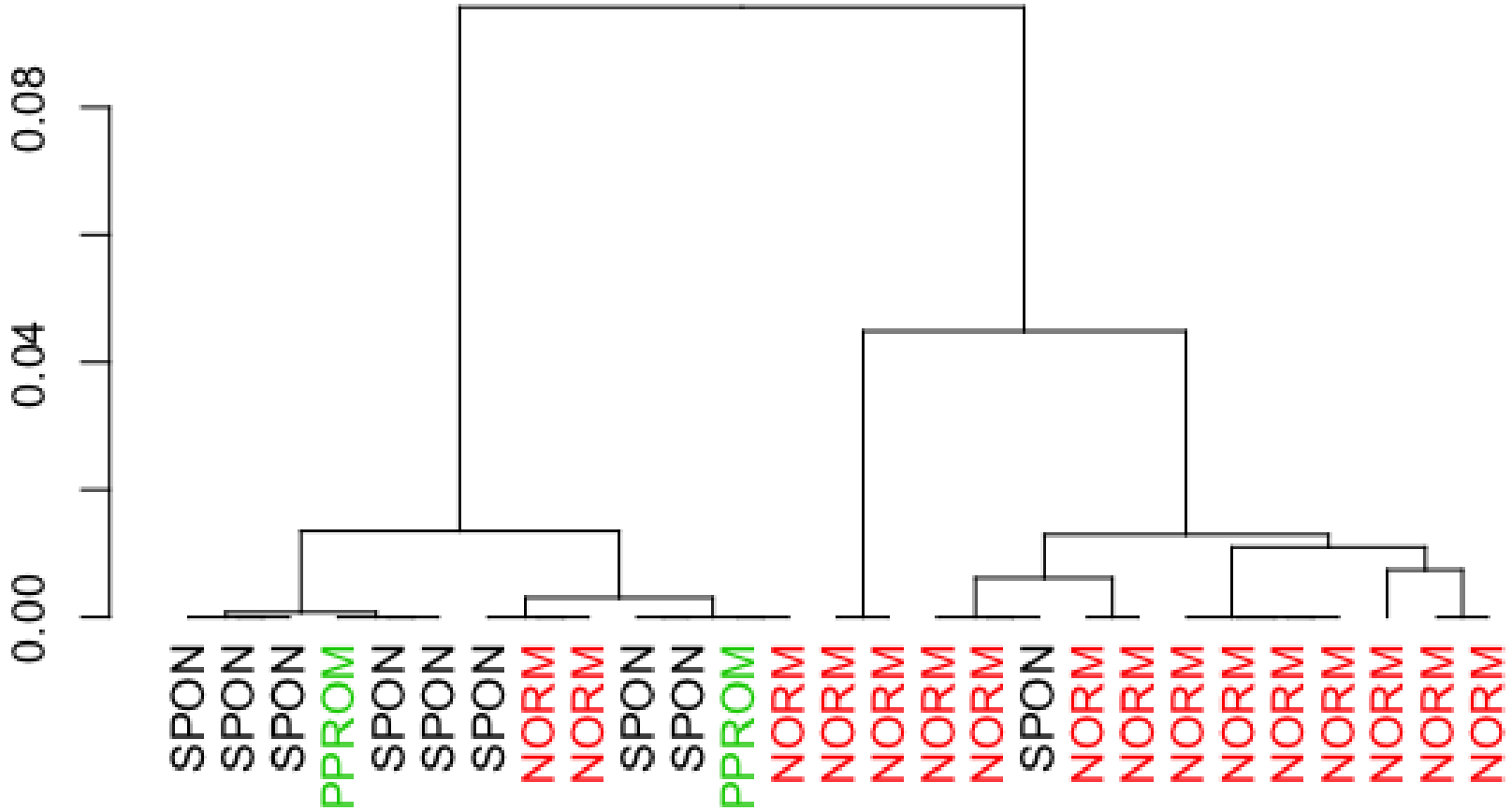


Figure 5.7. Hierarchical clustering of patients' expression profiles derived from the selenium pathway. The distance matrix was defined as 1-RDC, RDC denoting the randomized dependence coefficient developed by Lopet-Paz et al, 2013 (https://arxiv.org/abs/1304.7717) and extended by Jia & Müller-Myhsok (in preparation).

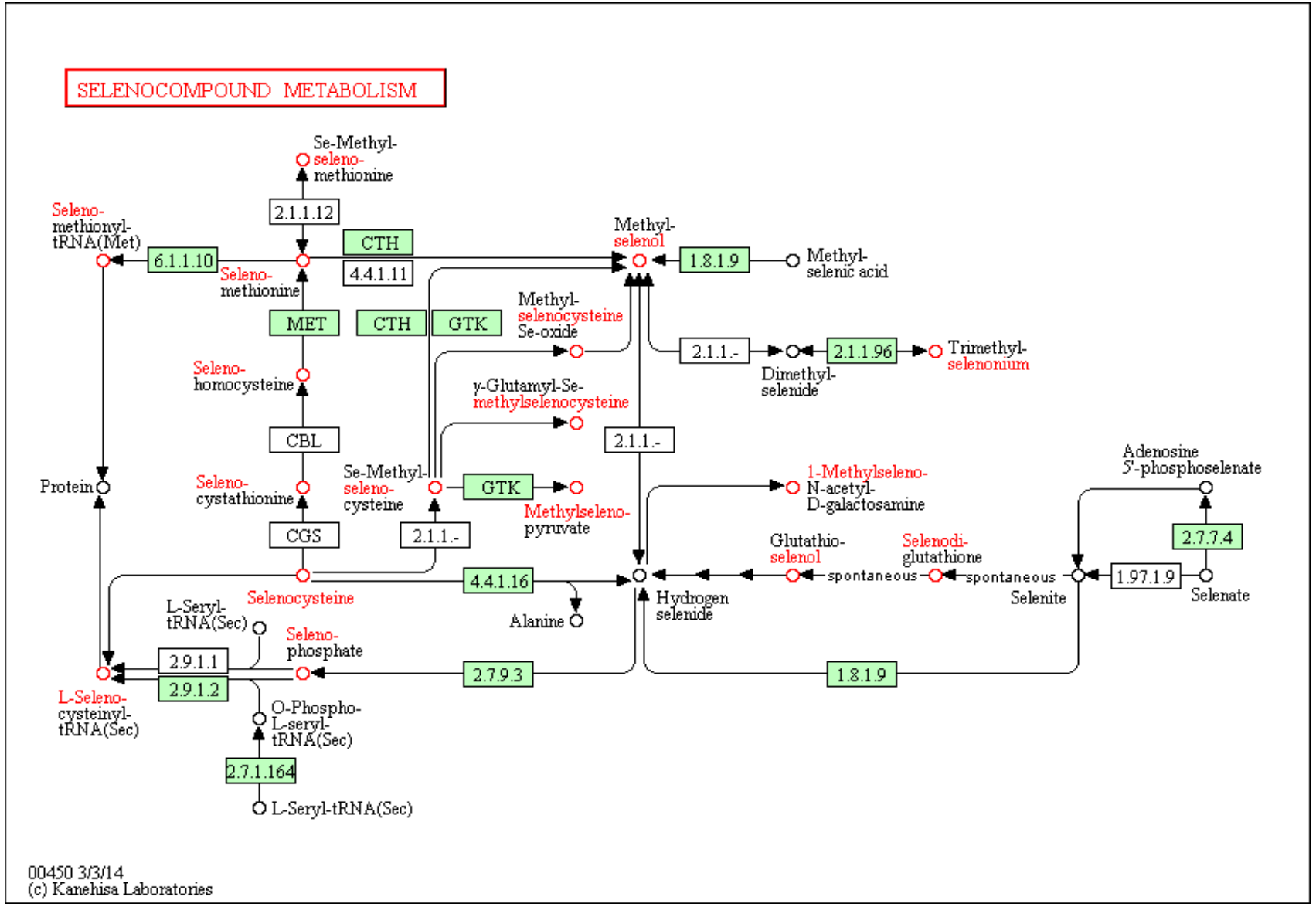


Figure 5.8. Selenium metabolism pathway from KEGG pathway database (Kanehisa et al. 2017) Compounds involved in selenium metabolism

5.4 Discussion

In this part of the thesis, we investigated the association of whole blood gene expression across two clinically relevant timepoints between women who had recurrent sPTB, PPROM and women who had term births following a history of a sPTB. The findings of our GSEA suggest a role or an association of selenium in the initiation of early labour.

After adjusting for multiple testing, three genes in the selenium pathway were found to be statistically significant: CTH, LCMT1, TRMT11.

- CTH gene encodes a cytoplasmic enzyme in the trans-sulfuration pathway that converts cystathionine derived from methionine into cysteine (Figure 5.8).
- LCMT1 catalyses the methylation of the carboxyl group of the C-terminal leucine residue (leu309) of the catalytic subunit of protein phosphatase-2A (De Baere et al. 1999).
- TRMT11 is a protein coding gene for tRNA methyltransferase 11 homolog that transfers a methyl group onto a single guanidine residue present in most tRNAs and thereby modifies them post transcriptionally (Hori H. 2014).

It is possible that the effects of a dysfunctional selenoprotein may change this metabolite pathway in gestational or maternal tissues affecting gene expression and may also associate with low maternal selenium status. Selenium is known to be involved in attenuating inflammation and low maternal levels have been reported in the literature as associated with preterm birth (Rayman et al. 2011).

Our pilot analysis was based solely on random forest profiling which are fast to perform, easy to implement, produce highly accurate predictions and can handle a

large number of input variables without overfitting (Biau. 2012, Ahmad et al. 2018.). Random forests are considered to be one of the most accurate general-purpose learning techniques available (Biau. 2012). It generates data trees at random and is therefore completely unbiased from prior knowledge of the possibility of selenium involvement in PTB pathways. Strengthening our interest in the Se hypothesis is:

1) the recent discovery of association of the EEFSEC gene discussed in chapter 4,
2) Dutch PTB cohorts demonstrating women in the lowest quartile of serum selenium having twice the risk of PTB as the women in the highest quartile (OR 2.0, 95% CI 1.19-3.47) (Rayman et al. 2011).
3) low serum selenium concentration being independently related to PTB (OR 2.18, 95% CI1.25-3.77). (Rayman et al. 2011).

Interestingly, there are comparably low serum concentrations of selenium that have been found in a UK obstetric population compared to the Dutch population (Rayman et al. 2003). This important result from our pilot work requires validation before more credence can be given to this theory, but there are potential translational implications for patients.

With regards to the science and plausibility of this theory, the trace mineral selenium plays a role in immune response and the body's resistance to infection. Enzymes containing selenoenzymes can attenuate the inflammatory response associated with sPTB by downregulating the expression of pro-inflammatory genes (Vunta et al. 2007). The amino acid residue selenocysteine (Sec) is a major form of Se in the cell and Sec is encoded by the UGA codon. Proteins containing Sec are thought to be largely responsible for the health benefits of Se. Sec is introduced into selenoproteins by a complex mechanism that requires trans-acting protein factors,

Sec-tRNA (Shetty et al. 2014). When UGA codon is encountered by a ribosome this normally signals as a “stop” codon and translation is terminated, however Sec machinery interacts with translational machinery to prevent premature termination. At least two trans-acting factors are required for efficient recoding of UGA as Sec in eukaryotes; SBP2 and EEFSEC. It is not clear if women with higher levels of this Sec-specific translation elongation factor are more genetically capable of translating selenoproteins that cascade to prevent inflammation or if there are the creation of selenoproteins that are causing sPTB birth or PPROM.

Our hierarchical clustering analysis suggests that this high-risk group could be differentiated on their expression profiles of the selenium pathway by machine learning (Figure 5.7).

Further studies are required to validate these findings, but our data are strengthened by our GWAS data from the same cohort in chapter 4 reflecting EEFSEC gene expression from pooled data analysis in this cohort.

5.5 Conclusion

This analysis used whole blood gene expression across two timepoints to differentiate sPTB and PPROM from women delivering at term in asymptomatic women. Using expression levels and random forest alone, no predictors were found. However, a gene set enrichment in this population demonstrates that the selenoamino acid pathway differentiates asymptomatic high-risk women. Hierarchical clustering in a non-linear distance matrix can differentiate all but one of the sPTB cases. More studies are required to validate the findings from our analysis.

Chapter 6: Metabolomic Profiling of Pregnant Women to Assess for Candidate Metabolites Useful for the Clinical Prediction of Spontaneous Preterm Birth

6.1 Introduction

The advantage of metabolomics for biomarker discovery is that this "omics" layer is the most downstream from gene expression and protein synthesis and may be more representative of physiology at a functional level (Romero et al. 2010). Pregnancy is a state of adaptation for the human female with many processes changing over the course of gestation (Lain KY. 2007). Glucose, protein, calcium and lipid metabolism change to accommodate the growing fetus, with additional adaptation in maternal respiratory, endocrine, renal and cardiac physiology. Maternal blood as a candidate biological fluid should detect these changing biochemical dynamics and remains easily accessible for study or screening. Blood also remains in constant exchange with the fetus through the placenta providing nutrients required for growth and development.

As discussed in Chapter 2, two main approaches to the generation of metabolomics data are nuclear magnetic resonance (NMR) and mass spectrometry (MS). I chose to use NMR for analysis as it is a fast and highly reproducible technique that requires very little sample preparation or manipulation and typically identifies around 50 metabolites in serum. It is based on energy absorption and re-emission from the atom nuclei due to variations in the external magnetic field (Bothwell and Griffin. 2011). Hydrogen is the most commonly targeted nucleus (^{1}H-NMR) due to its natural abundance in biological samples. The resulting spectral data allows for indirect quantification of the concentration of the metabolite but also provides information about the chemical structure (Alonso et al. 2015). The pattern of the spectral peaks informs the physical properties (chemical structure, oxidative state, phosphorylation etc.) of the metabolite and is used in metabolite identification, whilst spectral peak areas are an indirect measure of quantity of metabolite in the

sample. Aside from the high reproducibility and short acquisition times, other advantages of ^{1}H-NMR include requirement of small sample volumes, low cost of analysis, non-destructiveness of the sample which remains intact after analysis. As a spectroscopic technique (rather than spectrometric) metabolite profiles obtained are virtually independent of the operator and instrument. Its main disadvantages include a high instrument cost and relatively low sensitivity to molecules present in low volumes (Kamath-Rayne et al. 2013). This chapter will discuss using NMR metabolomic profiling to assess candidate metabolites in the prediction of sPTB and PPROM.

6.2 Aims

1) To compare the metabolite profiles at 16 and 20 weeks of gestation in all participants (sPTB, PPROM and TERM groups)
2) To compare temporal metabolite profile changes between 16 and 20 weeks (sPTB, PPROM and TERM groups).
3) To assess the differences in metabolite profiles between women experiencing sPTB and PPROM at 16 and 20 weeks.

6.3 Methods

Population

From the population described in chapter three, of 128 women who had given at least one serum sample for analysis 46 were excluded from participation. To summarise, women were excluded if they had a 1) caregiver initiated preterm birth (i.e. not spontaneous), 2) sPTB that had a likely identifiable cause (i.e. infection or placental abruption), 3) women with PPROM that were associated with polyhydramnios or chorioamnionitis, 4) genetic abnormality associated with spontaneous preterm birth or miscarriage, 5) late spontaneous preterm birth (34+1-36+6) and 6) women delivering >37 weeks who had treatment for short cervix.

Materials

500 µl aliquots of serum were securely stored at -80°C in the NMR Centre for Structural Biology laboratory after transfer from the Centre for Women and Children's Health, prior to processing in batches. On the day of analysis, samples were thawed for 1 hour and 330µl aliquots of serum and 330 µl of phosphate buffer; consisting of 66µl of 1M Sodium Phosphate ($Na^{+}PO_4^{3-}$) buffer prepared in 99.8% deuterated water (^2H_2O or D_2O); pH 7.4, and 264 µl of double distilled water (ddH_20) prepared in Eppendorf tubes following established protocols (Beckonert et al. 2007) (Figure 6.1). The Eppendorf tubes were centrifuged at 21,500g for 5 minutes at 4°C and 600µl of the supernatant was transferred into 5-mm NMR tube.

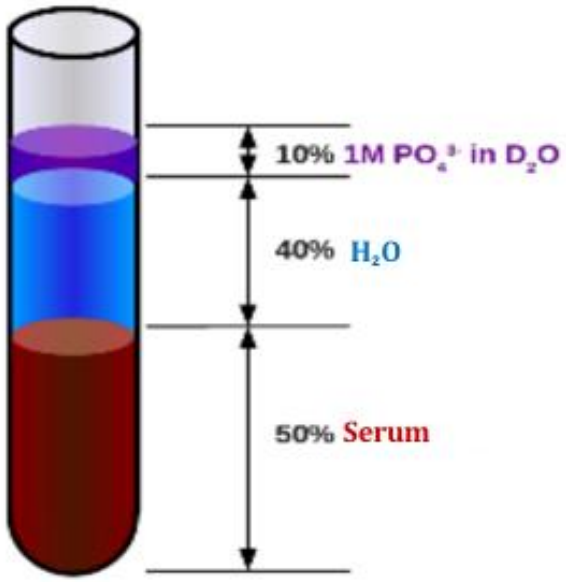


Figure 6.1 Ratio of Serum, Water and Phosphate Buffer

Spectral Acquisition

Spectra were acquired using Bruker Avance spectrometer operating at the proton frequency of 600MHz and equipped with a triple resonance TCI cryoprobe (Bruker, GmBH, Germany). A one-dimensional Carr-Purcell-Meiboom-Gill (CPMG) ^{1}H NMR echo pulse sequence with water suppression was employed to filter out broad spectral resonances arising from the macromolecules (such as lipoproteins and albumins). Spectra were manually phased and the baseline corrected using the Topspin 3.1 software (Bruker, GmBH, Germany). The whole spectrum was referenced and aligned to the glucose anomeric hydrogen signal (δ= 5.23 ppm). The residual water region (4.40-5.00ppm) was selectively removed.

Binning NMR spectra

NMR outputs are extremely complex spectra of metabolite resonances that are produced without any prior separation of the sample. Therefore, overlap of chemical signals makes it impossible to identify all the components in a spectrum. To analyse this type of data, reduction techniques are necessary. The practice of

splitting up spectra into integral regions is called “binning” or “bucketing”. Each sample spectra were subsequently divided into bins, assigning peak boundaries as they appeared on the spectra. Background noise had previously been removed using the QC steps discussed above.

Spectral Annotation and Metabolite Identification

Characteristic metabolite peaks have previously been identified and can now be used to annotate spectra via pattern recognition (Soininen et al. 2009). The schematic in Figure 6.2 shows A) the spectra produced from this analysis and B) an animation showing the characteristic identification peaks. Most spectra overlap, however some spectra showed obvious differences. Using Chenomx™ metabolomics software, compounds were annotated by comparison to databases of independently verified and externally validated metabolites. If a bin could not be annotated it was labelled as ‘unknown’ and each unknown bin was given a unique identifying number.

Statistical Analysis

SPSS v.24 (IBM Corporation, USA) was used for statistical analysis of clinical demographic data of the participants involved in this analysis. Descriptive statistics for the cohort included median and range values for age, BMI, smoking status, gestational age and cervical lengths at sampling.

The metabolomic data were analyzed using Metaboanalyst 4.0 Statistical Analysis (Xia and Wishart. 2016), an online tool for metabolomics analysis and interpretation. Following data upload, median intensity values were used for data filtering.

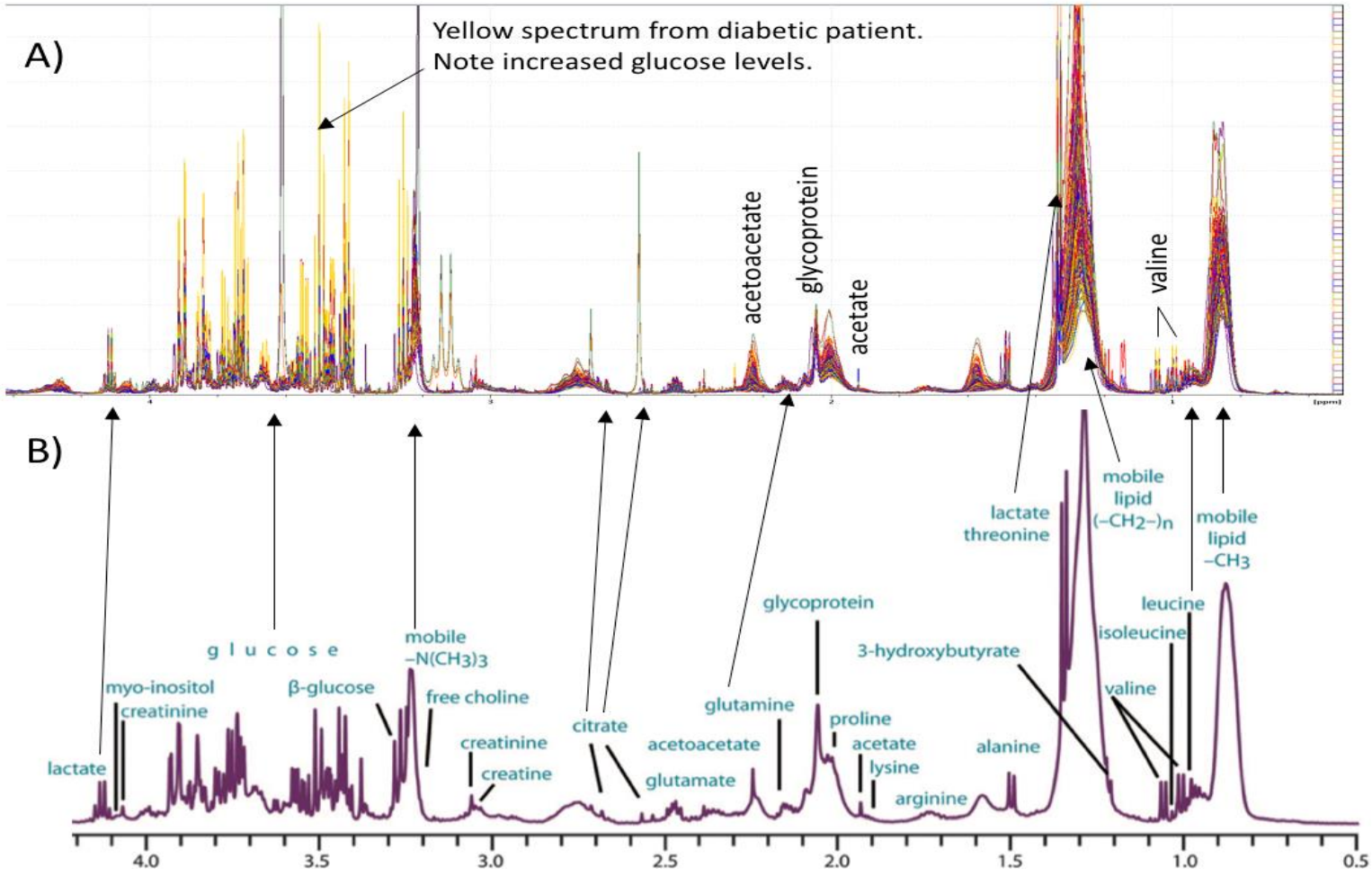


Figure 6.2. Annotation of Spectra. A) The aliphatic region (<5ppm) of all 128 patient samples. B) Animation showing recognised peaks (adapted from Soininen et al. 2009)

Data filtering identifies and removes variables unlikely to be of use during data modelling. No phenotype information is used in the filtering process, so the result can be used with any downstream analysis. This step is strongly recommended with untargeted datasets such as spectral binning, as the large number of variables provide baseline noise. (Hackstadt and Hess. 2009)

Preprocessing of the data matrix plays a crucial role in ensuring that subsequent data analysis is more robust and accurate. Two key steps in preprocessing are i) normalization and ii) scaling of the data. Probabilistic quotient normalization (PQN) was used to normalize our spectra (Dieterle et al. 2006, Kohl et al. 2012). The data were scaled using auto-scaling (mean-centred and divided by the standard deviation of each variable) (Jackson. 2006, Van den Berg et al. 2006). Together these normalization and scaling methods gave the most Gaussian distribution of the dataset prior to univariate statistical analysis.

Outlier Identification

An overview of the data was initially performed with a principal component analysis (PCA) to identify any outliers prior to further analysis. If an outlier was identified, then the participant clinical data was considered as a whole to explain why this was an outlier (i.e. medications taken within 48 hours of sampling). If no explanation was found for the outlier, the point was removed from the dataset in case of errors being introduced during sampling or the NMR process.

Univariate Analysis

Univariate analyses were used to identify difference in metabolite bins between groups. When all three groups (sPTB, PPROM, Term) were compared, statistically significant differences were evaluated using the one-way analysis of variance (ANOVA) test between all metabolites. A p-value of <0.05 was considered

significant. To address the problem of multiple testing, adjusted p values were determined using the False Discovery Rate (FDR) approach. Post hoc analyses using Tukey's Method was used to explore any statistically significant differences between multiple group means while controlling the experiment-wise error rate to assess between which groups were the biggest differences When comparing two groups (SPTB and PPROM) fold change (FC) analysis, t-test and volcano plots were used. The purpose of fold change was to compare absolute value change between two group means per metabolite. The result is plotted in log2 scale, so the same +/- fold change was plotted the same distance from the zero baseline. T-test was used to determine if there was a significant difference between the means of the two groups.

Multivariate Analysis

PCA was conducted for detection of inherent trends and separation of group data. PCA is a powerful method of data extraction, which finds combination of variables that describe trends in large data, called principle components visualized in scores and loading plots.

To attempt to get sharper separation between the two groups a Partial Least Squares- Discriminant Analysis (PLS-DA) was performed. This is a supervised multivariate analysis that attempts to maximise the variances that separate the groups and minimise the differences within the patient groups to build a discriminant model. The model required cross-validation to ensure that the data was not overfit. A leave-one-out cross-validation was used to establish a sum of squares value (R2) and a predicted sum of squares value (Q2) to describe the sample clustering and the accuracy of prediction to which cluster each sample belongs. (Xia and Wishart. 2011).

6.4 Results

Demographics

From 128 women recruited to the biomarker pilot study, 46 participants were excluded prior to analysis. (Figure 6.3) Following exclusions three groups remained:

1) Women with sPTB $\leq 33^{+6}$ without evidence of infection; n=12
   (16wk n= 11, 20 wk. n=11, paired samples n=9)
2) Women with PPROM $\leq 33^{+6}$ without known cause for PPROM; n=10
   (16 wk. n=10, 20 wk. n=9, paired samples n=9)
3) Term controls; TERM n = 60
   (16wk n=55, 20wk n=53, paired sample n=50)

Metabolites Identified

Background noise was removed using the QC steps outlined. A data matrix of 145 bins per sample, with 101 bins assigned to 35 metabolites was produced. Appendix M details the full list of metabolites identified.

Removing Outliers

Once phenotype exclusions had occurred (Figure 6.3), the initial PCA for 16 weeks (Figure 6.4a) and 20 weeks (Figure 6.4b) demonstrated outlying samples in the term control group.

Both samples for participant 129 were flagged as outliers at both 16 and 20 weeks. This makes a measuring error during NMR processing less likely. Her clinical data was examined for potential causes. Apart from reporting eczema, smoking 6-10 cigarettes and taking pregnancy vitamins, no unusual medications or drug use was reported. Specific questions regarding diet were not explored beyond type of diet (vegetarian or non-vegetarian).

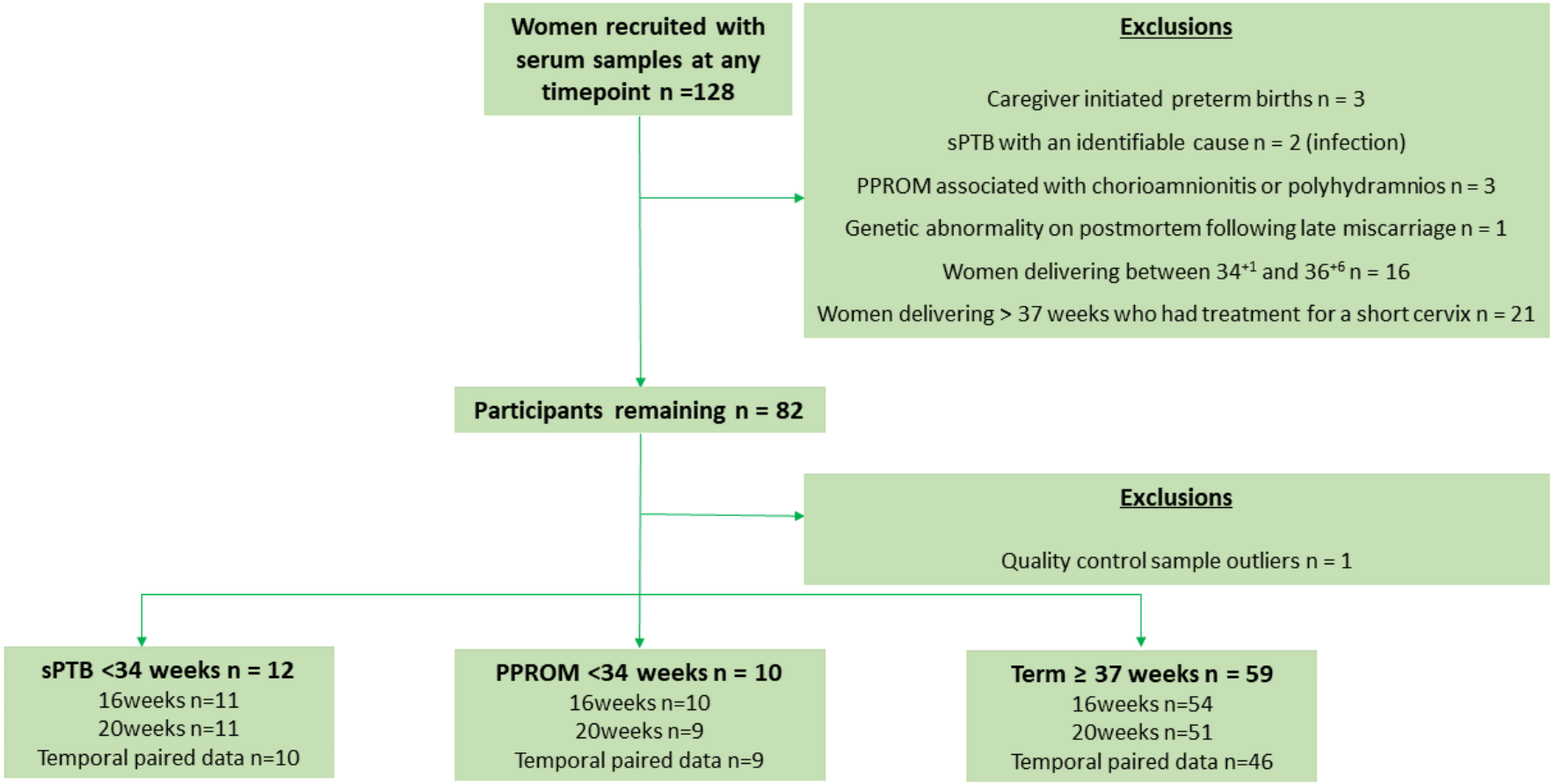


Figure 6.3. Flowchart to show final numbers of participants in metabolomic analysis

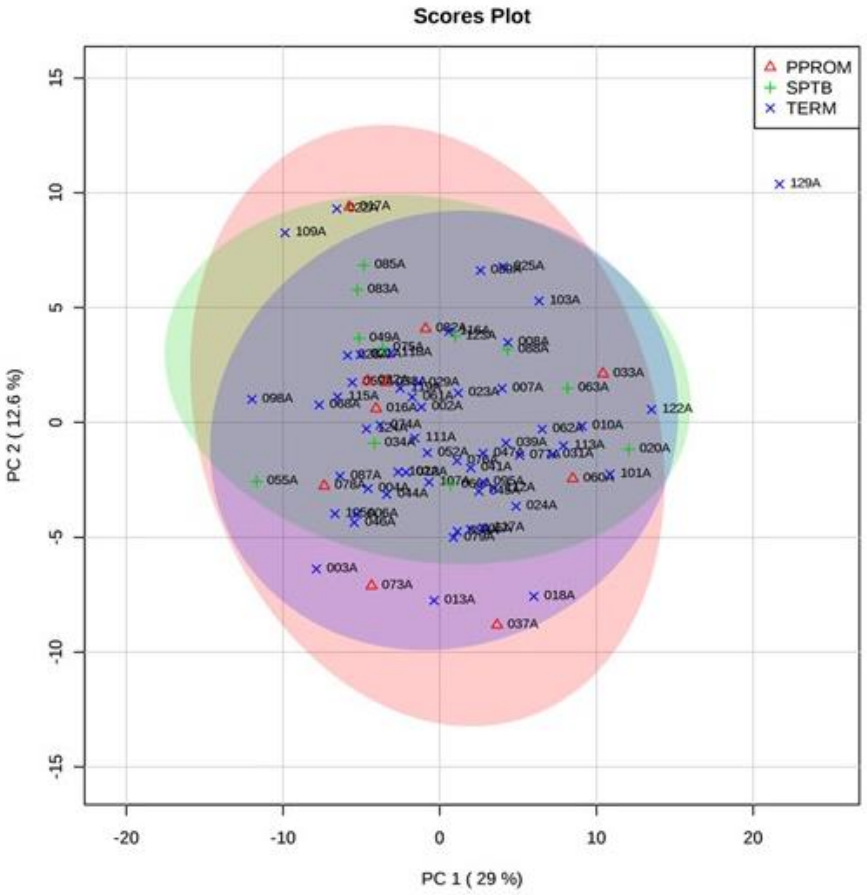


Figure 6.4a) PCA for all metabolomic samples in analysis at 16 weeks

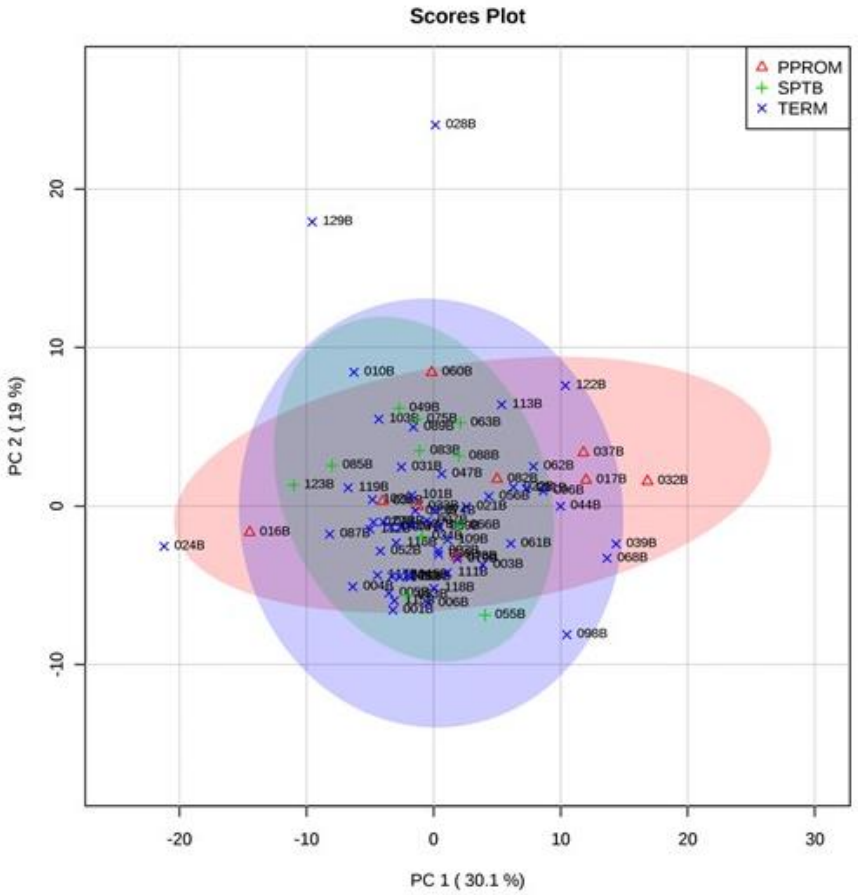


Figure 6.4b) PCA for all metabolomic samples at 20 weeks. Outliers shown away from the main clusters formed by the other samples and removed

Therefore, a specific diet or undisclosed drug consumption may have affected this profile. As no explanation was found, this patient was excluded. The sample for control participant no. 28 was also noted to be an outlier. In an otherwise fit and healthy individual, it was considered that this could be a sampling or measuring error and was removed.

Participant Demographics

Table 6.1 shows the clinical characteristics of these study participants. There are no significant differences between groups based on age, BMI, smoking or the gestation the samples were taken. The cervical length measured at 16 weeks does not show a statistical difference, but at 20 weeks it does fitting with this as a known predictor of sPTB.

Table 6.1 Characteristics of Study Participants

Data	sPTB (n=12)	PPROM (n=10)	TERM (n=59)	p-val
Age, yr.*	31 (20-40)	31 (21-36)	31 (19-40)	.557
BMI, *	28 (20-39)	25 (23-33)	25 (18-35)	.312
Smoking, (%)	1 (8)	3 (30)	16 (27)	.322
n, samples A visit	11	10	54	NA
GA at sampling, 16wk visit*	16^{+2} (14^{+5}-17^{+1})	16^{+0} (15^{+3}–17^{+3})	16^{+2} (14^{+1}–18^{+1})	.370
CL at sampling, 16wk visit*	33 (20-42)	39 (19-52)	36 (25-60)	.277
n, samples 20wk visit	11	9	51	NA
GA at sampling, 20wk visit*	20^{+1} (17^{+5}–21^{+6})	20^{+3} (19^{+2}–21^{+3})	20^{+1} (18^{+1}–23^{+1})	.693
CL at sampling, 20wk visit*	25 (5-37)	28 (0-37)	35 (23-56)	*<.001*
n, paired samples	10	9	46	NA
GA at delivery, wk.*	32+2 (22^{+4}–33^{+5})	33+0 (17^{+2}–34^{+2})	39+5 (37^{+0}–41^{+5})	*<.001*

*median(range). sPTB Spontaneous labour <34 weeks, PPROM PPROM <34 weeks, TERM women with term delivery. NA Not Applicable. GA Gestational Age. CL Cervical Length. Statistical p values calculated by Kruskal-Wallis

Comparison of all participants at 16-week gestation (TERM n=54, PPROM n=10, sPTB n=11).

The uploaded data matrix file contained 75 samples by 145 spectra bins for analysis, with zero missing values. Normalisation and scaling were performed before analysis (Figure 6.5).

The one-way ANOVA is shown in Figure 6.6. The metabolites meeting the p value threshold of 0.05 are labelled in blue. To show the differences in normalised metabolite concentrations between the groups, Figure 6.7 presents the boxplots of the assigned metabolites per group (PPROM – red, sPTB – green, TERM – blue). The median values of the PPROM group and TERM group appear consistently more similar in this analysis compared to the sPTB group coloured in green. However, none of these metabolites were statistically significant after adjusting for multiple testing (FDR < 0.05), therefore no post hoc analysis was performed.

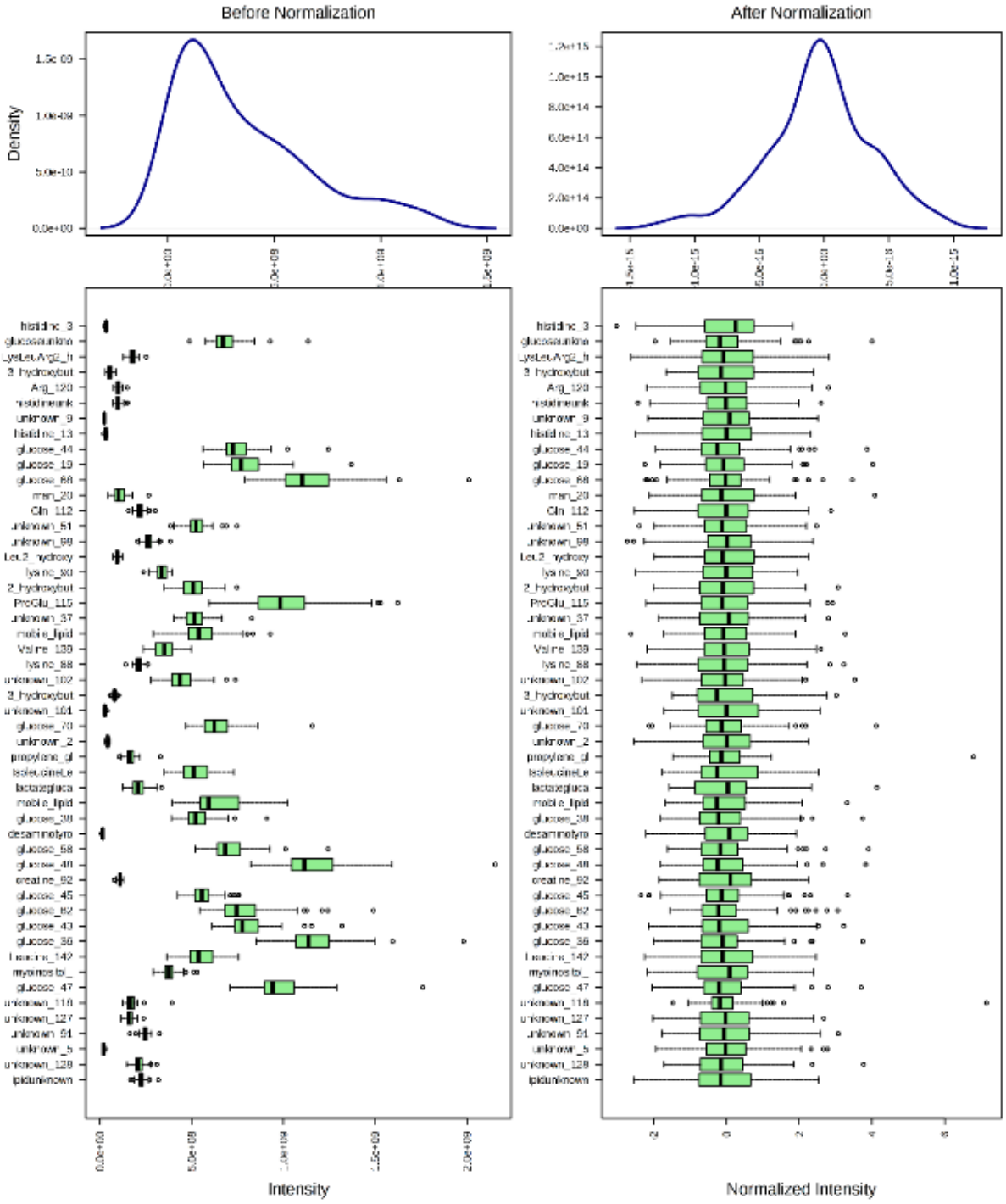


Figure 6.5. Box plots and kernel density plots before and after normalisation prior to data analysis at 16 weeks. Normalisation: Probabilistic quotient normalization (PQN). Data scaling; autoscaling

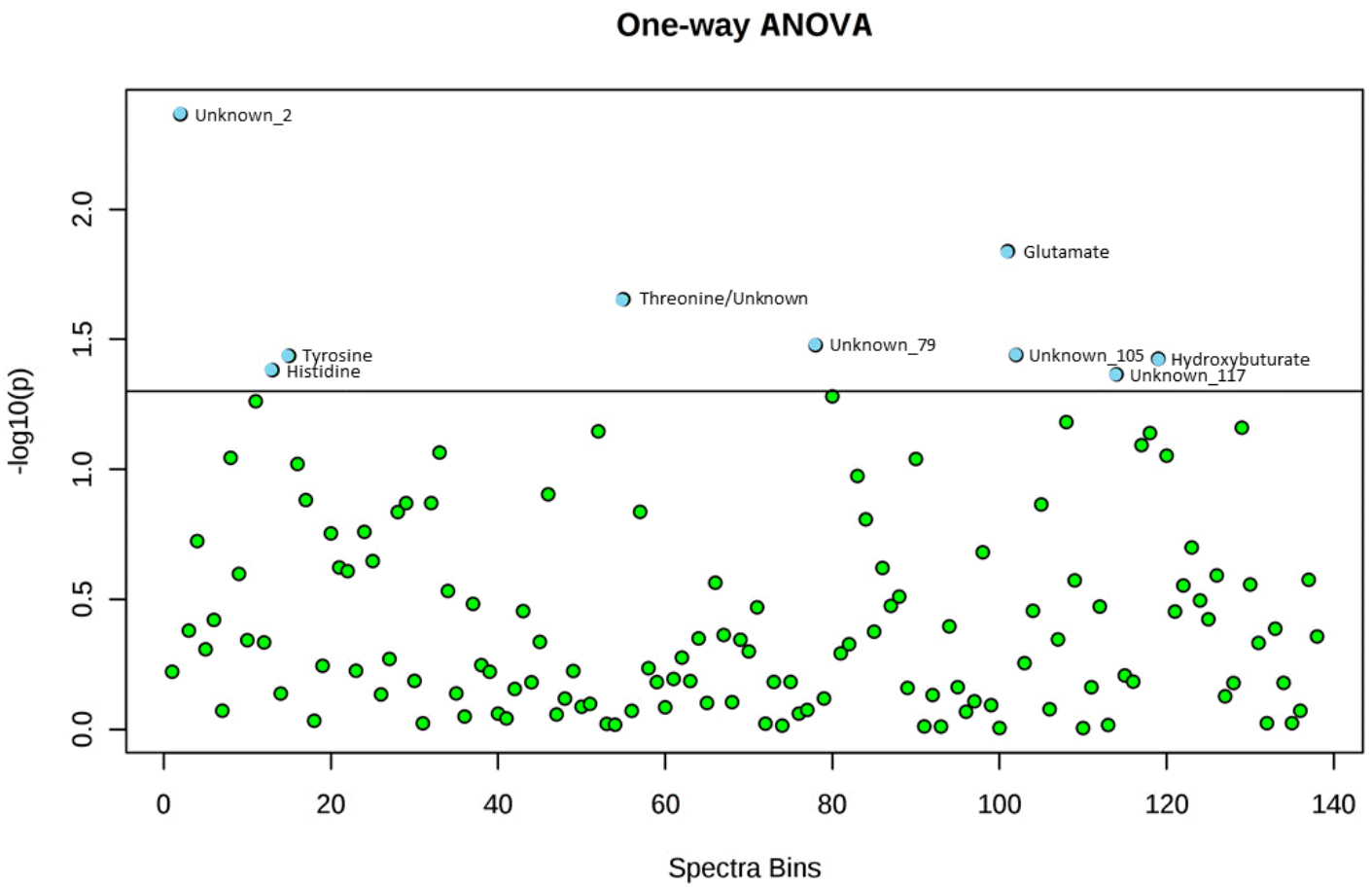


Figure 6.6. One-way ANOVA plot comparing samples taken at 16 weeks. The nominal (p val) 0.05 error rate is drawn across the graph as a black line. Metabolites above this line are highlighted in blue and labelled.

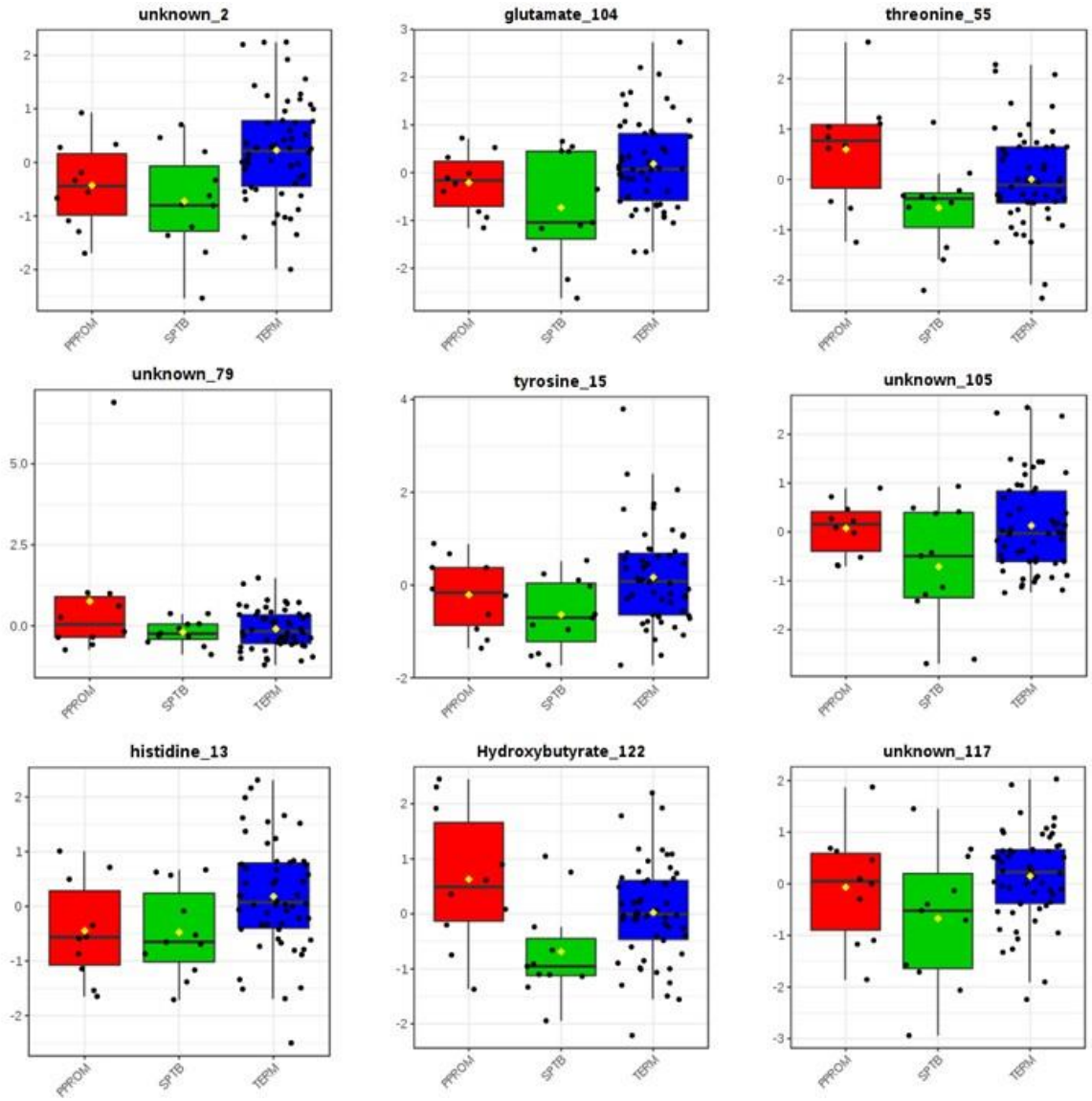


Figure 6.7. Boxplots of the assigned metabolite per group. The bottom and top of the box are the 25th and 75th percentile. The black dots represent the concentrations of the selected metabolite from all the samples. The mean concentration of each group is indicated with a yellow diamond.

Comparison of all participants at 20-week gestation (TERM n=51, PPROM n=9, sPTB n=11).

The uploaded data matrix file contained 73 samples by 145 spectra bins for analysis with zero missing values for the three analysis groups. Normalisation and scaling of the data were performed as described.

The one-way ANOVA comparing assigned metabolites is shown in Figure 6.8. Ten metabolite peaks annotated above the black line represent the features that showed a statistical difference based on p value <0.05. The box plots to allow comparison of groups is shown in Figure 6.9. Although the PPROM group (labelled in red) appears to have consistently lower mean values than sPTB and term groups, the variance around normal is large and there is overlap between the concentrations of metabolite in the PPROM/sPTB and the term group. There were no metabolites with an FDR <0.05.

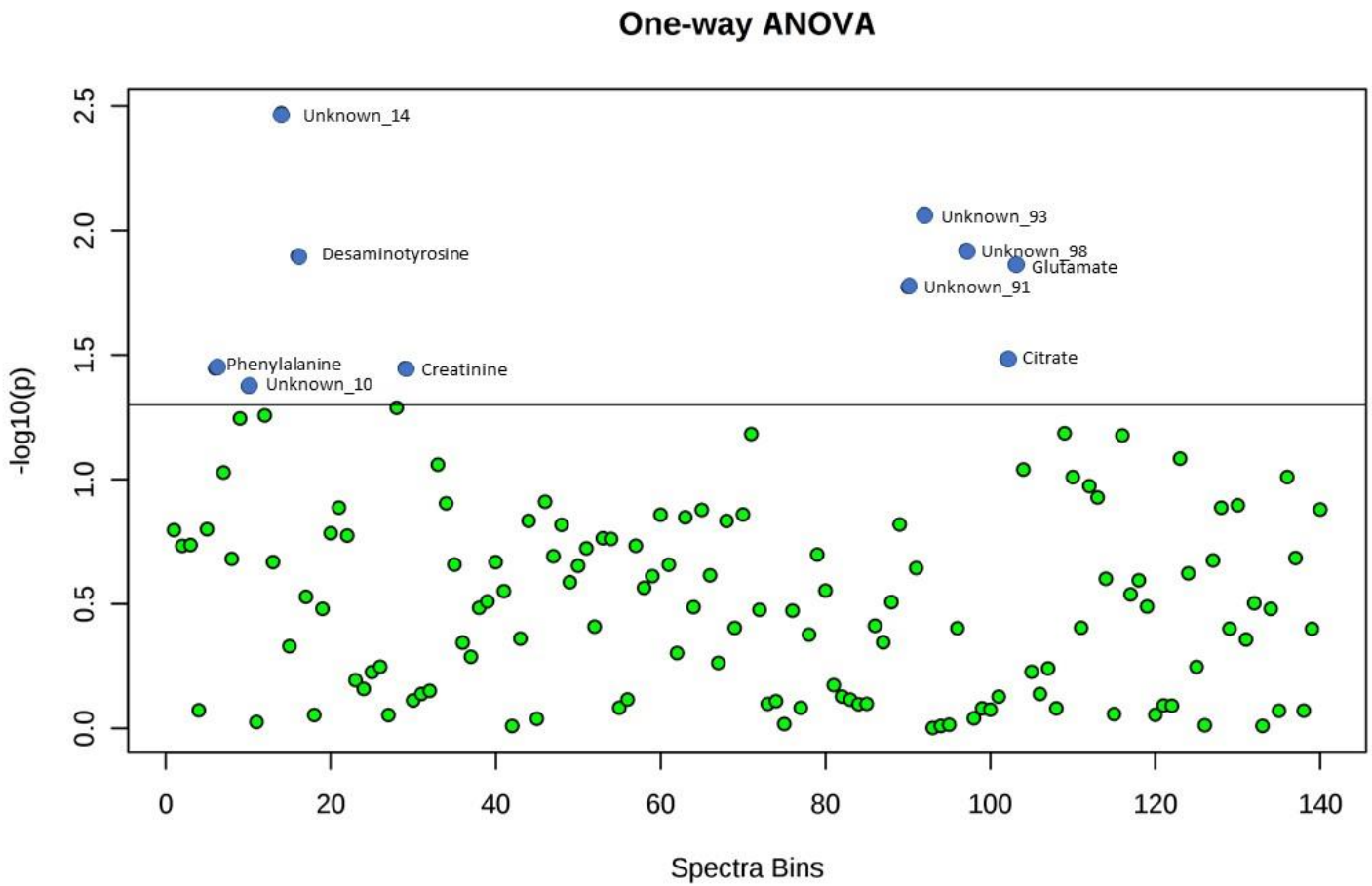


Figure 6.8. ANOVA comparing samples at 20 weeks gestation. The nominal (p val) 0.05 error rate is drawn across the graph as a black line. Metabolites above this line are highlighted in blue and labelled

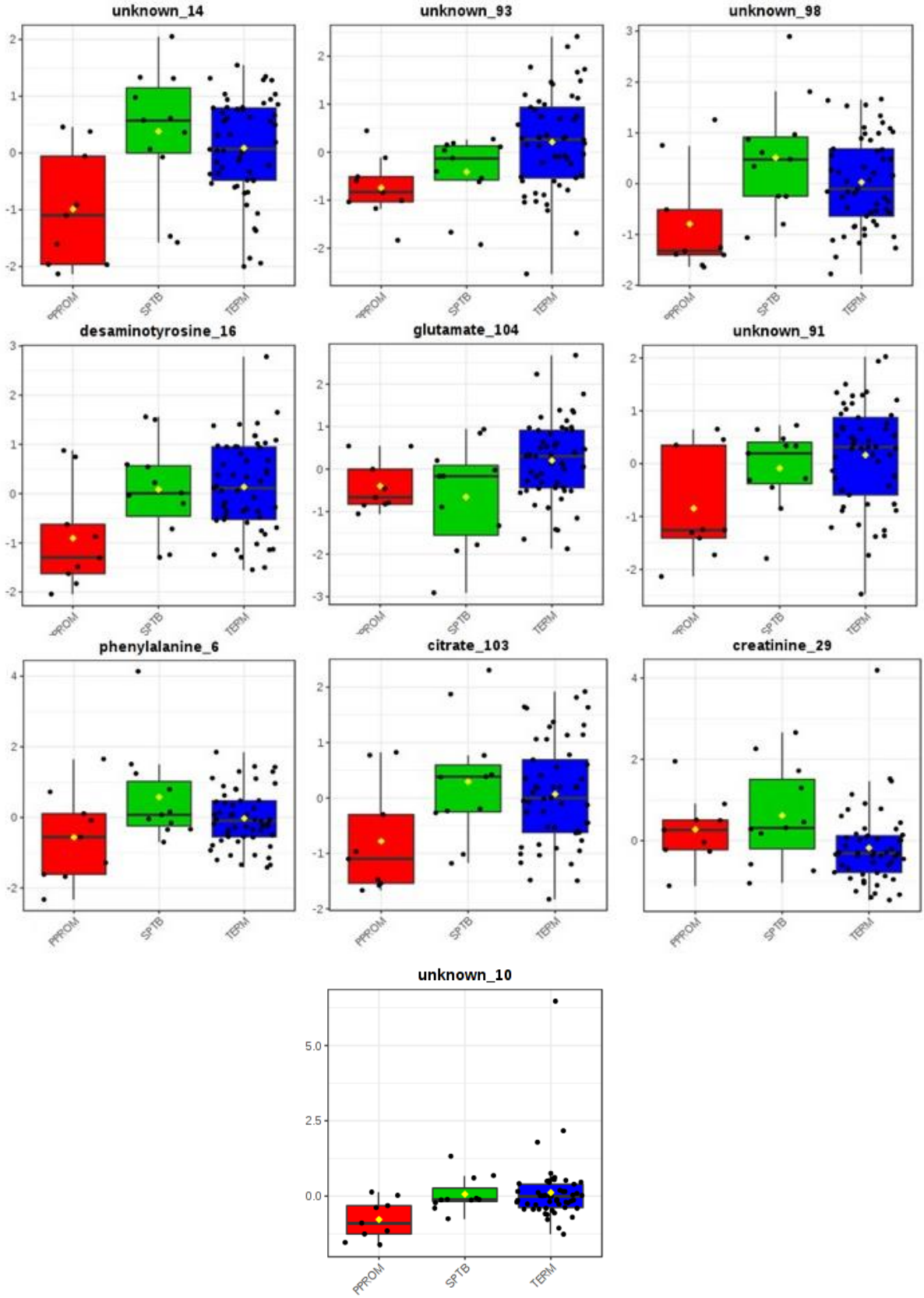


Figure 6.9. Boxplots of the assigned metabolite per group. The bottom and top of the box are the 25th and 75th percentile. The black dots represent the concentrations of the selected metabolite from all the samples. The mean concentration of each group is indicated with a yellow diamond.

Time-paired Sample Analysis

The next analysis compared temporal changes in the metabolites from week 16 to week 20. For this analysis a balanced design was required to perform one way repeated measures (within subjects) ANOVA, the groups had to be equal and therefore nine pairs of samples sPTB (n=9), PPROM (n=9) and TERM (n= 9) were analysed, with the TERM samples selected at random. 54 samples were included in the analysis matrix of 145 bins per sample, with 101 bins assigned to 35 metabolites. Following normalisation of the data, the data was viewed as 3D PCA plot (Figure 6.10), no outliers were identified and no clear separation between groups was demonstrated. The results of the temporal analysis ANOVA are shown in Figure 6.11 and Table 6.2.

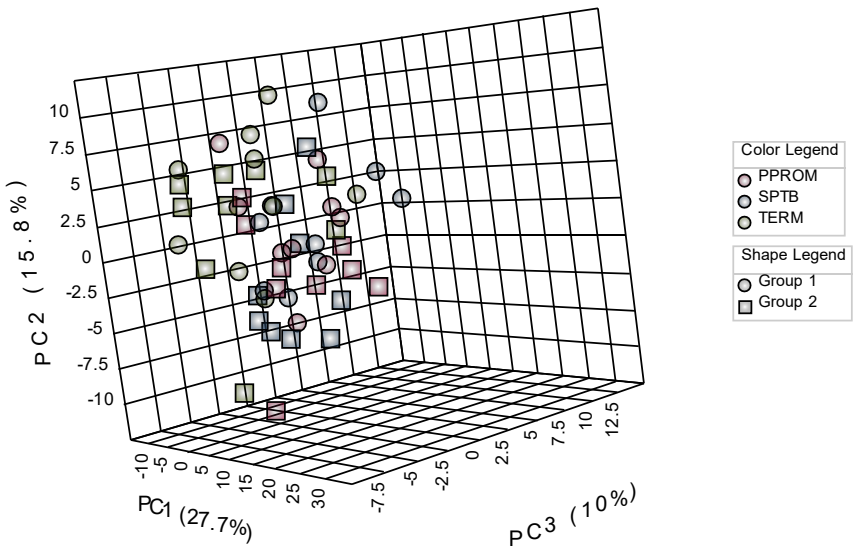


Figure 6.10. Still image taken from an interactive 3D PCA plot. No clear separation can be seen between the three groups.

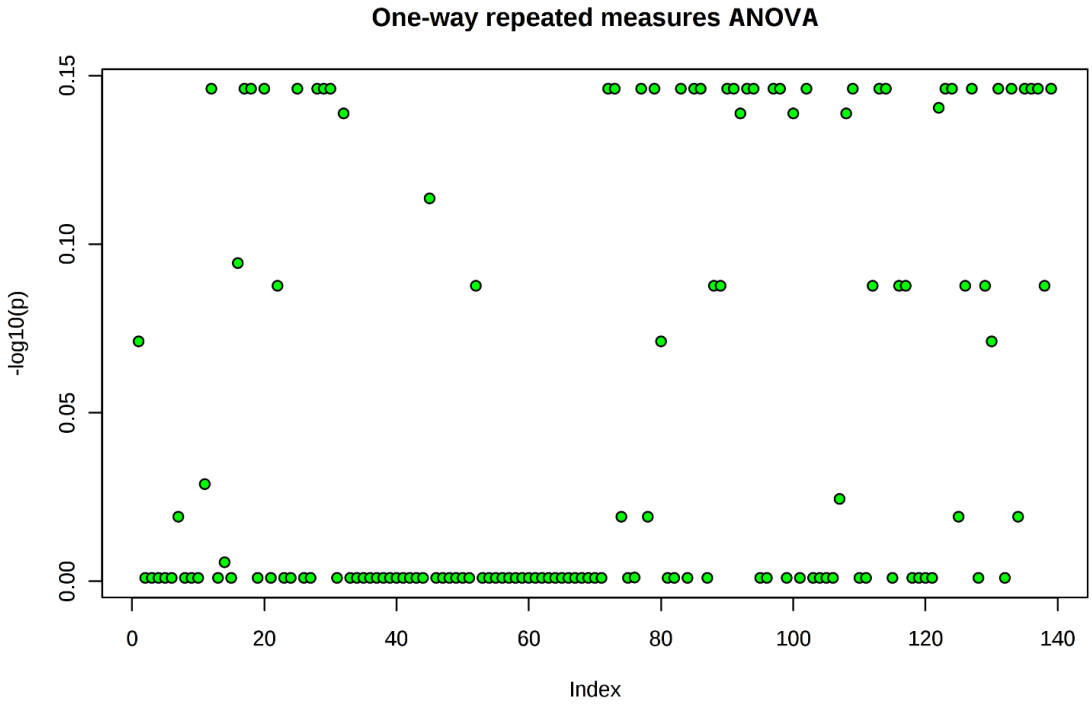


Figure 5. ANOVA for metabolites between 16 and 20 weeks. The table below the graph shows the p values and FDR (adjusted P-val) for the metabolites showing the biggest difference between groups over time. None remain statistically significant (<0.05) following adjustment for multiple testing

Table 6.2. Metabolites showing the largest temporal change between groups, p values and FDR (adjusted P-val) shown. None remain statistically significant (<0.05) following adjustment for multiple testing

Name	F-value	Raw P-val	Adjusted P-val
desaminotyrosine_12	38.642	**0.024915**	0.71431
Isoleucine_138	31.609	**0.030211**	0.71431
unknown_73	30.251	**0.031503**	0.71431
Leucine_141	18.043	0.051203	0.71431

Comparison of sPTB (n=11) versus PPROM (n=10) groups at 16 weeks

Samples taken at 16 weeks for women with sPTB (n=11) and women with PPROM (n=10) were compared. Fold change (FC) analysis compared absolute values of group means per assigned metabolite. The four metabolites reaching the FC threshold (1.2) are highlighted in pink (Figure 6.12). Boxplots of concentration values for these four metabolites, unknown 21/22, proline and acetoacetate are shown in Figure 6.13.

T test results found no significant features between the two groups for an adjusted p-value cut-off of 0.05 (FDR). Figure 6.14 shows the summary volcano plot of FC change and t test results (p values). Only acetoacetate showed a difference between the two groups with lower concentrations on average in the PPROM group (Figure 6.15). However, if using adjusted p values, acetoacetate is not associated with a difference between groups.

Principal component analysis score plot (Figure 6.16) shows no separation between the groups at 16 weeks. As no obvious differences can be seen between the two groups, no further analysis was performed.

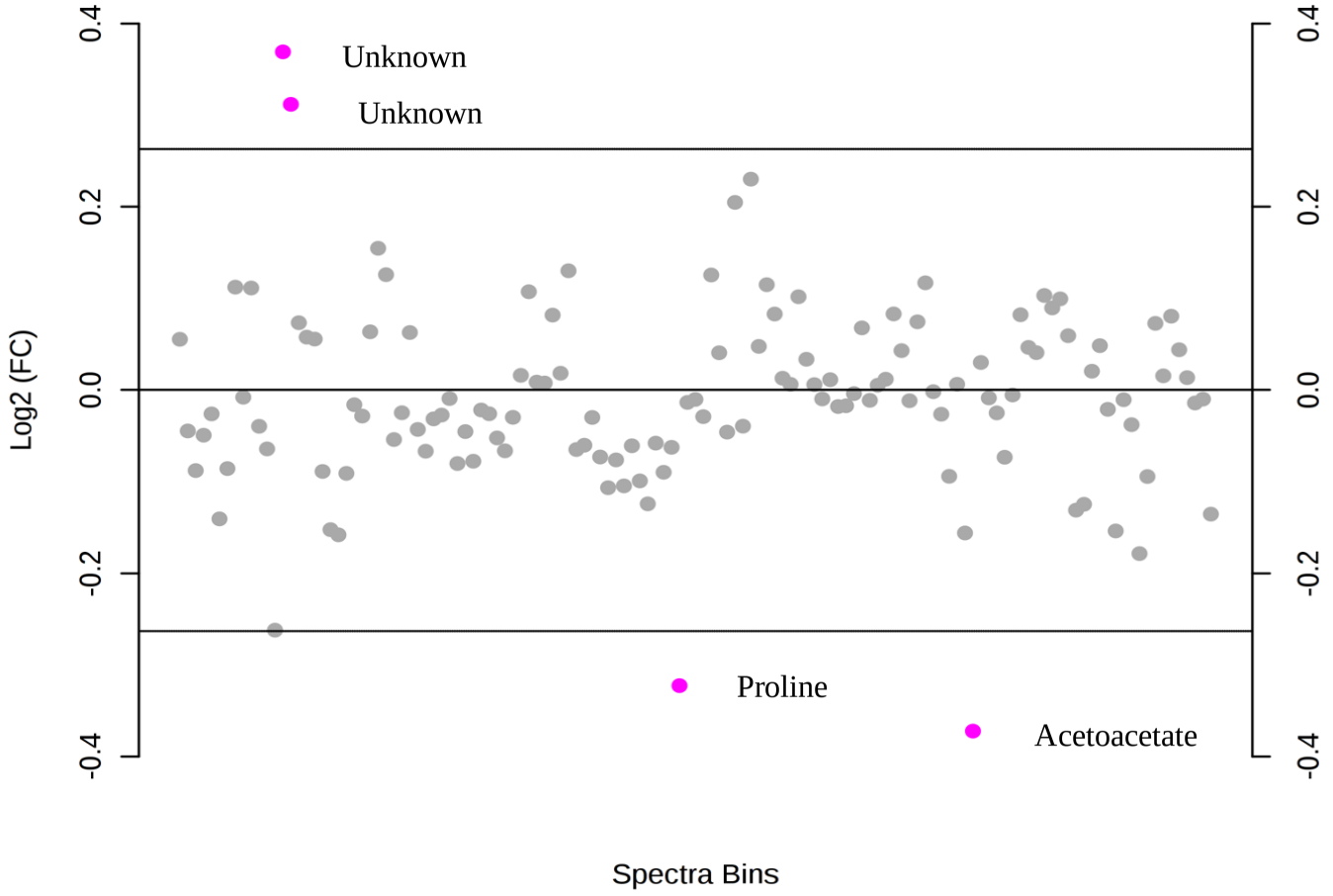


Figure 6. Results of fold change analysis comparing metabolite profiles between PPROM and sPTB, the fold change threshold is set to 1.2. Significant variables are shown in pink.

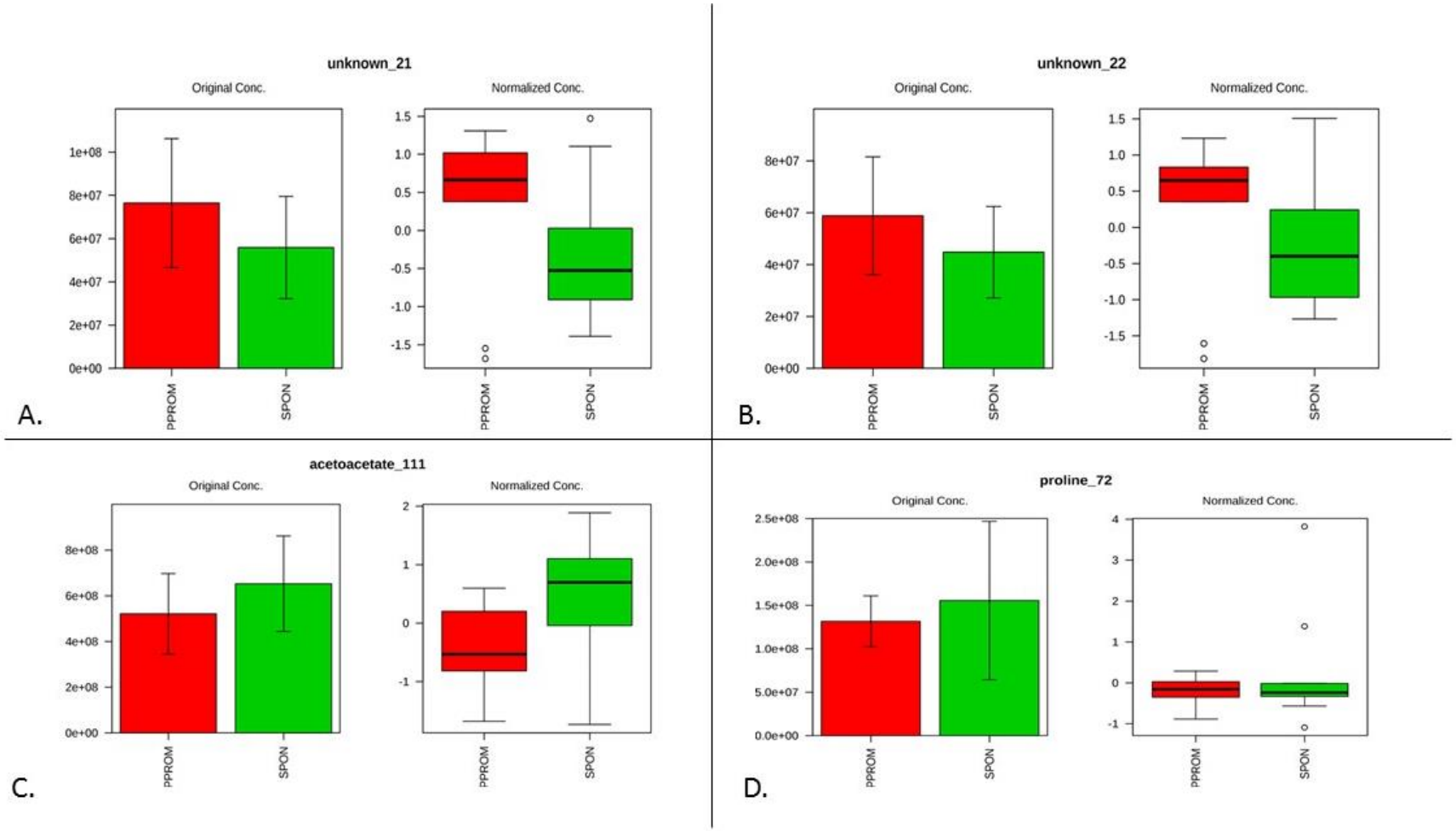


Figure 6.13. Bar plots on the left show the original value (mean +/- SD). The box and whisker plots on the right summarize the normalised values. a) Unknown metabolite (21) b) Unknown metabolite (22) c) Acetoacetate d) Proline.

Volcano Plot for 16 week metabolite comparison between sPTB and PPROM groups

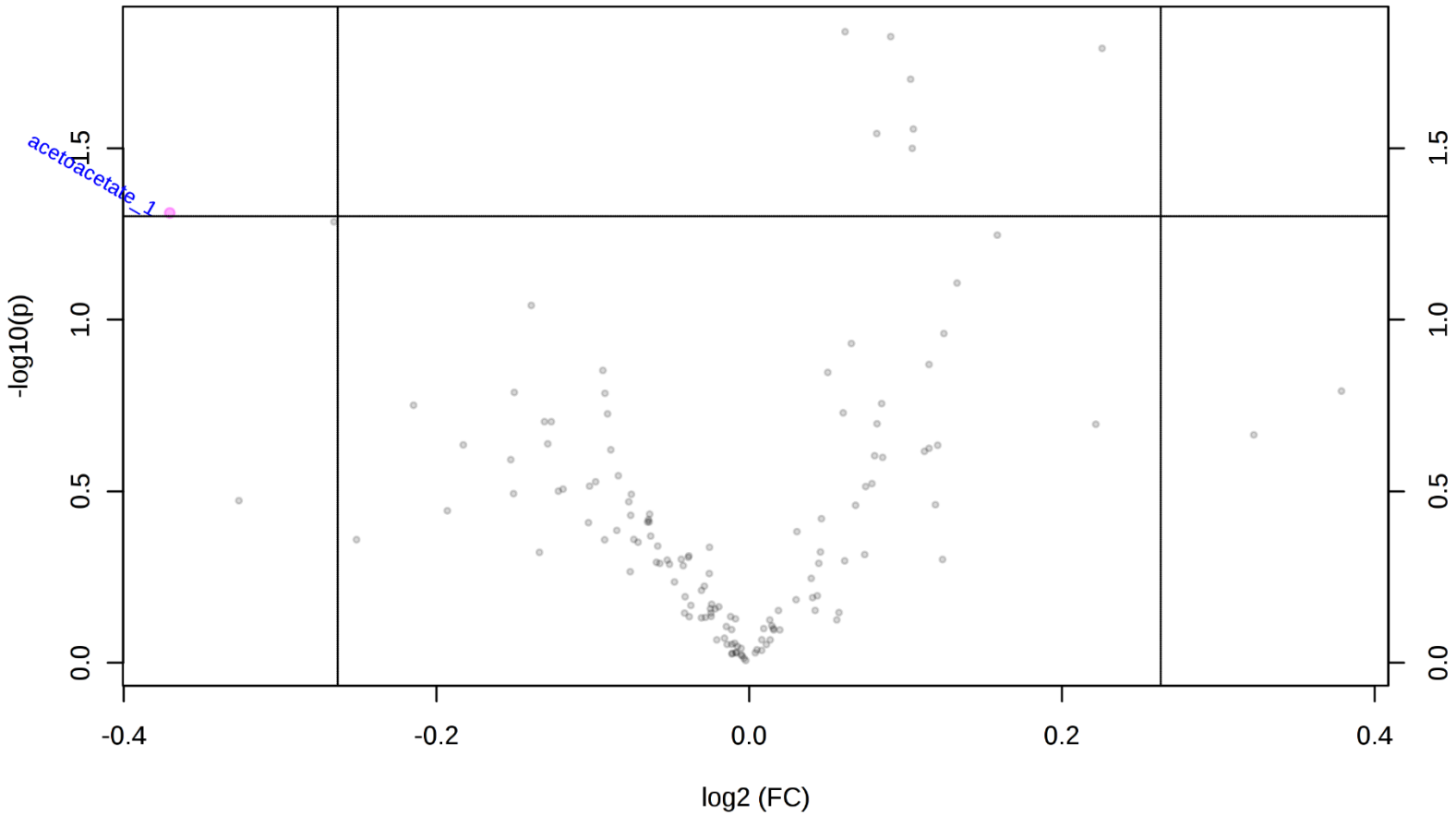


Figure 6.14. Volcano plot. The important features selected by fold change (x axis) threshold 1.2 and t-test raw p value threshold (y) 0.05. The pink circles represent features above the threshold

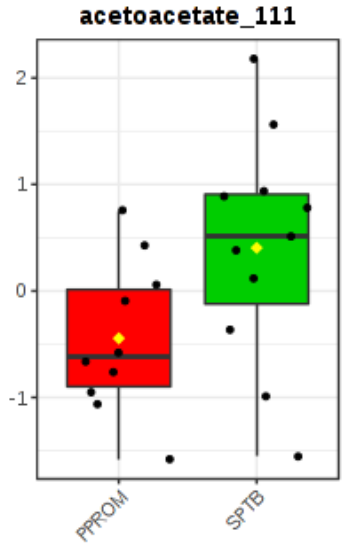


Figure 6.15. Boxplot of acetoacetate. The black dots represent the concentrations of acetoacetate in all samples. The mean concentration of the group is represented with a yellow diamond. The whiskers extend to the highest and lowest observations.

Score plot for top PCA to differentiate sPTB and PPROM at 16 weeks

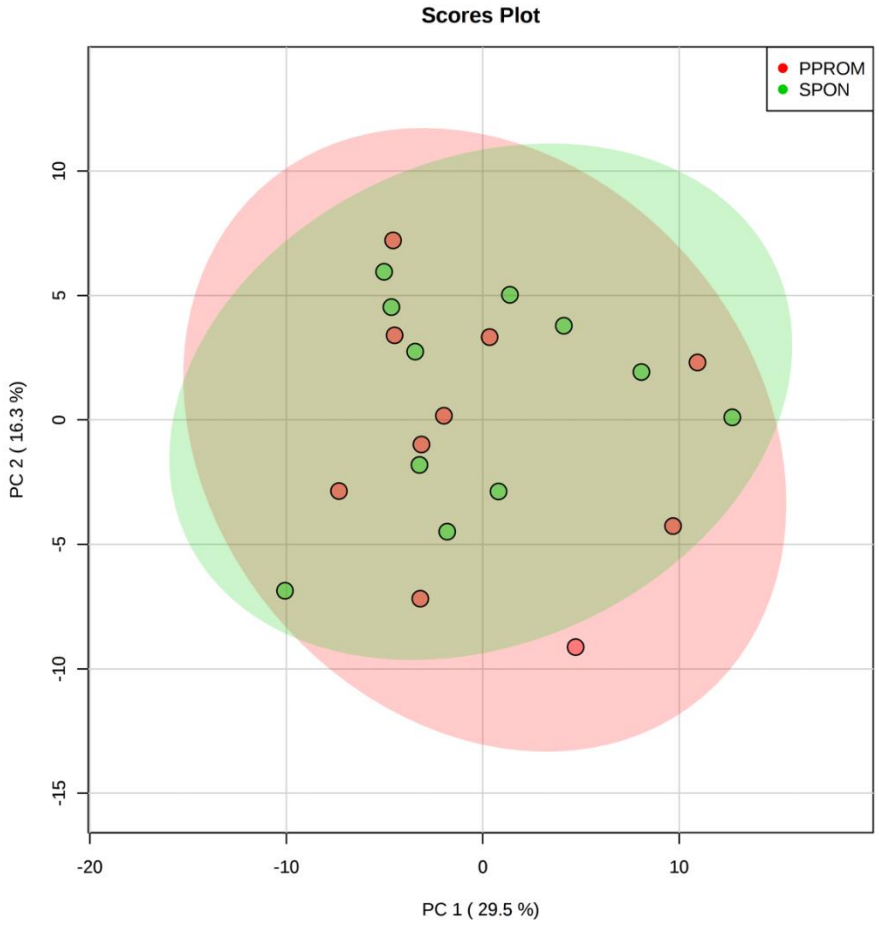


Figure 6.16. Score plot between top principal components (PC). The explained variances are shown in brackets.

Comparison of sPTB (n=11) versus PPROM (n=9) group metabolites at 20 weeks gestation

Serum analysis comparing metabolite spectra from the second 20-week timepoint was performed on 11 sPTB samples and 9 PPROM samples. Normalisation and scaling were performed as previously to achieve Gaussian distribution of data.

FC analysis (Figure 6.17) echoed the findings for 16 weeks and with a threshold of 1.2 both unknown metabolite 21 and 22 were differentiated between the groups (higher in PPROM group compared to sPTB). In addition, mannose, desaminotyrosine and unknown metabolite peak 1 had lower concentrations in the PPROM group compared to sPTB. T test analysis found no statistically significant differences between the two groups with an FDR of 0.05. The volcano plot (Figure 6.18) are shown below, with desaminotyrosine shown to be the metabolite with the largest difference between the two groups. The normalised concentrations are shown by boxplots in Figure 6.19.

A principal component analysis (PCA) was performed but shows little separation between the two groups with mostly overlapping 95% CI (Figure 6.20). A discriminant analysis was performed to sharpen the separation of the groups based on the metabolomic observations. This gives a better distinction between the groups (Figure 6.21a). However, when the cross validation of the model is investigated, it shows the model is a poor predictor and almost certainly overfitted due to the small number of samples (Figure 6.21b).

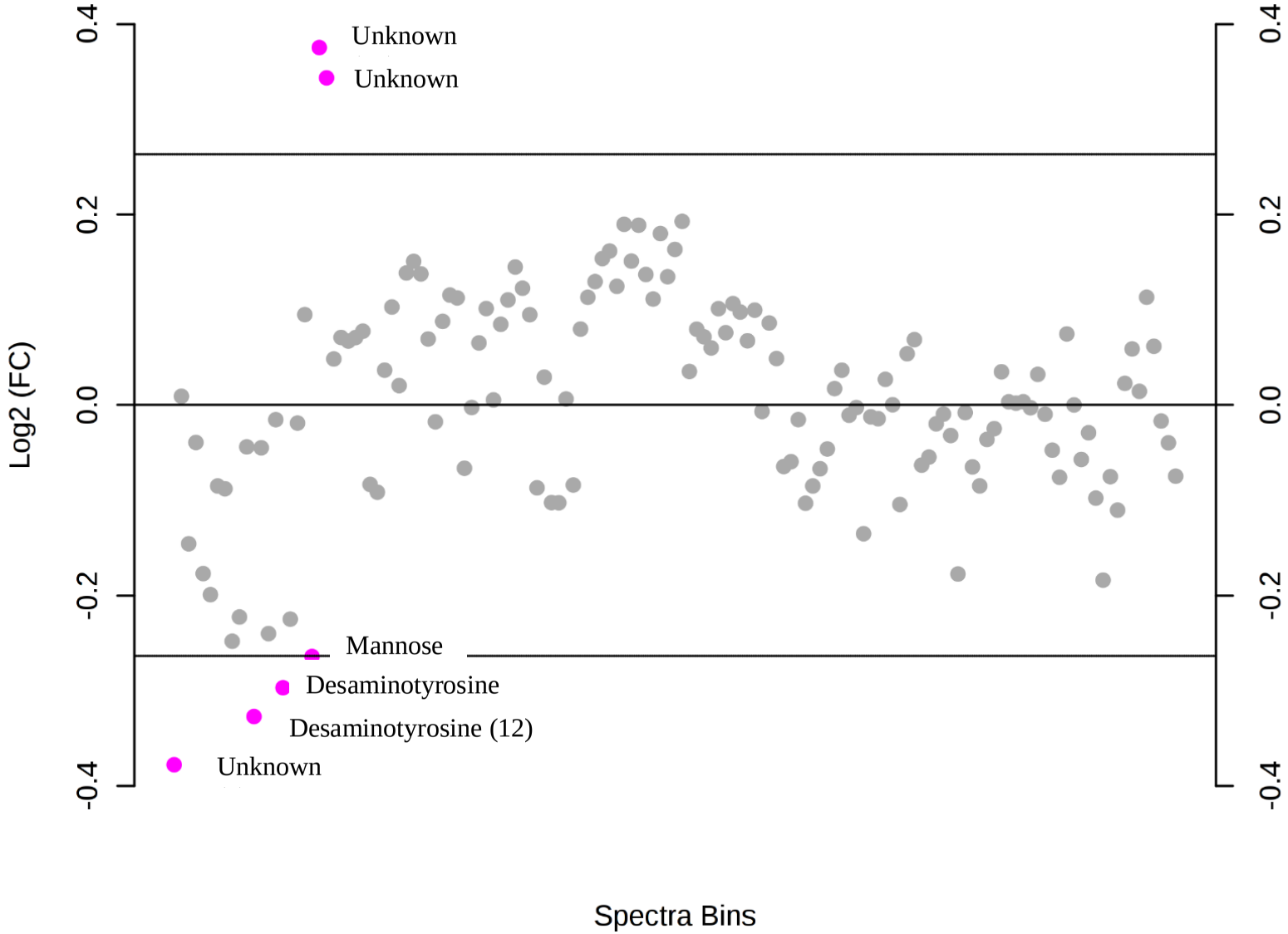


Figure 6.17. Results of fold change analysis comparing group means per metabolite profiles at 20 weeks between PPROM and sPTB, the fold change threshold is set to 1.2. Metabolites above the threshold (higher or lower) are shown in pink

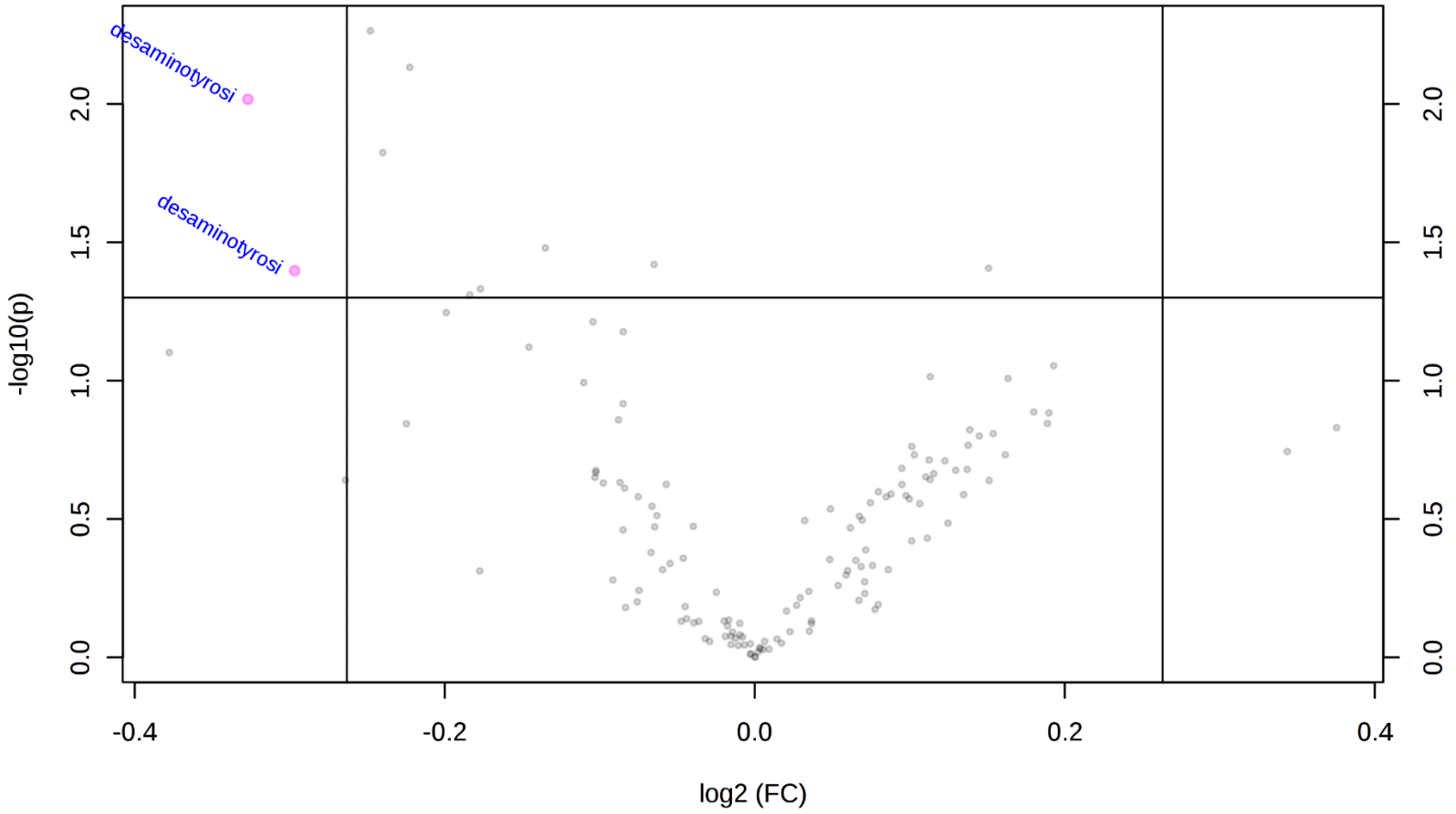


Figure 6.18. Volcano plot comparing sPTB and PPROM 20 week metabolites . The important features selected by fold change(x) threshold 1.2, and t-test (y) raw p value threshold 0.05. Desaminotyrosine is the metabolite that shows the biggest difference between the two groups and is highlighted in pink

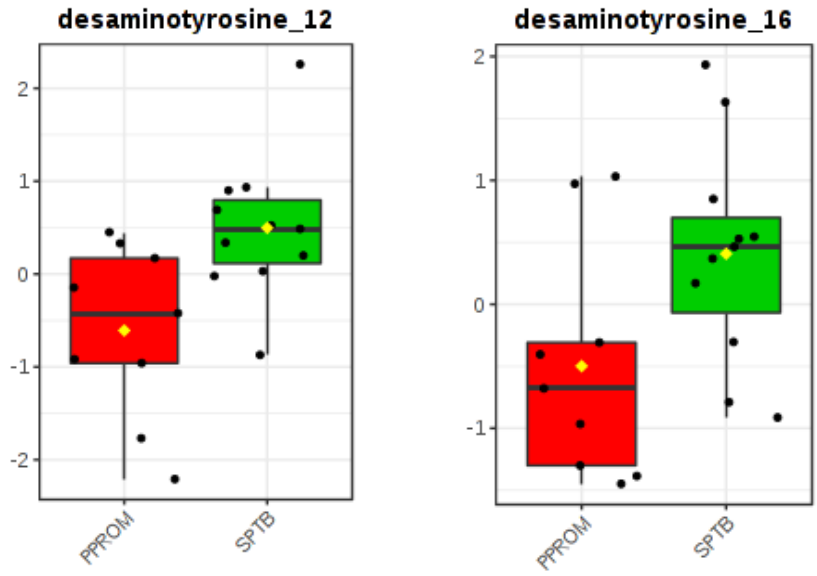


Figure 6.19. Boxplots to show differences in the metabolomic feature desaminotyrosine between PPROM and sPTB. The black dots show the normalised concentration for each sample. The mean concentration of each group is represented with a yellow diamond

Score plot for top PCA to differentiate sPTB and PPROM at 20 weeks

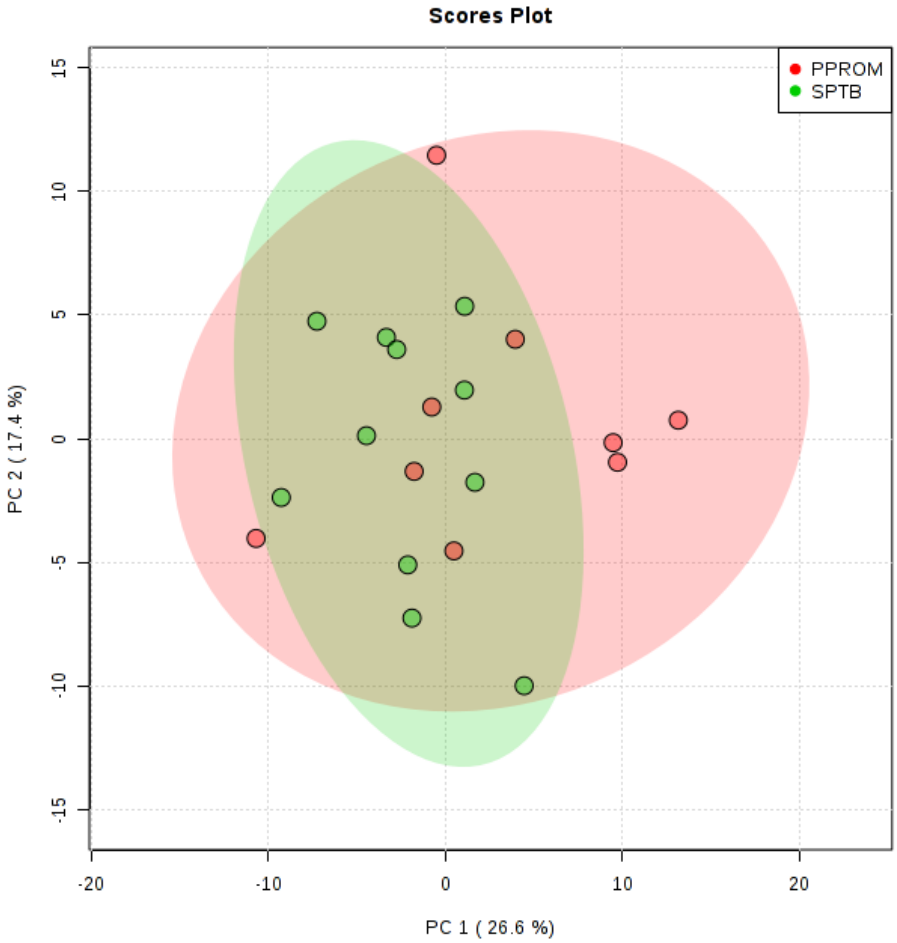


Figure 6.20. Score plot between top principal components (PC). The explained variances are shown in brackets. 95% CI are shown by coloured circles.

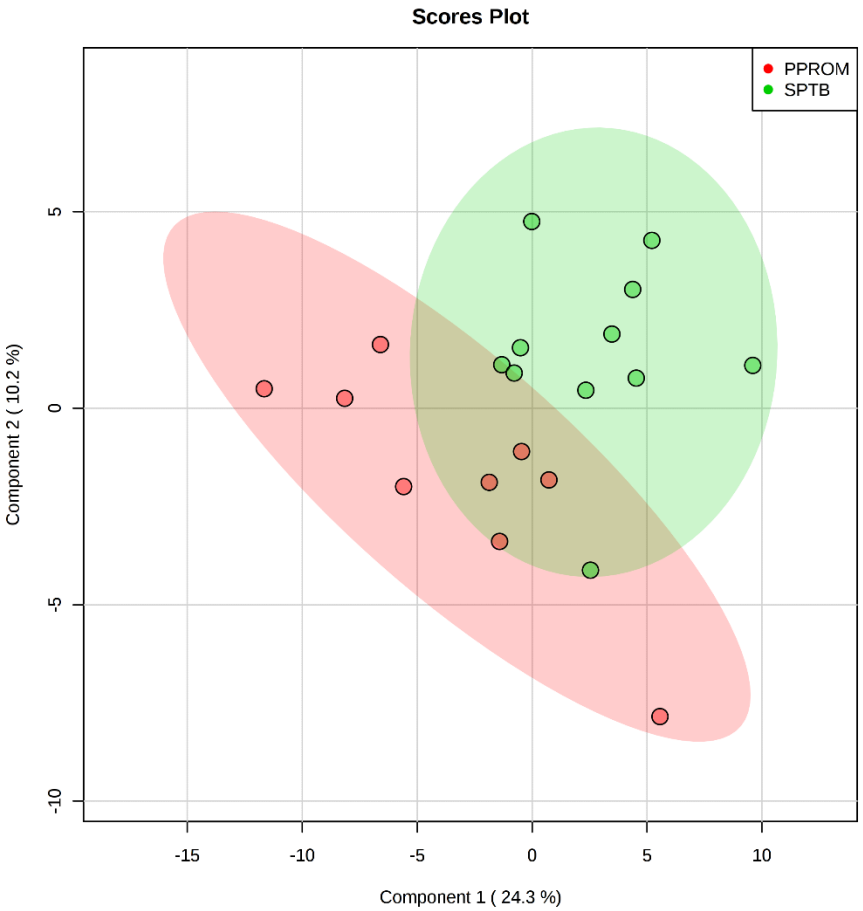


Figure 7A) PLS-DA. Score plot between the selected principle components. The explained variances are shown in brackets

.

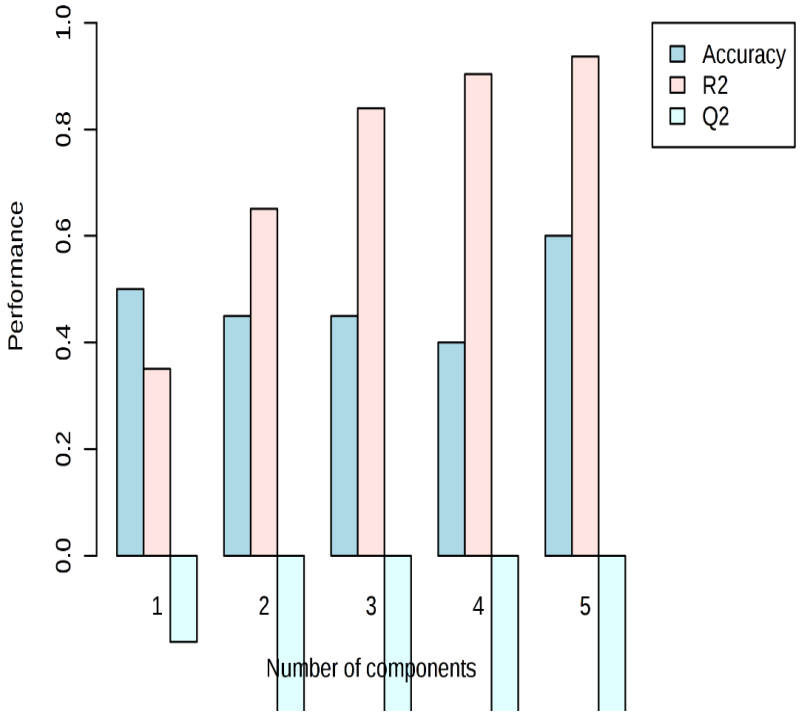


Figure 6.21B) PLS-DA Cross validation. Q2 estimates the predictive ability of the PLS-DA model, calculated by cross validation (CV). In each CV (1-5), the predicted data are compared with the original data, and the sum of the squared errors is calculated. Good predictions will have a high Q2. This model is not at all predictive and most likely overfitted.

6.5 Discussion

A detailed assessment has been made of the NMR profile of serum collected at two different time points in the second trimester from women at high risk of sPTB.

This study found that overall maternal serum samples could not differentiate between women experiencing PPROM, sPTB and TERM deliveries based on their metabolite profiles at 16 or 20 weeks. A subset of data to give balanced groups were used to perform a time paired analysis across 16 to 20 weeks in all three groups; sPTB, PPROM and TERM. No metabolites with temporal change showed any association with sPTB or PPROM.

No metabolites differentiated between PPROM and sPTB groups at 16 weeks. Acetoacetate showed a greater magnitude fold change than other metabolites *and* statistical significance (p value), however, it did not associate with sPTB following adjustment for multiple testing, which is not surprising given the small number of samples tested. However, this finding was not seen at the 20-week comparative analysis, which might be expected if this metabolite was predictive of a difference between sPTB and PPROM. It has also not been reported in other metabolomic studies of PTB. Acetoacetate is the conjugate base of acetoacetic acid and is released into the bloodstream during periods of fasting (Berg et al. 2002). As we have no information on time from last meal, we cannot be sure if the difference in fasting times between groups may have caused this result.

At 20 weeks, desaminotyrosine differentiated the groups after fold change and t-test analysis showed a lower maternal serum concentration level in women subsequently experiencing PPROM. Again, given the small sample size, it is not surprising that the FDR was >0.05. In Figure 6.9 the boxplot of desaminotyrosine

comparing all three groups (PPROM, sPTB and TERM), showed that mean levels for sPTB and TERM were similar, and only PPROM appears to have lower levels. Adding strength to this finding was the differentiation at multiple metabolite peaks for the same metabolite on the spectra. The NMR spectral signature for desaminotyrosine is shown in Figure 6.22. Desaminotyrosine (DAT) has four peaks on the spectra,

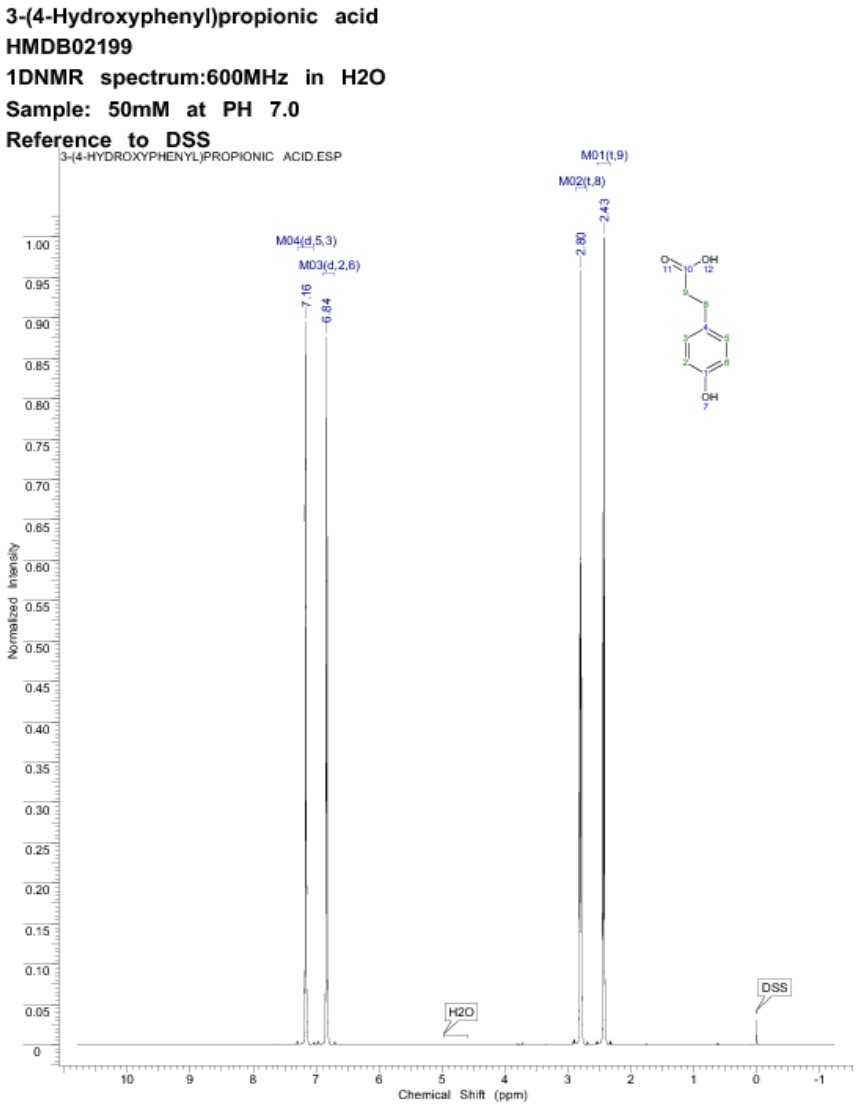


Figure 8. Figure 60. 1H NMR Spectrum (HMDB0002199). Compound name 'Desaminotyrosine'. Image taken from the Human Metabolome Database (accessed at http://www.hmdb.ca/spectra/nmr_one_d/1862 on 3 October 2019)

however only the two peaks in the ppm range 2-3 were detected by NMR, and both differentiated between groups.

It is difficult to understand biological plausibility behind participants with PPROM having lower DAT. Desaminotyrosine, also known as 4-hydroxyphenylpropionic acid, is a degradation product of flavonoids, a compound most commonly found in plants. The gut microbiota generates many small metabolites that enter the systemic circulation, and DAT is produced by human enteric bacteria from flavonoids and amino acids (Steed et al 2017). In mice, DAT produced by gut bacteria has been shown in mice to be protective against influenza through the augmentation of type 1 interferon (IFN) (Steed et al. 2017). However, when type 1 IFN has been studied in a pregnant mouse model, upregulation of type 1 IFN sensitises the animals to bacterial products predisposing to spontaneous PTB. (Cappelletti et al. 2017). Therefore, it is not clear how reduced DAT serum levels could contribute to PPROM. We did not collect detailed dietary information and low levels of DAT may be simply a surrogate for a poor diet containing low quantities of plants (i.e. low flavonoids) which also contain other antioxidant properties. If DAT is protective against viral illnesses through augmentation of type 1 IFN, it could be surmised that potentially there is a viral cause for PPROM and the sPTB and TERM groups are protected. This is highly unlikely as there has been very little evidence for the role of viral infection in PPROM to date. Out of 174 AF samples from patients with PPROM tested with PCR for human cytomegalovirus (HCMV), herpes simplex virus (HSV), parvovirus B19, human adenoviruses (HAdV), enteroviruses

(EV) and human parechovirus (HPeV), only 1 was positive for a viral genome (Bopegamage et al. 2013).

Of interest are the two unknown metabolites (or a single metabolite with two closely related peaks) found to differentiate the groups on fold change analysis at both 16 and 20 weeks increases the possibility that this may not be a chance finding. If this is truly a novel metabolite or drug degradation product to be identified, then the next step would be to try to perform a full structure elucidation. This would typically entail isolation and purification of the metabolite from serum and using techniques of NMR spectroscopy, MS, infrared spectroscopy and ultraviolet spectroscopy detail the metabolites full structure. (Dona et al. 2016).

Overall principal component analysis showed comparable variance between the two groups at both gestations tested and a predictive model could not be generated without overfitting data. This may be due to the small sample size and the community recognition of PLS-DA as a limited technique (Gromski et al. 2015).

This study is the first metabolomic study that has compared multiple timepoints in high risk women with a preterm birth <34 weeks, that has also attempted to establish if there are different metabolomic phenotype profiles between sPTB and PPROM.

Strengths of this study are the unique high-risk population from which these samples are obtained, the extent to which births have been phenotyped and that samples are available for multiple timepoints in the second trimester. Prior studies examining metabolites associated with preterm birth have noted the importance of collecting samples at the same timepoints instead of immediately prior to labour which is a limitation of some metabolomic studies (Menon et al. 2014, Lizewska et al. 2018).

A particular strength of this study is focussing on predictors in the early preterm birth group (<34 weeks). Naturally it is a harder group from which to obtain samples as there is a much lower incidence than late preterm births (34-37 weeks). Many studies which at first glance appear to present impressive numbers of participants, often have included all sPTBs <37 weeks which clinically has less value due to the burden of morbidity and mortality associated with the earlier gestations of delivery. An example would be the largest metabolomic study from the SCOPE cohort that recruited 5,690 low risk nulliparous women across several hospital sites in four different countries (Souza et al. 2019). Data was available from 55 sPTB from Cork in Ireland (discovery) and 55 women with sPTB from Auckland in New Zealand (validation). Approximately half of whom had been classified as PPROM, but unlike our study not analysed as a separate group. Only 16 women in Auckland and 13 women in Cork delivered <34 weeks, similar to the numbers in our study (n=21 at 16 weeks, n=20 at 20 weeks if PPROM and sPTB were combined). Using GC-MS analysis they found that elevated alkanes (decane, undecane and dodecane) were higher in sPTB <37 weeks in the Cork cohort only, but there was no evidence these alkanes were associated with sPTB < 34 weeks in either study site. (Souza et al. 2019)

The strength of NMR analysis is that it isn't hypothesis driven and can provide new insights into the pathology of complex diseases such as preterm birth. The entire visible metabolome is taken into account and metabolites with both large and small effect contribute to differentiation of groups and predictive modelling.

Appropriate scaling and normalisation was performed as part of the study design. PQN normalisation has been shown to be more accurate than integral normalization for 1H NMR metabolomics (Dieterle et al., 2006).

Limitations of this study include 1) small sample size, 2) methodological limitations in identifying all metabolites, 3) absent validation cohort, 4) no controlling for type of diet or fasting times.

Our findings are somewhat limited by the relatively small number of cases used, despite the well phenotyped cohort of births <34 weeks. Our small number of cases and controls may increase the risk of reporting false negatives. This is particularly important for our smaller sub-groups of sPTB (n=12) and PPROM (n=10) that did not differentiate in this analysis when false positives were properly controlled for however, some differences have been suggested by the data that warrants further exploration. Therefore, the null hypothesis cannot be confidently excluded.

The disadvantages of the 1H NMR method using multiple bins are that it is time consuming to assign metabolite peaks to get maximum identification of metabolites and can be subject to data interpretation errors due to overlapping peaks. However overall, NMR is highly reproducible but limitations in our current knowledge meant that not all metabolite peaks could be identified at this time, illustrated by unknown samples 21 and 22.

We did not perform a validation study as the purpose of this thesis was to combine different omic layers. However, an important progression of this study is to identify a validation cohort to establish the reliability of these results. Additionally, studying samples from low risk women (i.e. no previous preterm births) may see bigger discrepancies in effect sizes that are captured more easily.

We did not control for other probable sources of variability in the metabolite profile such as time of day, type of diet and time from most recent meal. Normalising for these many variables may have improved separation between our groups.

However, we did not collect this type of self-reported data from our participants and therefore this depth of analysis was not possible.

6.6 Conclusion

From 1H NMR analysis of different birth phenotypes there is no data to suggest there are metabolite differences at 16 and 20 weeks or between 16 and 20 weeks, that can contribute to prediction of birth phenotype. Our data suggests some potential differences between sPTB and PPROM groups, but these findings should be interpreted with caution given the small study size. Further work in this area should be pursued before ruling in our out any differences. Overall, there was no purely NMR metabolic model that could clearly discriminate the groups leading to the possibility that any successful model would require additional information (enzyme, gene, physiological) or more sensitive metabolite measurements (MS).

Chapter 7: Integromics for the Prediction of Spontaneous Preterm Birth

7.1 Introduction

The objective of integrating multi-omic data (*Integromics*) is to construct a model that can be used to predict women who will have sPTB. By integrating the omics data from the same women, as opposed to data sets from different populations, we are increasing the chance of identifying 'cross-talk' across omics platforms and potentially identifying predictive biomarkers. The previous four chapters explored the data from individual "omics" layers and phenotypes from this cohort of pregnant individuals and highlighted the quality control methods employed to ensure a reliable input dataset. This chapter will outline the combination of these data.

Different analysis methods of combining omics data were previously discussed in chapter 2. In summary, the meta-dimensional approaches to combine different omics dataset for analysis included concatenation, transformation and model-based integrations, each with their own strengths and weaknesses. We chose a concatenation-based model as it's a) is relatively easy to apply statistical tests for categorical data analysis and b) does allow for crosstalk between omics layers via interaction. We tested seven different machine learning strategies on our dataset to identify the most predictive strategy and provided a prediction of sample sizes for a new multi-omic study required for replication. Finally, we interrogated the most predictive model for novel biological hypothesis.

7.2 Methodology

Study Population

From the cohort described in chapter 3 we included only women that had data genomics, transcriptomics and proteomics once quality control processes were complete. How these results were obtained is described in chapters 4, 5 and 6. Women with a sPTB or PPROM <34 weeks (as per chapter 3 definition) were included as cases and women with delivery >37 weeks were controls. All women delivering between 34-37 weeks (late sPTB) were excluded.

Multivariate Modelling

Using concatenation-based integration, a large matrix of all data types was created from the raw data. Data combination followed by machine learning analysis performed by Professor. Bertram Müller-Myhsok, an expert in statistical genetics and machine learning based at the Max-Planck Institute of Psychiatry, Munich, Germany.

Cross Validation

In this analysis, while the subjects were independent, the samples collected from various trimesters of the same subject were not. The unsupervised nature of the multi-omic methods makes it difficult to determine whether a method is overfitting or identifying a true biological relationship. To account for this, we designed a '*leave-one-out'* cross-validation strategy. A model was trained on all available samples except for all the samples at 16 and 20-week timepoints of a given subject. The model was then tested on all samples of the subject that it was blinded to. This process was repeated for all subjects until a blinded prediction was produced for all samples to confirm that the reconstruction error of the model on the left-out points is

close to the error in the training set. Final results are reported using these blinded predictions. This ensures complete independence from any inter-subject correlations.

Multi-Omic Machine Learning Method Comparisons

We ran prediction algorithms on the cross-validation data. The seven statistical approaches used were 1)'linear discriminant analysis' (Fisher. 1936) 2)'Genetic expression programming' (Ferreira. 2001), 3)'K-means' (MacQueen. 1967), 4)'Support vector machine with a linear kernel', (Cortes and Vapnik. 1998, Bradley and Mangasarian. 1998), 5)'Support vector machine with a Gaussian Kernel' (Cortes and Vapnik. 1998, Bradley and Mangasarian. 1998), 6)'Probabilistic neural network (Specht. 1990)' and 7)'Random Forest' (Ho. 1995). These were chosen as representatives of commonly used types of methods. We estimated the area under the curve (AUC) of the receiver operating characteristic (ROC) in cross-validation data with the null set equal to an AUC score of 0.5 (equal to random guess), which enabled comparison of the performance of the different methods on this dataset. This methodology also permits calculation of sample sizes for future multi-omics project.

Linear Discriminant Analysis

Discriminant analysis is a popular method for multiple class classification (sPTB, PPROM and TERM). In its original form it goes back to Fisher (1936). It focuses on reducing the original variable data matrix into a lower dimensional space and maximises the separability among known categories. This is performed through three steps; the first is calculating between-class variance, the second is calculating within-class variance. The third step is to reduce the dimensional space to maximise between and within-class variance. Essentially LDA maximises the distance between

separate variable means whilst minimising the variation (or scatter) within each category (Tharwat et al. 2017).

Genetic Expression Programming

Genetic Expression Programming (GEP) is a learning algorithm invented by Candida Ferrerira in 2001 (Ferreira. 2001). GEP learns specifically about relationships between variables in sets of data and subsequently builds models to explain these relationships. It functions through an architecture based on two entities; 'the chromosome' and the 'expression tree (ET)'. This is a full-fledged genotype/phenotype system with expression trees of different sizes and shapes encoded in linear chromosomes of fixed length. GEP chromosomes are multigenic, encoding multiple expression trees or sub-programs that can be organized into a much more complex program. Figure 7.1 shows the reproduction process containing the modifications performed by the algorithm operations that allows for the evolution of a simple replicator system.

K-means

The term "k-means" was first coined by James MacQueen in 1967 (MacQueen. 1967). K-means is the most popular clustering algorithm (Jain. 2010). It is an unsupervised learning algorithm. K-means clustering attempts to divide data into "k" number of separate groups and is effective at uncovering novel patterns (Theobald. 2017). In our analysis, K was set to two (sPTB and TERM). In the first step, the algorithm examines the unclustered data and selects a central point or centroid for each of the clusters. The rest of the datapoints are then assigned to a centroid using the Euclidean distance. Once all datapoints are allocated, the mean value of the datapoints in each cluster is aggregated. These are then used to update the centroid co-ordinates, which may affect the Euclidean distances of the datapoints resulting in

some datapoints switching clusters. If this happens the whole process is repeated, until there is no more movement (Theobald. 2017).

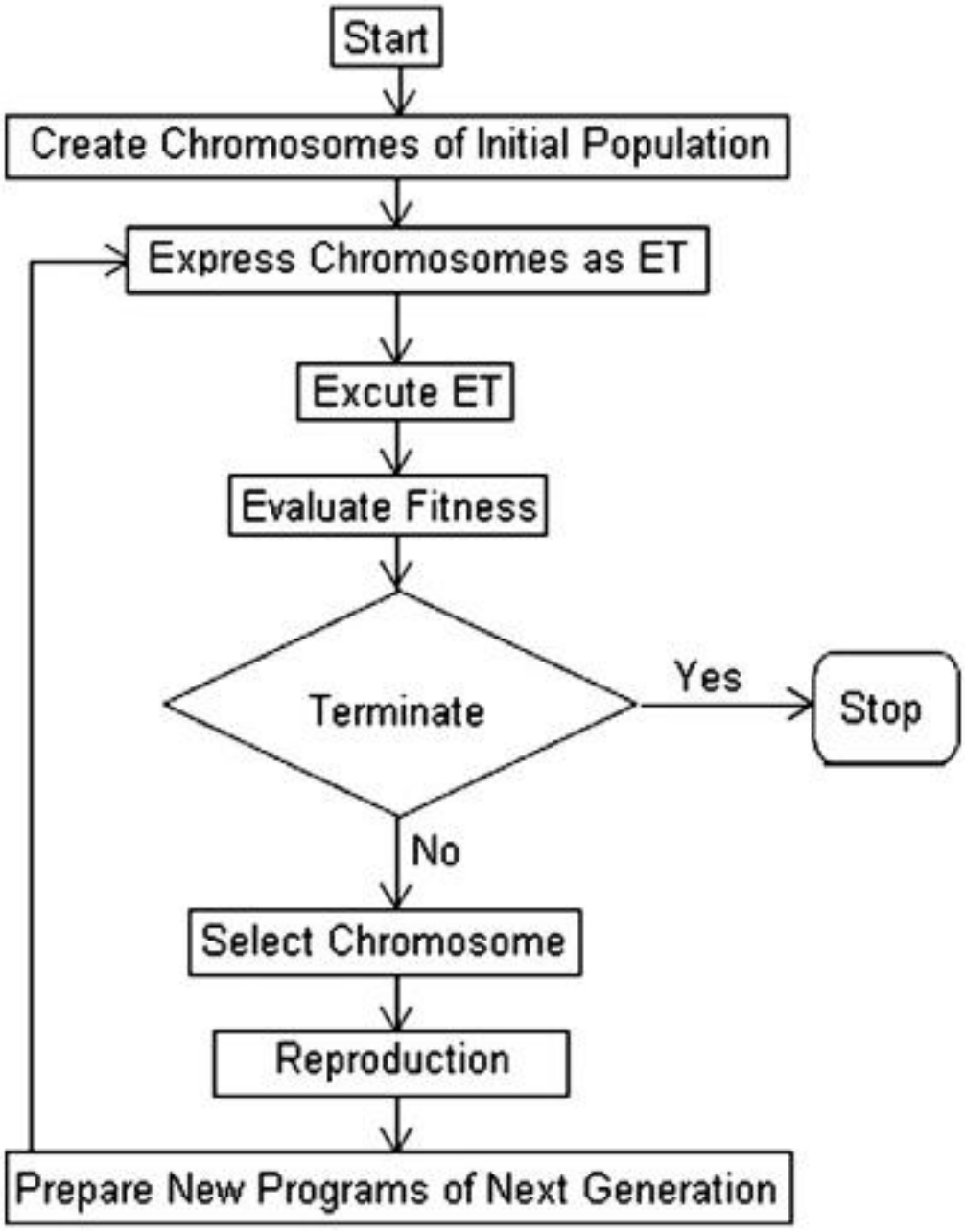


Figure 7.1. Flowchart from GEP algorithm. From Gullu, H. 2012. Prediction of peak ground acceleration by genetic expression programming. *Engineering Geology*. 141-142: 92-113. Adapted from Ferreria, C. 2001. Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. *Complex Systems*. 13(2): 87-129.

Support Vector Machine with a Linear Kernel

Support vector machine formulation is used to enlarge the feature space of predictors to create a decision boundary that is linear, where the original decision boundaries are non-linear. With already so many features, enlarging a feature space even further may make the data unmanageable. However, the support vector machine allows for enlargement of the feature space in a way that leads to efficient computation (James et al. 2013). It creates a fast, linear programming algorithm that will discriminate between massive datasets in n-dimensional space and results in an optimal separating plane for the entire dataset (Bradley and Mangasarian. 1998).

Support Vector Machine with a Gaussian Kernel

A Kernel is a computational function that quantifies the similarity of two observations (James et al. 2013). A kernel does not always have to be linear, it could be polynomial, radial, or as used here, gaussian. The advantage of using a kernel, rather than enlarging the feature space, is that it makes it computationally faster to perform these functions on data pairs, avoiding always having to work in the enlarged feature space. In summary, this is another way of creating hyperplanes to capture decision boundaries between groups.

Probabilistic Neural Network

This is a popular classification technique in machine learning to process data through layers of analysis. Like neurons in the human brain, networks are formed by interconnecting neurons, called nodes, which interact with each other through axons, called edges (Figure 7.2). There is an “all or nothing” arrangement as the sum of the connected edges must satisfy an activation threshold to communicate with the node at the next layer (Theobald. 2017).

The final decision node will output the class with the highest summed activation. This has the advantage of being a very flexible type of analysis, but computationally takes much longer to complete. The specific implementation used here goes back to Specht (1990).

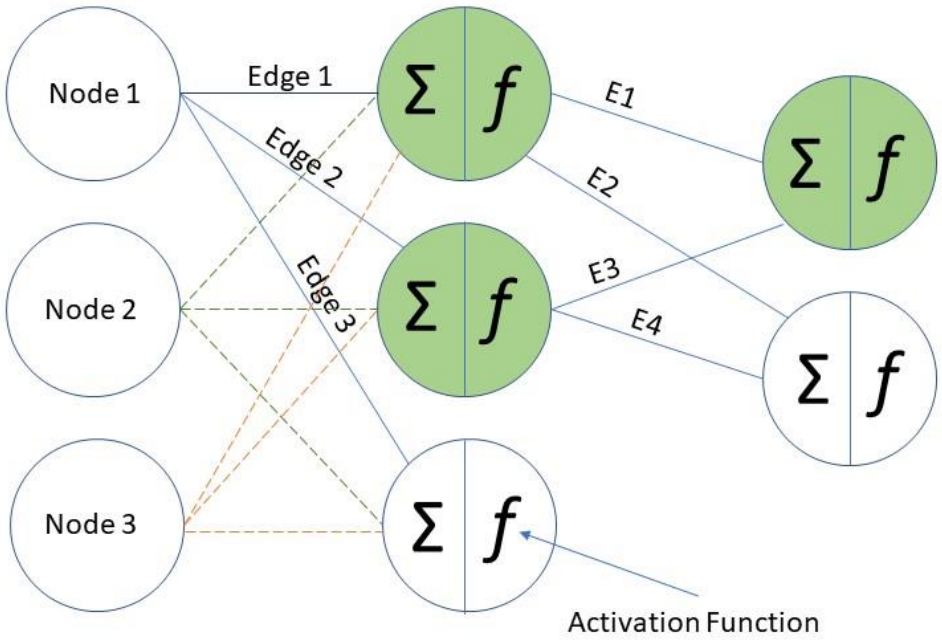


Figure 7.2 Schematic of basic neural network. Nodes are stacked in layers. The first layer of input is the raw omics data divided into nodes. Each node sends information to the next layer via edges. If the sum of the connected edges satisfies a set threshold (activation function) this activates the node at the next layer.

Random Forest

The random forest methodology has been previously explained in detail in section 5.2. In short, this technique uses multiple decision trees with an artificial cap on the number of variables that can be considered for each split. It has the advantage of being relatively easy to perform and computationally fast.

Sample Size Estimations

Once all classification algorithms were run, we then calculated the predicted sample size required for a three-layer multiomics study against different accuracies of AUC. Under the assumption that we would have 80% term controls and the same biomarker effect sizes as well as phenotypic distribution, we estimated the sample sizes needed via the R function power.roc.test from the R package pROC.

7.3 Results

Four women with sPTB, one woman with PPROM and ten women delivering at term had full omic data available at both timepoints analysis. As there was only one PPROM participant this could not constitute a group and would not have been classified by machine learning. Having had such poor differentiation of PPROM and sPTB classes in the omics analysis performed in the other chapters, we kept this as a case and included it as a case of sPTB (Figure 7.3).

The AUCs for all six tests can be seen in Figures 7.4-7.10. A summary of the results of the machine learning analysis are shown in table 7.1. K-means and probabilistic neural network were the most predictive statistical tests with an AUC of 1.00. The variable importance's for each test are included in Appendix M.

Figure 7.11 demonstrates that for a clinically predictive AUC (>90) it would be possible to obtain reasonable power with a relatively small number of women recruited to a multi-omics study. Using this graph, if we expected a minimum AUC of 0.9 to be found with a repeat study, with an alpha level of 0.001 (probability of rejecting the null hypothesis when true) being desirable, then we can estimate that a sample size of 50 would be required This calculation is based on 80% term controls which would equate to 40 term controls and 10 sPTB cases in a two group comparison.

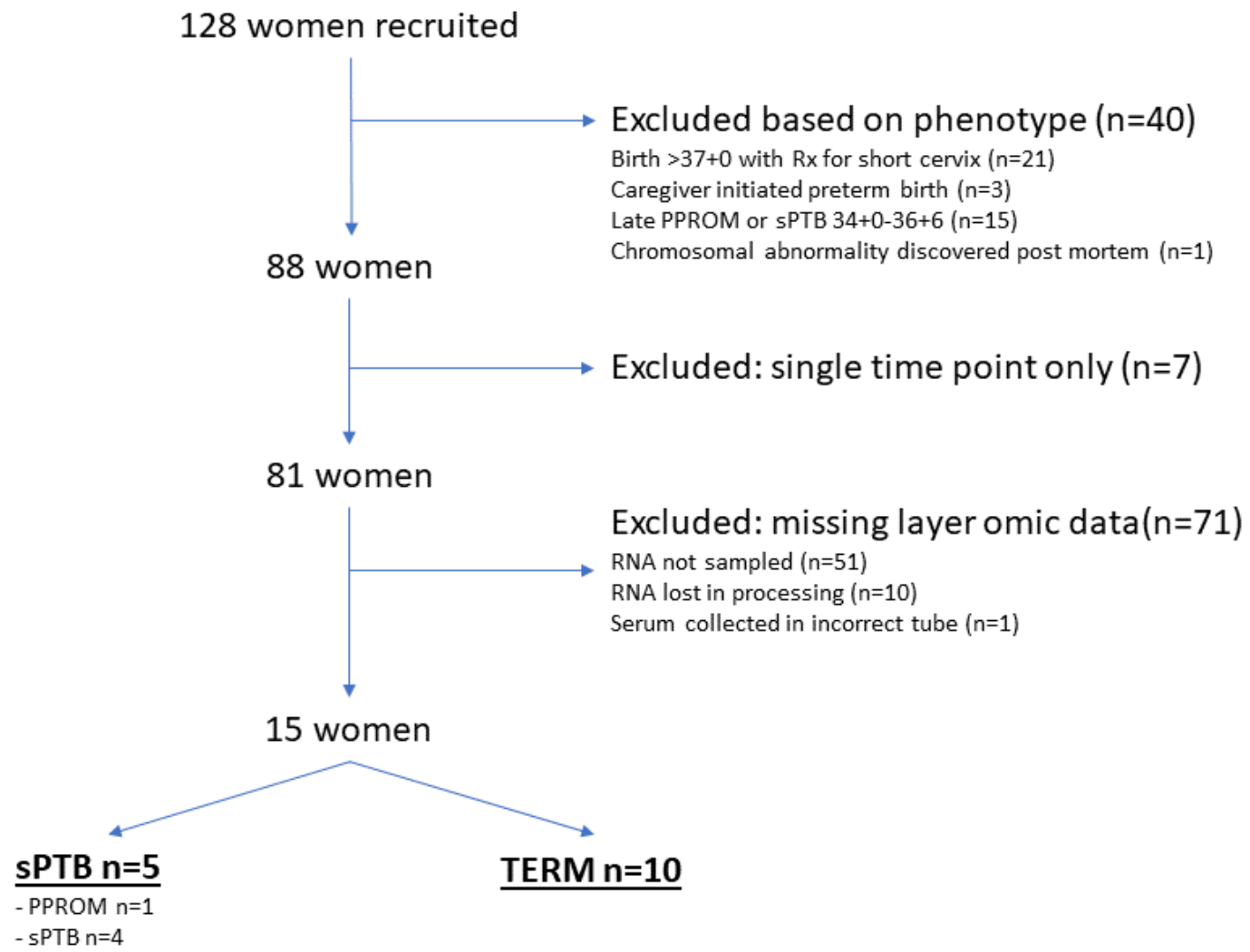


Figure 7.3. Flowchart of included participants in final omic analysis

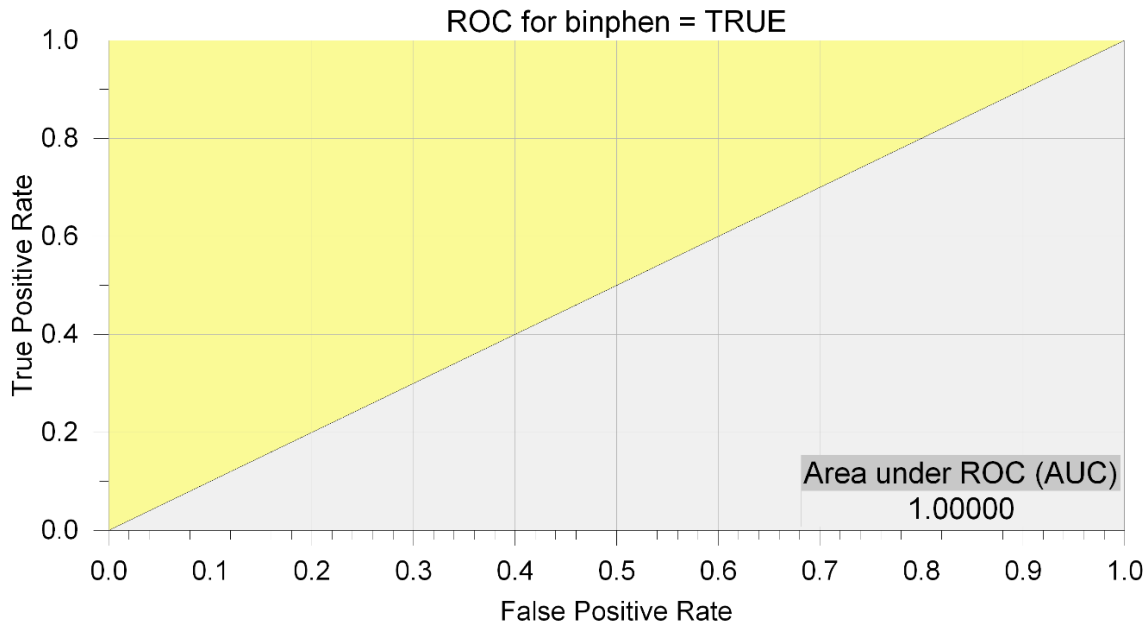


Figure 9. AUC for Linear Discriminant Analysis

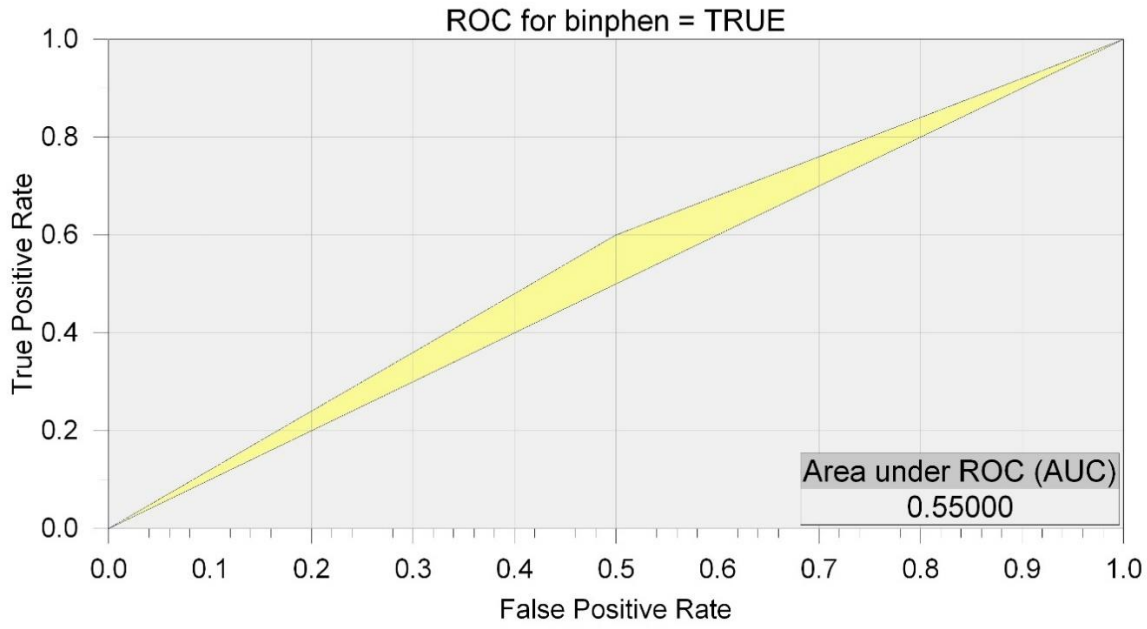


Figure 10. AUC for Genetic Expression Profiling

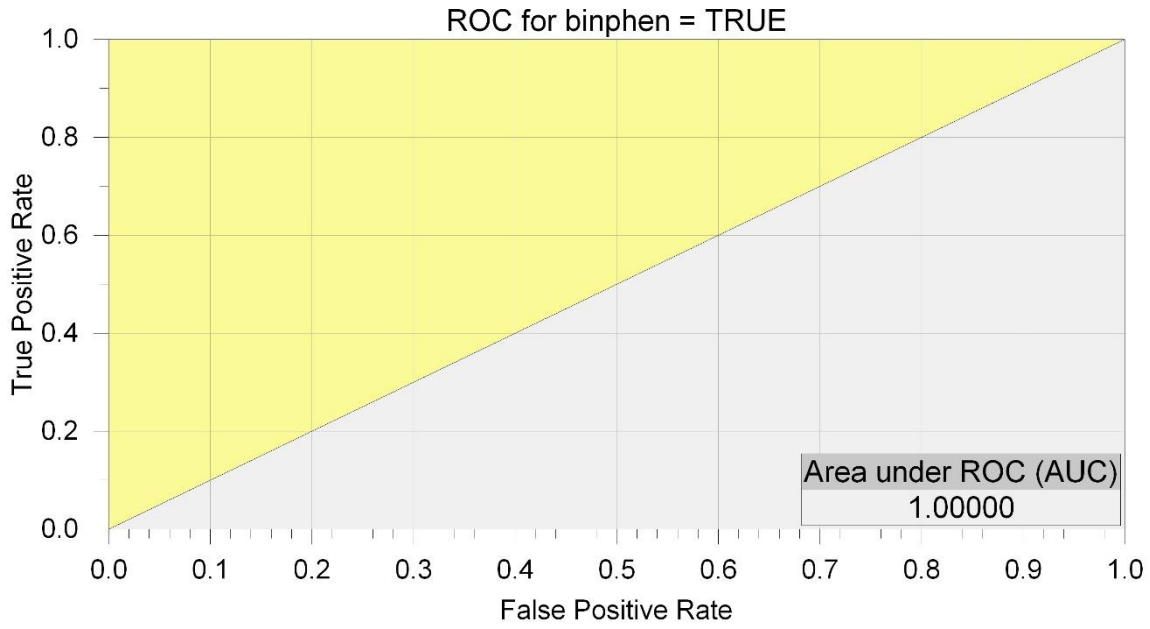


Figure 7.6. AUC for K-means

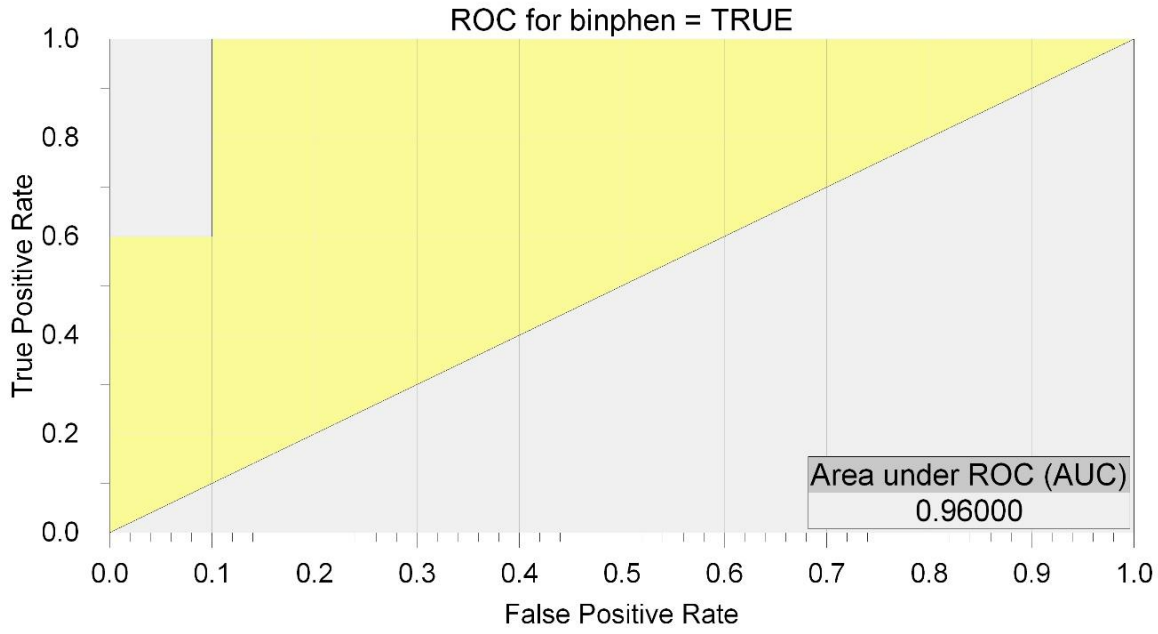


Figure 7.7. AUC for Linear Support Vector Machine

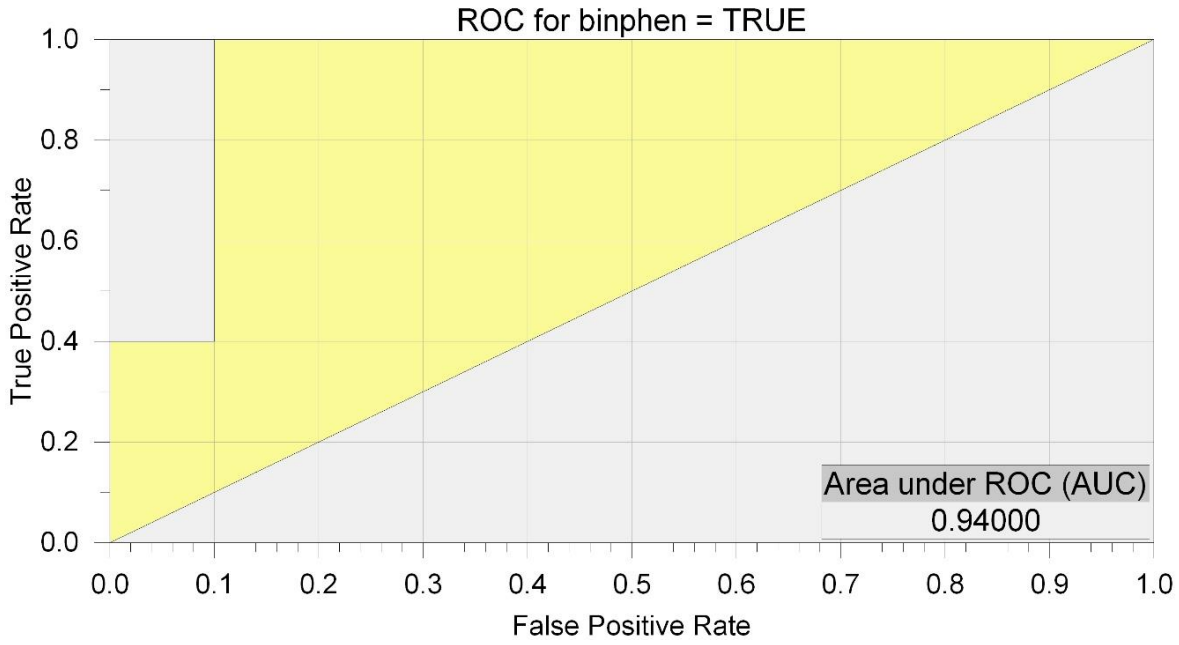


Figure 7.8. AUC for Support Vector Machine with Gaussian Kernel

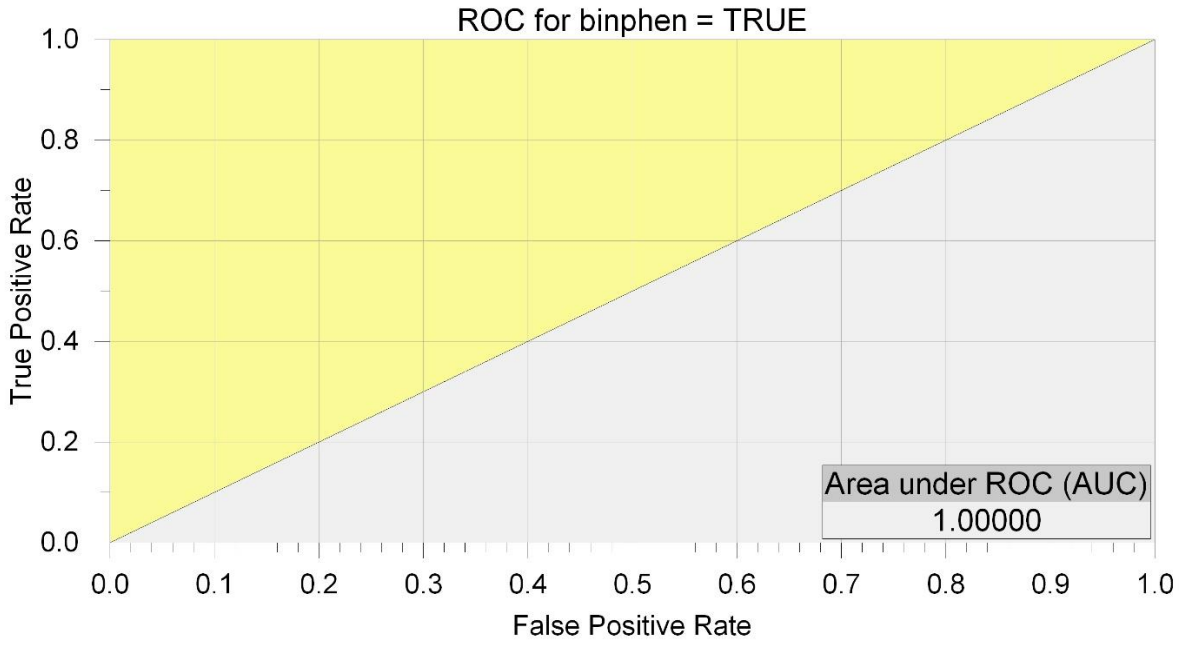


Figure 7.9. AUC for Probabilistic Neural Network

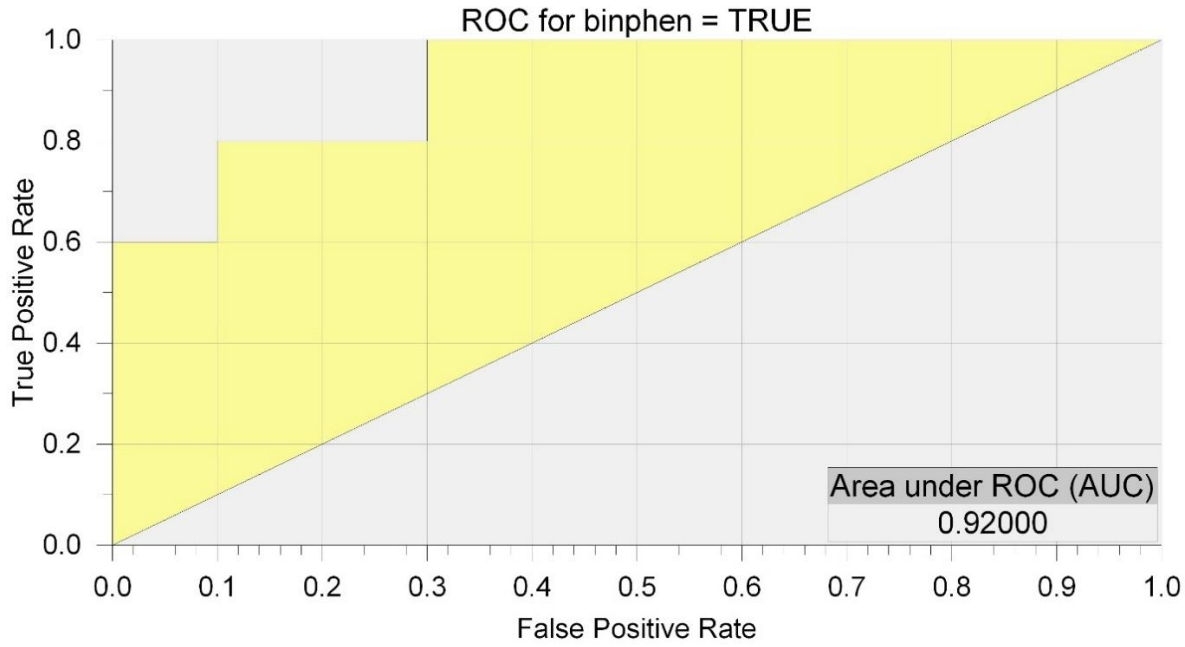


Figure 7.10. Random Forest

Table 7.1. Summary of results of machine learning algorithms to predict sPTB in our cohort.

Algorithm used	**AUC of ROC obtained**
Linear discriminant analysis	0.90
Genetic expression programming	0.70
K-Means	1.00
Linear support vector machine	0.96
Support vector machine with a Gaussian Kernel	0.94
Probabilistic neural network	1.00
Random Forest	0.92

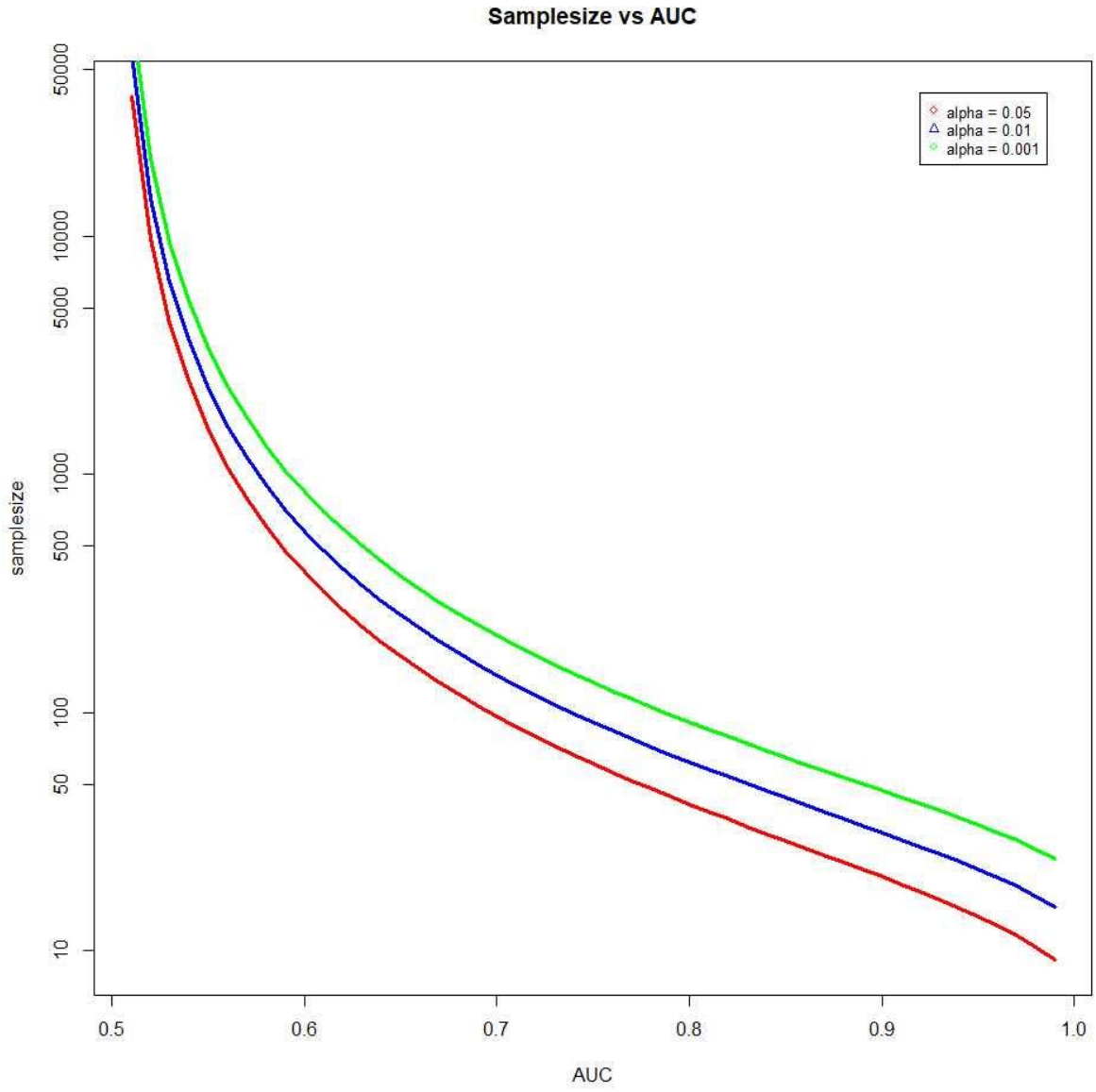


Figure 11. Estimates of sample size across AUC's with significance levels colour coded (at power = 0.8 assuming 80% term controls)

7.4 Discussion

This chapter shows that combination of multiomics from individuals to predict sPTB is possible. Nearly all analysis methods that were tested on our dataset found good levels of prediction except for genetic expression profiling (GEP). GEP only gave an AUC of 0.7, which was much lower than other methods with excellent prediction (≥0.9). These findings clearly need to be validated in another multiomics study.

A challenge for designing multiomic studies is obtaining enough samples to generate enough statistical power, particularly when testing a relatively rare population or condition that is hard to predict, such as sPTB. Simulation tools such as OmicsSIMLA have the ability to combine omics and calculate sample size and power for a new multiomics study (Chung and Kang. 2019). McKeigue (2019) has described a simple method based on a Gaussian approximation for calculating the predictive performance of the learned classifier, given the size of the biomarker panel, the size of the training sample, and the optimal predictive performance of the biomarker panel if a training sample of unlimited size were available, however these are not based on real data. We have provided a more accurate estimate for the purposes of trying to calculate a sample size that would provide enough power to predict sPTB classification.

A criticism of the data would be that we have such small groups in our final analysis. RNA analysis was only collected and performed for a subset of 56 patients in this cohort (of which 29 controls, 5 sPTB, 6 PPROM) which has been the largest limiting factor for numbers in our overall omic analysis. However, following sample storage, extraction and laboratory quality control checks using our methodologies the available data for a three set omic integration (genomics, transcriptomics and

metabolomics) at 16 weeks were limited to only 10 term controls and 5 sPTB cases (4sPTB and 1 PPROM). This means, at this stage, that we cannot clearly answer if omics can differentiate between sPTB and PPROM as we have insufficient sized subgroups. We had to combine the sPTB and PPROM to create a single sPTB group, predominantly represented by sPTB. This is disappointing as the purpose of rigorous classification was to avoid heterogenous groupings as much as possible. However, when compared to other sPTB studies in the literature, it is not uncommon to have such small comparison groups. For example, Chan *et al* (2014) compared RNA sequencing data in the myometrium collected at caesarean section of n=5 sPTB cases and n=5 term births, Gray *et al* (2017) compared miRNA of n=7 sPTB with n=8 term controls, Pereyra *et al* (2019) published a transcriptomic analysis of fetal membranes comparing n=15 term birth cases and n=9 sPTB cases and chapter 2 of this thesis contains many more examples. Not only did these examples study only one type of omic, but for two of these studies listed above the samples were taken at delivery, for which they do not have a gestational age matched control. A clear strength of this study is that our samples were not only taken at the same gestational age timepoints, we also have different layers of omics data from the same individual which we have shown can reduce the number of samples required to power this type of study.

After discussing different methods of combining omics data for further analysis in chapter 2, we chose a concatenation-based model as it's a) is relatively easy to apply statistical tests for categorical data analysis and b) does allow for crosstalk between omics layers via interaction.

The obvious question is, why is it necessary to evaluate so many different statistical approaches? Is there not one single *best* method that can be decided a

priori? As the different methods work better or worse depending on the dataset, there is no one method that dominates all the others on every dataset. (James et al. 2013) With a specific dataset, it is an important task to decide which method produces the best result, but this is not necessarily transferable to a similar but different dataset.

With prediction accuracy and model interpretability there will always be a trade-off between the flexibility of the model to fit the available data points and interpretability. In general, as the flexibility increases the interpretability decreases. (James et al. 2013) Some methods can lead to very complicated estimates of the unknown function of predictor variables. It can become too difficult to work out how individual predictors are associated with outcome. This can be seen clearly with support vector machines with non-linear kernels which are higher in flexibility than trees, but also much harder to interpret. There are many more variable importance's produced from the random forest than the support vector machines (Appendix M). The type of model that we would ultimately choose may depend on how much inference is important once prediction was established.

To try and better understand sPTB, particularly to design new preventative interventions, it is important that we ultimately understand how our dependent variable changes as a result of important predictors. However, the primary purpose of this thesis was first and foremost to establish accurate prediction, therefore the interpretability when choosing models was a secondary consideration. This does not necessarily mean that the most flexible model is either the most accurate or best choice. Highly flexible models can model so closely to the known data points (and errors), that they become poor predictors when attempting to classify new datasets – a concept called '*overfitting*' a model.

7.5 Conclusion

This analysis shows that it is possible to combine multiomics data to predict sPTB. For a future study using multiomics methodology we would expect an AUC of >0.9 for prediction of sPTB. Only a sample size of approximately 50 participants (with 10 sPTB cases) would be required for validation at a power of 0.8. However, in view of the small size of this discovery set, there is a possibility of an overfitted model. In addition, the practical issues in obtaining the samples and sample dropout rate should be taken into consideration when planning a validation study. This recruitment figure would need to be doubled, as a minimum, to allow for a complete omics data set at the analysis. Lastly, this calculation only accounts for the combination of sPTB and PPROM cases as a single group – if further analysis of different subtypes were to be performed this calculation would no longer apply.

Chapter 8: Discussion and Conclusion

8.1 **Addressing Aims**

The primary aim of the thesis was to establish a way of combining three different types of 'omics' analyses used in a pilot study for the prediction of sPTB. Chapters 3, 4, 5 and 6 showed the analysis of the individual layers of omics data including how the participants were divided based on phenotype. Chapter 7 addresses the multi-omic machine learning comparisons that were used for 'prediction' using this pilot cohort group. Overall, this thesis shows that this type of analysis is possible. It may not only be a useful tool to establish predictors of preterm birth but may allow us to evaluate the most important cross-omic biological signals that weight these predictions.

I also aimed to establish if there were distinct differences in biomarkers between PPROM and sPTB subgroups of sPTB within my results. Contrary to my expectations, I did not find many signals suggesting differences between the subgroups, but I feel there is further work to perform in this area. The GWAS data analysis require large sample size to find differences in groups, and for this analysis the groups were kept as large as possible (sPTB and PPROM cases were combined into one larger group) – therefore genetic data could not contribute to finding differences. The transcriptomic data found no differences between sPTB and PPROM groups. After FUMA analysis demonstrated enrichment of the selenoamino acid metabolism pathway in the high-risk preterm birth population, the hierarchical clustering analysis that followed clustered the PPROM cases among the sPTB cases (Figure 5.7). Only metabolomic data showed a suggestion that there might be differences between sPTB and PPROM, but these data were not controlled for diet or time from fasting and there is a high chance that any differences seen are potential false positives until validation testing is performed.

8.2 **Key Findings**

Perhaps the most interesting, clinically relevant and unexpected finding from this work was the support for selenium playing a key role in sPTB. Existing work from a GWAS study (Zheng et al. 2017) had already suggested that the EEFSEC (Selenocysteine elongation factor) gene is associated with sPTB. When we compared odds ratios for a SNP on the EEFSEC gene with published data from the Zheng et al. study, our data agreed that this gene is associated with sPTB <37 weeks. Independently, the findings of our GSEA of the transcriptomic data also suggest a role of selenium in the initiation of sPTB. After adjusting for multiple testing three genes in the selenium pathway were found to be statistically significant; CTH, LCMT1, TRMT11. The machine learning method of analysis we used could not have been influenced by any prior knowledge. Having both omic layers (genomic and transcriptomic) support this data independently is far more powerful than either on its own and suggests this is an important area for future study.

Another important output from this study is the 'feasibility' of performing a multi-omic study with well phenotyped groups. There are several stages at which samples are 'lost' despite good overall recruitment of participants to the study. Patients not attending (n=6) or declining participation (n=2) at follow up occurred. Declining participation occurred at follow up as one woman had experience of a difficult venepuncture at her GP practice and remained bruised. She did not wish to have any further unnecessary blood tests. Another woman could not wait to see the recruitment team following her clinic appointment due to time pressures with childcare. The biggest area to impact recruitment Figures was establishing a clear delivery phenotype, which frequently resulted in exclusion from the study. Implementing a strict phenotype for inclusion meant that the samples available for

analysis are fewer than those recruited. Of 128 women recruited, 41 (32%) women were excluded once delivery phenotypes were available (Table 3.4, page 126). For future multi-omic studies, this data will be useful for setting recruitment targets and remaining realistic about what any single centre can achieve. Collaboration of multiple centres is likely to be necessary to achieve a study size of value in multi-omic work to allow for discovery and validation and allow for subgroup analysis of sPTB and PPROM.

8.3 **Discussion Points**

One of the major strengths of this study is the unique nature of the sPTB population. All the participants had a history of sPTB and therefore this is a study of *recurrent* sPTB versus a *high-risk* control. The advantage of recruiting cases experiencing a recurrent preterm birth is there may be positive selection of aetiologies of sPTB inherent to the mother rather than an individual pregnancy. For example, women with a genetic predisposition to sPTB may be more likely to have two pregnancies affected by sPTB than if infection was the cause, which may only affect a single pregnancy.

Although recruitment was performed in a high-risk population for sPTB < 34 weeks gestation, recruiting both low and other high-risk populations for sPTB were considered. Opening recruitment to the whole obstetric (low risk) population would enable recruitment of women with sPTB/PPROM <34 weeks and additionally women with *only* term deliveries >37 weeks (i.e. no previous preterm births, a low risk control). However, the incidence of preterm birth in a low risk population <37 weeks is approximately 7.1% in the UK (Office of National Statistics 2017). When considering only early sPTB or PPROM between 23 and 34 weeks this figure decreases to approximately 1-2% (Beta et al. 2012). Recruiting for cases would be

difficult in the low risk population as too many participants overall would be required to achieve the same number of cases. Additionally, the design of our study would have changed if we had decided to recruit from the general obstetric population. Scheduled hospital contact for ultrasound scans occur at 12 and 20 weeks, with additional midwifery appointments frequently occurring in the community; therefore 12 and/or 20 weeks would pragmatically have been the best recruiting timepoints at hospital for a low risk population. Compared to recruiting from the preterm birth clinic, difficulties with this method of recruiting include identifying and contacting women for participation in advance of their first scan, particularly as the viability of the pregnancy may not be known. Recruitment would take a long time, and bias may be introduced by the type of women choosing to participate in the research study, potentially leading to fewer preterm births amongst the participants than planned. However, adding a low risk control group to this study in the future is an area to consider and would be feasible if only "healthy" volunteers with a history of term birth are approached. To get samples at the 16- and 20-week gestation timepoints, women would have to be invited to participate and a separate research appointment scheduled. Women could be identified after viability was confirmed at the 12-week scan, there is a singleton gestation and no abnormal or concerning USS features are present. Hospital records could be searched to ensure that only women with previous term pregnancies are approached. Even if only 50% of women approached from this population accept participation in the study, they are all highly likely to have a subsequent term birth and the prevalence of this population is high, therefore recruitment will be faster than the high risk cohort even if take up to the study is low.

There were other risk factors for sPTB/PPROM that I chose to exclude when recruiting. Women with short cervix without a history of sPTB, women who had 'significant' cervical surgery and multiple pregnancies are three of the most significant examples. One of the strengths of this cohort was the homogeneous ("clean") phenotype. Given that our analysis involved so many variables within the omics layers, the more we could minimise the differences in the aetiologies of sPTB, the more the groups may cluster and give us significant findings or new results that may have been missed in more heterogenous cohorts. Additionally, in the UK, screening for short cervical length does not occur routinely at present and therefore recruiting women with short cervix, but no history of sPTB is not feasible for our recruitment setting. Multiple pregnancies are seen frequently throughout pregnancy in the multiple pregnancy clinic (MPC) therefore recruitment would be feasible; however, the aetiology of preterm delivery in twin pregnancies is likely different from that of singletons.

We also considered obtaining samples from large UK pregnancy biobanks. On enquiry, the samples obtained frequently did not have multiple timepoints from the same individual, the samples were not collected in the second trimester, the type of biological fluid we wished to use was not available or there would only be enough sample available for one type of omic analysis. The heterogeneity would have been too large and further highlights the uniqueness and importance of our recruited cohort.

It was challenging to decide how to analyse the group of women who had a treatment for a short cervix, but ultimately delivered at term. It could be argued that these women had "successful" treatment i.e. the treatment prevented another sPTB. To the contrary, these women were always going to have a term birth and the

treatment given was superfluous. The truth is likely to be somewhere in between these two positions; therefore, we excluded these women.

A criticism of our phenotyping is that despite being as thorough as possible, we were dependent on information recorded in clinical notes and made retrospective judgements on groupings. Relevant information may have been missed from the records. In the UK we do not perform specific screening for polyhydramnios or infection in pregnancy. It is possible that there may have been other cases of polyhydramnios where PPROM occurred before the increased amniotic fluid volumes were ever noted. Additionally, women who bled throughout pregnancy may have never reported this symptom or might not have been recorded clearly

We did not see the differences between sPTB and PPROM that we had hypothesised that we might. A possible explanation, and another criticism of this data, is that our groups were too small to see significant differences at an individual omic level. We also did not have enough samples in the multi-omic comparison (chapter 7) to have PPROM as its own group, and therefore differences could not be tested. Another possible explanation is that there are no differences between these groups, i.e. we are not seeing a difference because there isn't one. However, our phenotyping of the PPROM group was strict and we may have excluded cases with an aetiology that would have shown significant differences like infection or polyhydramnios. Without these, sPTB and PPROM are in fact the same when defined by omics signatures. This may also explain why other studies in the literature have suggested there are differences as they were not phenotyped with this degree of precision. To address this, further research with larger groups of sPTB and PPROM should be performed. Other methods of analysis might be better at differentiating

these groups, such as mass spectroscopy which typically identifies more metabolites than NMR.

Of note, our population in this study is predominantly Caucasian (90%), as is typical for a Liverpool population. On one hand this makes our cohort increasingly homogenous, reducing known genetic variation based on ethnicity. On the other hand, this may reduce the generalisability of our results to a wider population.

The issue of the small dataset was addressed in the discussion in Chapter 7 (section 7.4). Despite being an obvious criticism of the data across all 'omic' analysis in this thesis, the evidence in the literature demonstrates how common other studies with comparatively small figures are. These studies are frequently only a single time point, and only investigating one specific omic analysis or panel of biomarkers. This highlights again the strength of this dataset, including the well-designed study methodology, use of multiple omics to increase predictive value and data from multiple timepoints.

RNA was only collected for a subset of 56 patients in this cohort (29 controls, 5 sPTB, 6 PPROM) and this is certainly the largest limiting factor for the *integromics*, as many women were excluded as they didn't have RNA despite samples for genetics and metabolomics being available. Following all other exclusions only 15 women were left for analysis in Chapter 7. This data is essential for planning further omic studies as we would expect approximately a quarter of the women recruited to proceed to the final analysis, based on our estimates.

Should we have collected a different sample for omics analysis such as urine or saliva, or used a more pregnancy specific tissue such as placenta or amniotic fluid instead of blood? I think it is reasonable to use blood as a source for biomarker prediction as it is easily accessible and generally acceptable to pregnant women who

are required to have blood taken at various stages during pregnancy. It is common for genetic studies to use whole blood for DNA extraction, but would gene expression from blood truly reflect a difference between the pregnancy cohorts? Popular belief initially assumed that mammalian erythroid cells, lacking a nucleus, were devoid of mRNA and therefore incapable of protein synthesis, making it a poor fluid for omics analysis, particularly transcriptomic analysis. Microarray studies on mRNA isolated from whole blood demonstrated that up to 70% of the total RNA isolate was actually from haemoglobin/erythrocytes and *not* the leukocyte fraction. (Tian et al. 2009). Blood quickly became a surrogate for tissue-specific RNA. Studies using microarray analysis have shown that blood cells share more than 80% of the transcriptome with each of nine tissues studied (brain, colon, heart, kidney, liver, lung, prostate, spleen and stomach), and estimates are that the blood transcriptome contains 16,000–20,000 transcripts (Liew et al. 2006). Therefore, we felt that this was an appropriate fluid to study. As although it is not pregnancy specific like for example amniotic fluid, it may still detect signals from the uterus/placenta and has the advantage of being safe to collect in the mid-trimester to provide a gestational age matched control.

A criticism of using cross-omics data as a predictive tool is that there could be a long way to go before any positive results of this type of research become translatable to a clinical setting for patient benefit. Omics analysis is unlikely to ever be fast enough for a bedside test. Omic services will remain highly specialised, samples require time and care to process, pass strict quality control thresholds and frequently require shipment to a specialist laboratory which may have a wait time for results that is incompatible with clinical need. Interpretation of omic data requires time and expertise that are not yet widely available and by the time samples are

analysed using the techniques here, a woman would have already delivered. That is without mention of the prohibitive cost of the tests.

A more realistic goal of data integration using advanced omic technologies is to identify key variables that are used by machine learning to classify women based on their outcome. These variables may subsequently be used to create a panel of predictive biomarkers that can be used in the clinic.

Therefore, it would be prudent to move away from machine learning analyses that are very difficult to interpret. One disadvantage of neural networks is the "black box" dilemma. Despite the networks ability to approximate accurate outcomes, there is no way of tracing its decision structure, which then reveals very little insight about the variables that impact the final classification outcome (Theobald. 2017). Even given a dataset and network topology, there can be two neural networks with different weights and same result. This makes the analysis hard. The 'Random Forest' analysis struck the best balance in our dataset between being a good predictor, not being prone to overfitting and remaining interpretable. It is also unaffected by prior knowledge and completely at random generates trees from the predictor variables.

8.4 **Implications for Future Research**

Validation of results and training our models on new cohorts is required to understand the reliability of our results. Although we have used very advanced cross-validation methods on all our machine learning algorithms to ensure accurate model performance, this should be repeated on an independent cohort to ensure the results are the same. Although we have focussed on the accuracy of prediction in this thesis,

a future direction would be to examine model interpretability to assess the key variables leading to prediction.

We have only used three omics in this pilot study. More omic layers may be included in future analysis, particularly omics that are starting to show some promise towards differentiating sPTB from term outcomes. The vaginal microbiome discussed on pages 23 to 25 is a promising area of research and proteomics using electrophoresis techniques or mass spectroscopy may complement the transcriptomics and metabolomics analysis. Mitochondrial RNA might also be a useful analysis in addition to messenger RNA. MiR223, a mitochondrial RNA, has been found to be increased in sPTB when compared to term (Hassan et al. 2015, Sanders et al. 2015, Gray et al. 2017). However, this would also require significantly more funding and may become prohibitively expensive or require too many vials of blood to be taken from the participant. Additionally, for each layer of omic analysis there is the potential to create more "noise" as hundreds, and sometimes thousands, of variables are added

Assessing other groups such as a low risk pregnant cohort would also be useful. Examining omics in women who have not had a history of preterm birth may show a greater discrepancy between some of the most significant variables between cases and controls.

Our initial results are suggestive of a link between selenium and sPTB. The next logical research step is to evaluate the selenium pathway or markers of selenium in our groups. If there is a large difference in selenium concentrations between the sPTB group and the term group this result may suggest that there are possible clinical implications for selenium supplements in high risk women.

8.5 Final Conclusions

This research aimed to combine different types of 'omics' analysis for the prediction of sPTB. Following integration of genomic, transcriptomic and metabolomic data, six out of seven machine learning algorithms to predict sPTB provided excellent prediction, making individualized preterm prediction a realistic possibility and a research area that should be pursued. This research provides valuable insight into planning future omic studies.

A key limiting factor in our study was the small overall number of samples that were included as cases and controls in the final analysis. This did not allow us to investigate the differences between sPTB and PPROM groups and this remains an area for future study. While the sample size limits the generalisability of the results, our results did provide new insights into the role of selenium in the prediction of sPTB and identified the selenium pathway as an important avenue for future study with the potential for clinical impact.

Bibliography

AAGAARD, K., RIEHLE, K., MA, J., SEGATA, N., MISTRETTA, T.A., COARFA, C., SABEEN RAZA, ROSENBAUM, S., VAN DEN VEYVER, I., MILOSAVLJEVIC, A., GEVERS, D., HUTTENHOWER, C., PETROSINO, J., VERSALOVIC, J. (2012). A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLOS ONE*. 7(6):E36466.

ABBOTT, D.S., RADFORD, S.K., SEED, P.T., TRIBE, R.M., SHENNAN, A.H. (2013). Evaluation of a quantitative fetal fibronectin test for spontaneous preterm birth in symptomatic women. *American Journal of Obstetrics and Gynecology*. 208(2):122.E1–122.E6.

ACOG Practice Bulletin No.142. (2014). Cerclage for the management of cervical insufficiency. American College of Obstetricians and Gynecologists. *Obstetrics & Gynecology.* 123:372.

AGGARWAL, R., GATHWALA, G., YADAV, S., KUMAR, P. (2016). Selenium Supplementation for Prevention of Late-Onset Sepsis in Very Low Birth Weight Preterm Neonates. *Journal of Tropical Pediatrics*. 62 (3), 185–193.

AGRAWAL, V., HIRSCH, E. (2012). Intrauterine infection and preterm labor. *Seminars in fetal & neonatal medicine*. 17 (1), 12-9.

AHMAD, I., BASHERI, M., IQBAL, M.J., RAHIM, A. (2018). Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. *IEEE Access.* 6; 33789-33795

ALFIREVIC, Z. & HASSAN, S. S. (2012). Vaginal progesterone in women with an asymptomatic sonographic short cervix in the midtrimester decreases preterm delivery and neonatal morbidity: a systematic review and meta-analysis of individual patient data. *American Journal of Obstetrics and Gynecology*. 206(2), 124 e1-19

ALFIREVIC, Z., STAMPALIJA, T., ROBERTS, D., & JORGENSEN, A.L. (2012). Cervical stitch (cerclage) for preventing preterm birth in singleton pregnancy. *The Cochrane database of systematic reviews*. 18 (4), CD008991.

ALLEMAN, B.W., MYKING, S., RYCKMAN, K.K., MYHRE, R.,FEINGOLD, E.,FEENSTRA, B.,GELLER, F.,BOYD, H.A.,SHAFFER, J.R.,ZHANG, Q.,BEGUM, F.,CROSSLIN, D.,DOHENY, K.,PUGH, E., PAY, A.S.,OSTENSEN, H.,MORKEN, N.H.,MAGNUS, P.,MARAZITA, M.L.,JACOBSSON B,MELBYE M,MURRAY JC. GENE, ENVIRONMENT ASSOCIATION STUDIES (GENEVA) CONSORTIUM; NORWEGIAN MOTHER AND CHILD COHORT STUDY (MOBA) GENOME-WIDE ASSOCIATION STUDY GROUP. (2012). No observed association for mitochondrial SNPs with preterm delivery and related outcomes. *Pediatric Research.* 72(5):539-44.

ALONSO, A., MARSAL, S., JULIÀ, A. (2015). Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in Bioengeering and Biotechnology.* 3:23.

ALTHUISIUS, S.M., DEKKER, G.A., VAN GEIJN, H.P., HUMMEL, P. (1999). The effect of therapeutic McDonald cerclage on cervical length as assessed by transvaginal ultrasonography. *American Journal of Obstetrics and Gynecology.* 180(2 Pt 1), 366-9.

ALTHUISIUS, S.M., DEKKER, G.A., HUMMEL, P., VAN GEIJN, H.P. (2003). Cervical incompetence prevention randomized cerclage trial: emergency cerclage with bed rest versus bed rest alone. *American Journal of Obstetrics and Gynecology*. 189(4):907-10.

ALTMANN, A., TOLOŞI, L., SANDER, O., LENGAUER, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–134.

ANDERSON, C.A., PETTERSON, F.H., CLARKE, G.M., CARDON, L.R., MORRIS, A.P., ZONDERVAN., K.T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols.* 5(9): 1564-73.

ANDRAWEERA, P.H., DEKKER, G.A., THOMPSON, S.D., NORTH, R.A., MCCOWAN, L.M., ROBERTS, C.T., SCOPE CONSORTIUM. (2012). The interaction between the maternal BMI and angiogenic gene polymorphisms associates with the risk of spontaneous preterm birth. Molecular Human Reproduction. 18(9):459–65.

ANNUNZIATO, A. (2008) DNA Packaging: Nucleosomes and Chromatin. *Nature Education* 1(1):26

ARMITAGE E.G & BARBAS, C. (2014). Metabolomics in cancer biomarker discovery: current trends and future perspectives. *Journal of Pharmaceutical and Biomedical Analysis*. 87:1-11.

BACK, S.A., RIDDLE, A., MCCLURE, M.M. (2007). Maturation dependent vulnerability of perinatal white matter in premature birth. *Stroke*. 38(2 suppl): 724-30.

BARTHA, J.L., FERNANDEZ-DEUDERO, A., BUGATTO, F., FAJARDO-EXPOSITO, M.A., GONZALEZ-GONZALEZ, N., HERVIAS-VIVANCOS, B. (2012). Inflammation and cardiovascular risk in women with preterm labor. *Journal Women's Health.* 21 (6), 643-8.

BARALDI, E., GIORDANO, G., STOCCHERO, M., MOSCHINO, L., ZARAMELLA, P., TRAN, M.R., CARRARO, S., ROMERO, R., GERVASI, M.T. (2016). Untargeted Metabolomic Analysis of Amniotic Fluid in the Prediction of Preterm Delivery and Bronchopulmonary Dysplasia. *PLoS One.* 11(10); e0164211.

BENJAMINI, Y. HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.* 57 (1): 289–300.

BENNETT, P. (2007). Preterm Labour. In: Edmonds D.K., editor. *Dewhurst's Textbook of Obstetrics & Gynaecology*: Blackwell Publishing. p. 177-91.

BERG, J.M., TYMOCZKO, J.L., STRYER, L. (2002) Biochemistry. 5th edition. Section 30.3, Food Intake and Starvation Induce Metabolic Changes. New York: W H Freeman. Available from: https://www.ncbi.nlm.nih.gov/books/NBK22414/

BETA, J., ISSAT, T., NOWICKA, M.A., ANDZIAK, M., JAKIMIUK, A.J. (2012). Early spontaneous preterm deliveries before 34 weeks' gestation in a tertiary care centre: analysis of maternal factors and obstetric history. *The Journal of Maternal-Fetal and Neonatal Medicine.* 76:720-723.

BITNER, A., SOBALA, W., KALINKA, J. (2013). Association between maternal and fetal TLR4 (896A > G, 1196C > T) gene polymorphisms and the risk of pre-term birth in the Polish population. *American Journal of Reproductive Immunology*. 69:272–80.

BLAND, J.M., ALTMAN, G.D. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal.* 310:170.

BLENCOWE, H., COUSENS, S., CHOU, D., OESTERGAARD, M., SAY, L., MOLLER, A., KINNEY, M., LAWN, J. (2013). Born too soon: the global epidemiology of 15 million preterm births. *Reproductive Health*. 10, Suppl 1:S2.

BOOTS, A.B., SANCHEZ-RAMOS, L., BOWERS, D.M., KAUNITZ, A.M., ZAMORA, J., SCHLATTMANN, P. (2014). The short-term prediction of preterm birth: a systematic review and diagnostic metaanalysis. *American Journal of Obstetrics and Gynecology.* 210 (1), 54.e1-e10.

BOPEGAMAGE, S., KACEROVSKY, M., TAMBOR, V., MUSILOVA, I., SARMIROVA, S., SNELDERS, E., DE JONG, A.S., VARI, S.G., WILLEM, J., MELCHERS, G., GALAMA, J.M.D. (2013) Preterm prelabor rupture of membranes (PPROM) is not associated with presence of viral genomes in the amniotic fluid. Journal of Clinical Virology. 58(3);559-563.

BOTHWELL J.H. & GRIFFIN, J.L. (2011). An introduction to biological nuclear magnetic resonance spectroscopy. *Biological Reviews of the Cambridge Philosophical Society*. 86 (2):493-510.

BOYD, H.A., POULSEN, G., WOHLFAHRT, J., MURRAY, J.C., FEENSTRA, B., MELBYE, M. (2009). Maternal contributions to preterm delivery. *American Journal of Epidemiology.* 170 (11), 1358-64.

BOYLE, E.M., POULSEN, G., FIELD, D.J., KURINCZUK, J.J., WOLKE, D., ALFIREVIC, Z., QUIGLEY, M.A. (2012). Effects of gestational age at birth on health outcomes at 3 and 5 years of age: population-based cohort study. *British Medical Journal*. 344, e896.

BRADLEY, P.S. AND MANGASARIAN, O.L. (1998) Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Machine Learning Proceedings of the Fifteenth International Conference* (ICML '98), pages 82–90, San Francisco, California.

BREAM, E.N., LEPPELLERE, C.R., COOPER, M.E., DAGLE, J.M., MERRILL, D.C., CHRISTENSEN, K., SIMHAN, H.N., FONG, C.T., HALLMAN, M., MUGLIA, L.J., MARAZITA, M.L., MURRAY, J.C. (2013). Candidate gene linkage approach to identify DNA variants that predispose to preterm birth. *Pediatric Research*. 73:135–41.

BREIMAN, L. (2001). Random forests. *Machine Learning.* 45(1):5.

BROCKLEHURST, P., GORDON, A., HEATLEY, E., MILAN, S.J. (2013). Antibiotics for treating bacterial vaginosis in pregnancy. *Cochrane Database of Systematic Reviews*. 1465-1858

BROWN R.G., MARCHESI, J.R., LEE, Y.S., SMITH, A., LEHNE, B., KINDINGER, L.M., TERZIDOU, V., HOLMES, E., NICHOLSON, J.K., BENNETT, P.R. AND MACINTYRE, D.A. (2018). Vaginal dysbiosis increases risk of preterm fetal membrane rupture, neonatal sepsis and is exacerbated by erythromycin. *BMC Medicine.* 16 (9); 1-15.

BRUBAKER, D. LIU, Y. WANG, J., TAN, H., ZHANG, G., JACOBSSON, B., MUGLIA, L., MESIANO, S., CHANCE M.R. (2016). Finding lost genes in GWAS via integrative-omics analysis reveals novel sub-networks associated with preterm birth. *Human Molecular Genetics.* 25 (23): 5254-5264.

BUKOWSKI, R., HANKINS, G.D., SAADE, G.R., ANDERSON, G.D., THORNTON, S. (2006). Labor-associated gene expression in the human uterine fundus, lower segment, and cervix. *PLOS Medicine*. 3 (6): e169.

BUKOWSKI, R., SADOVSKY, Y., GOODARZI, H., ZHANG, H., BIGGIO, J.R., VARNER, M., PARRY, S., XIAO, F., ESPLIN, S.M., ANDREWS, W., SAADE, G.R., ILEKIS, J.V., REDDY, U.M., BALDWIN, D.A. (2017). Onset of human preterm and term birth is related to unique inflammatory transcriptome profiles at the maternal fetal interface. *The Journal of Life and Environmental Sciences.* 5:e3685.

CADENAS, E. & DAVIES, K.J. (2000). Mitochondrial free radical generation, oxidative stress, and aging. *Free Radical Biology and Medicine*. 29:222–230.

CALLAHAN, B.J., DIGIULIO, D.B., GOLTSMAN, D.S.A., SUN, C.L., COSTELLO, E.K., JEGANATHAN, P., BIGGIO, J.R., WONG, R.J., DRUZIN, M.L., SHAW, G.M., STEVENSON, D.K., HOLMES, S.P., RELMAN, D.A. (2017). Replication and refinement of a vaginal

microbial signature of preterm birth in two racially distinct cohorts of US women. *Proceedings of the National Academy of Sciences U S A* 114: 9966e71.

CANNIE, M.M., DOBRESCU, O., GUCCIARDO, L., STRIZEK, B., ZIANE, S., SAKKAS, E., SCHOONJANS, F. DIVANO, L., JANI, J. C. (2013). Arabin cervical pessary in women at high risk of preterm birth: a magnetic resonance imaging observational follow-up study. *Ultrasound in Obstetrics and Gynecology*. 42 (4), 426-33.

CAPECE, A., VASIEVA, O., MEHER, S., ALFIREVIC Z & ALFIREVIC A. (2014). Pathway analysis of genetic factors associated with spontaneous preterm birth and pre-labor preterm rupture of membranes. *PloS One*. 9 (9), e108578.

CAPPELLETTI, M., PRESICCE, P., LAWSON, M.J., CHATURVEDI, V., STANKIEWICZ, T.E., VANONI, S., HARLEY, I.T., MCALEES, J.W., GILES, D.A., MORENO-FERNANDEZ, M.E., RUEDA, C.M., SENTHAMARAIKANNAN, P., SUN, X., KARNS, R., HOEBE, K., JANSSEN, E.M., KARP, C.L., HILDEMAN, D.A., HOGAN, S.P., KALLAPUR, S.G., CHOUGNET, C.A., WAY, S.S., DIVANOVIC, S. (2017) Type I interferons regulate susceptibility to inflammation-induced preterm birth. *JCI Insight*. 2 (5), e91288.

CARE, A.G., SHARP, A.N., LANE, S., ROBERTS, D., WATKINS, L., ALFIREVIC, Z. (2014). Predicting preterm birth in women with previous preterm birth and cervical length >/= 25 mm. *Ultrasound in Obstetrics and Gynecology*. 43 (6), 681-6.

CARE, A., INGLEBY, L., ALFIREVIC, Z., SHARP, A. (2019). The influence of the introduction of national guidelines on preterm birth prevention practice: UK experience. *BJOG: An International Journal of Obstetrics and Gynaecology*. 126(6):763-769.

CARTER, J, TRIBE, R.M., WATSON, H.A., SHENNAN, A.H. (2016) Threatened preterm labour management: results of Delphi consensus on best practice: PL 37 *British Journal of Obstetrics & Gynaecology.* 123; 100-101.

CARTER, J., SEED, P.T., WATSON, H.A., DAVID, A.L., SANDALL, J., SHENNAN, A.H., TRIBE, R.M. (2020) Development and validation of prediction models for the QUiPP App v.2: a tool for predicting preterm birth in women with symptoms of threatened preterm labor. *Ultrasound in Obstetrics and Gynecology.* 55*;* 357-367.

CHAIX, M.A., KOOPMANN, T.T., GOYETTE, P., ALIKASHANI, A., LATOUR, F., FATAH, M., HAMILTON, R.M., RIOUX, J.D. (2016) CALM3 mutations in pediatric long QT syndrome patients support a CALM-3 specific calmodulinopathy. *Heart Rhythm Case Reports*. 2(3):250-254.

CHAN, Y.W., VAN DEN BERG, H.A., MOORE, J.D., QUENBY, S., BLANKS, A.M. (2014). Assessment of myometrial transcriptome changes associated with spontaneous human labour by high-throughput RNA-seq. *Experimental Physiology*. 99(3):510-524.

CHANDIRAMANI, M., SEED, P. T., ORSI, N. M., EKBOTE, U. V., BENNETT, P. R., SHENNAN, A. H., & TRIBE, R. M. (2012) Limited relationship between cervico-vaginal fluid cytokine profiles and cervical shortening in women at high risk of spontaneous preterm birth. *PloS one*, *7*(12), e52412.

CHAWANPAIBOON, S., VOGEL, J.P., MOLLER, A-B., LUMBIGANON, P., PETZOLD, M., HOGAN, D., LANDOULSI, S., JAMPATHONG, N., KONGWATTANAKUL, K., LAOPAIBOON, M., LEWIS, C., RATTANAKANOKCHAI, S., TENG, D.N., THINKHAMROP, J., WATANANIRUN, K., ZHANG, J., ZHOU, W., GÜLMEZOGLU, A.M. (2019) Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Health.* 7(1): PE37-E46.

CHAVES, J.H., BABAYAN, A., BEZERRA CDE M., LINHARES, I.M., WITKIN, S.S. (2008). Maternal and neonatal interleukin-1 receptor antagonist genotype and pregnancy outcome in a population with a high rate of pre-term birth. *American Journal of Reproductive Immunology*. 60:312–7.

CHUNG, K., KANG, C. Y. (2019). A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. *Gigascience.* 1;8(5).

CHRISTENSEN, K.E., DAHHOU, M., KRAMER, M.S., ROZEN, R. (2014). The MTHFD1 1958G > A variant is associated with elevated C-reactive protein and body mass index in Canadian women from a premature birth cohort. *Molecular Genetics and Metabolism.* 111(3):390–2.

CHUN, S., PLUNKETT, J., TERAMO, K., MUGLIA, L.J., FAY, J.C. (2013). Fine-mapping an association of FSHR with preterm birth in a Finnish population. *PLoS One.* 8(10):e78032.

CLAUSSON, B., LICHTENSTEIN, P. & CNATTINGIUS, S. (2000). Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG : an International Journal of Obstetrics and Gynaecology*. 107(3), 375-81.

CORTES C, VAPNIK, V N. (1995). Support-vector networks. Machine Learning. 20 (3): 273–297.

COSTELOE, K.L., HENNESSY, E.M., HAIDER, S., STACEY, F., MARLOW, N., DRAPER, E.S. (2012). Short term outcomes after extreme preterm birth in England: comparison of two birth cohorts in 1995 and 2006 (the EPICure studies). *British Medical Journal.* 345, e7976.

CRAWFORD, N., PRENDERGAST, D., OEHLERT, J.W., SHAW, G.M., STEVENSON, D.K., RAPPAPORT, N., SIROTA, M., TISHKOFF, S.A., SONDHEIMER, N. (2018) Divergent Patterns of Mitochondrial and Nuclear Ancestry Are Associated with the Risk for Preterm Birth. *Journal of Pediatrics.* 194, 40-46.e4.

DAS, S., FORER, L., SCHÖNHERR, S., SIDORE, C., LOCKE, A.E., KWONG, A., VRIEZE, S., CHEW, E.Y., LEVY, S., MCGUE, M., SCHLESSINGER, D., STAMBOLIAN, D., LOH, P.R., IACONO, W.G., SWAROOP, A., SCOTT, L.J., CUCCA, F., KRONENBERG, F., BOEHNKE, M., ABECASIS, G.R., FUCHSBERGER, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*. 48, 1284–1287.

DE BAERE, I., DERUA, R., JANSSENS, V., VAN HOOF, C., WAELKENS, E., MERLEVEDE, W., GORIS, J. (1999). Purification of porcine brain protein phosphatase 2A leucine carboxyl methyltransferase and cloning of the human homologue. *Biochemistry* 38: 16539-16547.

DEFRANCO, E.A., LEWIS, D.F. & ODIBO, A.O. (2013). Improving the screening accuracy for preterm labor: is the combination of fetal fibronectin and cervical length in symptomatic patients a useful predictor of preterm birth? A systematic review. *American Journal of Obstetrics and Gynecology.* 208(3), 233.e1-6.

DELANEAU, O., MARCHINI, J. & ZAGURY, J.F. (2011). A linear complexity phasing method for thousands of genomes. *Nature Methods*. 9(2): 179–181.

DEVI, S.G., KUMAR, A., KAR, P., HUSAIN, S.A., SHARMA, S. (2014). Association of pregnancy outcome with cytokine gene polymorphisms in HEV infection during pregnancy. *Journal of Medical Virology.* 86:1366–76.

DIETERLE, F., ROSS, A., SCHLOTTERBECK, G., SENN, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical chemistry.* 78(13):4281-90.

DIGIULIO, D.B., CALLAHAN, B.J., MCMURDIE, P.J., COSTELLO, E.K., LYELL, D.J., ROBACZEWSKA, A., SUN, C.L., GOLTSMAN, D.S., WONG, R.J., SHAW, G., STEVENSON, D.K., HOLMES, S.P., RELMAN, D.A. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences U S A.* 112:11060-5.

DOBBIN, K.K. & SONG, X. (2013). Sample size requirements for training high-dimensional risk predictors. *Biostatistics*. 14(4):639–652.

DODD, J.M., JONES, L., FLENADY, V., CINCOTTA, R., CROWTHER, C.A. (2013). Prenatal administration of progesterone for preventing preterm birth in women considered to be at risk of preterm birth. *The Cochrane database of systematic reviews*. (7):CD004947.

DOLAN, S.M., CHRISTIAENS, I. (2013). Genome-wide association studies in preterm birth: implications for the practicing obstetrician-gynaecologst. *BMC Pregnancy and Childbirth.* 13 (S1):S4.

DONA, A.C., KYRIAKIDES, M., SCOTT, F., SHEPHARD, E.A., VARSHAVI, D., VESELKOV, K., & EVERETT, J. R. (2016). A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Computational and Structural Biotechnology Journal*, *14*, 135–153.

DREWS-PIASECKA, E., SEREMAK-MROZIKIEWICZ, A., BARLIK, M., KURZAWIŃSKA, G., WOLSKI, H., WOYCIECHOWSKA, A., CZERNY, B., DREWS, K. (2014). The significance of TNF-alpha gene polymorphisms in preterm delivery. *Ginekologia Polska*. 85:428–34.

DUNDAR, B., DINCGEZ CAKMAK, B., OZGEN, G., TASGOZ, F.N., GUCLU, T. & OCAKOGLU, G. (2018). Platelet indices in preterm premature rupture of membranes and their relation with adverse neonatal outcomes. *The Journal of Obstetrics and Gynaecology Research*. 44, 67-73.

EHN, N.L., COOPER, M.E., ORR, K., SHI, M., JOHNSON, M.K., CAPRAU, D., DAGLE, J., STEFFEN, K., JOHNSON, K., MARAZITA, M.L., MERRILL, D., MURRAY, J.C. (2007). Evaluation of fetal and maternal genetic variation in the progesterone receptor gene for contributions to preterm birth. *Pediatric Research.* 62:630–35.

EIDEM, H.R. ACKERMAN, W.E. 4th, McGARY, K.L., ABBOT, P., ROKAS, A. (2015). Gestational tissue transcriptomics in term and preterm human pregnancies: a systematic review and meta-analysis. *BMC Medical Genomics.* 5(8): 27.

EIDEM, H.R., RINKER, D.C., ACKERMAN, W.E. 4TH, BUHIMSCHI, I.A., BUHIMSCHI, C.S., DUNN-FLETCHER, C., KALLAPUR, S.G., PAVLIČEV, M., MUGLIA, L.J., ABBOT, P., ROKAS, A. (2016). Comparing human and macaque placental transcriptomes to disentangle preterm birth pathology from gestational age effects. *Placenta.* 41:74-82.

ELISIA, I., TSOPMO, A., FRIEL, J.K., DIEHL-JONES, W., KITTS, D.D. (2011). Tryptophan from human milk induces oxidative stress and upregulates the Nrf-2-mediated stress response in human intestinal cell lines. *Journal of Nutrition.* 141(8):1417-23.

ELOVITZ, M.A., BROWN, A.G., ANTON, L., GILSTROP, M., HEISER, L., BASTEK, J. (2014). Distinct cervical microRNA profiles are present in women destined to have a preterm birth. *American Journal of Obstetrics and Gynecology.* 210(3):221.e1-11.

ELOVITZ, M.A., ANTON, L., BASTEK, J., BROWN, A.G. (2015). Can microRNA profiling in maternal blood identify women at risk for preterm birth? *American Journal of Obstetrics and Gynecology*. 212:782.e1-5.

ENGEL, S.M., OLSHAN, A.F., SIEGA-RIZ, A.M., SAVITZ, D.A,. CHANOCK, S.J. (2006). Polymorphisms in folate metabolizing genes and risk for spontaneous preterm and small- for-gestational age birth. *American Journal of Obstetrics and Gynecology*. 195:1231. e1–11.

ENQUOBAHRIE, D.A., WILLIAMS, M.A., QIU, C., MUHIE, S.Y., SLENTZ-KESLER, K., GE, Z., SORENSON, T. (2009). Early pregnancy peripheral blood gene expression and risk of preterm delivery: a nested case control study. *BMC Pregnancy Childbirth*. 9:56.

FALAH, N., MCELROY, J., SNEGOVSKIKH, V., LOCKWOOD, C.J., NORWITZ, E., MURRAY, J.C. KUCZYNSKI, E., MENON, R., TERAMO, K., MUGLIA, L.J., MORGAN, T. (2013). Investigation of genetic risk factors for chronic adult diseases for association with preterm birth. *Human Genetics.* 132(1):57–67.

FARON, G., BALEPA, L., PARRA, J., FILS, J.F., GUCCIARDO, L. (2018). The fetal fibronectin test: 25 years after its development, what is the evidence regarding its clinical utility? A systematic review and meta-analysis. *The Journal of Maternal-Fetal & Neonatal Medicine*. 9:1-31.

FERREIRA, C. (2001). Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. *Complex Systems*. 13(2): 87-129.

FETTWEIS, J.M., BROOKS, J.P., SERRANO, M.G., SHETH, N.U., GIRERD, P.H., EDWARDS, D.J., STRAUSS, J.F., THE VAGINAL MICROBIOME CONSORTIUM., JEFFERSON, K.K., BUCK, G.A. (2014). Differences in vaginal microbiome in African American women versus women of European ancestry. *Microbiology*. 160(Pt 10):2272-2282.

FISHER, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*. 7 (2): 179–188.

FLENADY, V., HAWLEY, G., STOCK, O.M., KENYON, S., BADAWI, N. (2013). Prophylactic antibiotics for inhibiting preterm labour with intact membranes. *Cochrane Database of Systematic Reviews.* 5;(12):CD000246.

FONSECA, E.B., CELIK, E., PARRA, M., SINGH, M., NICOLAIDES K.H. (2007). Progesterone and the risk of preterm birth among women with a short cervix. *The New England Journal of Medicine*. 357(5), 462-9.

FRYDMAN, R., LELAIDIER, C., BATON-SAINT-MLEUX, C., FERNANDEZ, H., VIAL, M., BOURGET, P. (1992). Labor induction in women at term with mifepristone (RU 486): a double-blind, randomized, placebo-controlled study. *Obstetrics and Gynecology.* 80 (6), 972-5.

GARGANO, J.W., HOLZMAN, C.B., SENAGORE, P.K., REUSS, M.L., PATHAK, D.R., FRIDERICI, K.H., JERNIGAN, K., FISHER, R. (2009). Polymorphisms in thrombophilia and renin-angiotensin system pathways, preterm delivery, and evidence of placental hemorrhage. *American Journal of Obstetrics and Gynecology.* 2009;201:317. e1–9.

GHAEMI, M.S., DIGIULIO, D.B., CONTREPOIS, K., CALLAHAN, B., NGO, T.T.M., LEE-MCMULLEN, B., LEHALLIER, B., ROBACZEWSKA, A., MCILWAIN, D., ROSENBERG-

HASSON, Y., WONG, R.J., QUAINTANCE, C., CULOS, A., STANLEY, N., TANADA, A., TSAI, A., GAUDILLIERE, D., GANIO, E., HAN, X., ANDO, K., MCNEIL, L., TINGLE, M., WISE, P., MARIC, I., SIROTA, M., WYSS-CORAY, T., WINN, V.D., DRUZIN, M.L., GIBBS, R., DARMSTADT, G.L., LEWIS, D.B., PARTOVI NIA, V., AGARD, B., TIBSHIRANI, R., NOLAN, G., SNYDER, M.P., RELMAN, D.A., QUAKE, S.R., SHAW, G.M., STEVENSON, D.K., ANGST, M.S., GAUDILLIERE, B., AGHAEEPOUR, N. (2019). Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. *Bioinformatics*, **35**(1), 95–103.

GIBSON, C.S., MACLENNAN, A.H., DEKKER, G.A., GOLDWATER, P.N., DAMBROSIA, J.M., MUNROE, D.J., TSANG, S., STEWART, C., NELSON, K.B. (2007). Genetic polymorphisms and spontaneous preterm birth. *Obstetrics and Gynecology*. 109(2 Pt1):384–91.

GILLIS, J.C., GOA, K.L. (1995). Tretinoin. A review of its pharmacodynamic and pharmacokinetic properties and use in the management of acute promyelocytic leukaemia. *Drugs.* 50(5):897-923.

GOLDENBERG, R.L., MERCER, B.M., MEIS, P.J., COPPER, R.L., DAS, A., MCNELLIS, D. (1996). The preterm prediction study: fetal fibronectin testing and spontaneous preterm birth. NICHD Maternal Fetal Medicine Units Network. *Obstetrics and Gynecology.* 87(5 Pt 1):643-8.

GOLDENBERG, R.L., CULHANE, J.F., IAMS, J.D., ROMERO R. (2008). Epidemiology and causes of preterm birth*. Lancet*, 371(9606), 75-84.

GOLDENBERG, R.L., GRAVATT, M.G., IAMS, J., PAPAGEORGHIOU, A.T., WALLER., S.A., KRAMER, M., CULHANE, J., BARROS, F., CONDE-AGUDELO, A., BHUTTA, Z.A. KNIGHT, H.E., VILLAR, J. (2012). The preterm birth syndrome: issues to consider in creating a classification system. *American Journal of Obstetrics and Gynaecology.* 206: 113-118.

GOMEZ, R., ROMERO, R., MEDINA, L., NIEN, J.K., CHAIWORAPONGSA, T., CARSTENS, M. (2005). Cervicovaginal fibronectin improves the prediction of preterm delivery based on sonographic cervical length in patients with preterm uterine contractions and intact membranes. *American journal of Obstetrics and Gynecology*. 192(2), 350-9.

GOMEZ, L.M., SAMMEL, M.D., APPLEBY, D.H., ELOVITZ, M.A., BALDWIN, D.A., JEFFCOAT, M.K., MACONES, G.A., PARRY, S. (2010). Evidence of a gene-environment interaction that predisposes to spontaneous preterm birth: a role for asymptomatic bacterial vaginosis and DNA variants in genes that control the inflammatory response. *American Journal of Obstetrics and Gynecology*. 202(4);386, e1-6.

GONCALVES, L.F., CHAIWORAPONGSA, T. & ROMERO, R. (2002). Intrauterine infection and prematurity. *Mental Retardation and Developmental Disabilities Research Reviews*. 8(1), 3-13.

GOODFELLOW, L., CARE, A., SHARP, A., IVANDIC, J., POLJAK, B., ROBERTS, D., ALFIREVIC, Z. (2019) Effect of QUiPP prediction algorithm on treatment decisions in women with a previous preterm birth: a prospective cohort study**.** *BJOG : an International Journal of Obstetrics and Gynaecology*. 126: 1569– 75.

GOYA, M., PRATCORONA, L., MERCED, C., RODO, C., VALLE, L., ROMERO, A., JUAN, M., RODRIGUEZ, A., MUNOZ, B., SANTACRUZ, B., BELLO-MUNOZ, J.C., LLURBA, E., HIGUERAS, T., CABERO, L., CARRERAS, E. (2012) Cervical pessary in pregnant women with a short cervix (PECEP): an open-label randomised controlled trial. *Lancet.* 379: 1800–1806.

GRAVETT, M.G., RUBENS, C.E., NUNES, T.M., GAPPS REVIEW GROUP. (2010). Global Report on Preterm Birth and Stillbirth (2 of 7): discovery science. *BMC Pregnancy and Childbirth.* 10: S2

GRAY, C., MCCOWAN, L.M., PATEL, R., TAYLOR, R.S., VICKERS, M.H. (2017) Maternal plasma miRNAs as biomarkers during mid-pregnancy to predict later spontaneous preterm birth: a pilot study. *Scientific Reports.1*7(1):815.

GRISARU-GRANOVSKY, S., ALTARESCU, G., FINCI, S., WEINTRAUB, A., TEVET, A., SAMUELOFF, A. (2010). Prostanoid DP receptor (PTGDR) variants in mothers with post-coital associated preterm births: preliminary observations. *Journal of Perinatology.* 30:33–7.

GROMSKI R.S., MUHAMADALI, H., ELLIS, D.I., XU, Y., CORREA, E., TURNER, M., GOODACRE, R. (2015). A tutorial review: metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*. 879:10-23.

GUI, J., CHI, Y., ZHANG, Q., & BAO, X. (2017). A probability distribution kernel based on whitening transformation. *AMSE JOURNALS-AMSE IIETA publication-2017-Series: Advances B*; 60(1);93-109.

GUOYANG, L., MORGAN, T., BAHTIYAR, M.O., SNEGOVSKIKH, V.V., SCHATZ, F., KUCZYNSKI, E., FUNAI, E.F., DULAY, A.T., HUANG, S.T., BUHIMSCHI, C.S., BUHIMSCHI, I.A., FORTUNATO, S.J., MENON, R., LOCKWOOD, C.J., NORWITZ, E.R. (2008). Single nucleotide polymorphisms in the human progesterone receptor gene and spontaneous preterm birth. *Reproductive Sciences*. 15:147–55.

GUY, A., SEATON, S.E., BOYLE, E.M., DRAPER, E.S., FIELD, D.J., MANKTELOW, B.N., MARLOW, N., SMITH, L.K., JOHNSON, S. (2015). Infants born late/moderately preterm are at increased risk for a positive autism screen at 2 years of age. *The Journal of Pediatrics*. 166 (2); 269-75 e3.

HAATAJA, R., KARJALAINEN, M.K., LUUKKONEN, A., TERAMO, K., PUTTONEN, H., OJANIEMI, M., VARILO, T., CHAUDHARI, B.P., PLUNKETT, J., MURRAY, J.C.,

MCCARROLL, S.A., PELTONEN, L., MUGLIA, L.J., PALOTIE, A., HALLMAN, M. (2011). Mapping a new spontaneous preterm birth susceptibility gene, IGF1R, usink linkage, haplotype sharing, and association analysis. *PloS Genetics*. 7(2):e1001293

HACKSTADT, A.J. & HESS, A.M. (2009). Filtering for increased power for microarray data analysis. *BMC Bioinformatics.* 10, 11.

HADLEY, E.E., DISCACCIATI, A., COSTANTINE, M.M., MUNN, M.B., PACHECO, L.D., SAADE, G.R., CHIOSSI, G. (2017). Maternal obesity is associated with chorioamnionitis and earlier indicated preterm delivery among expectantly managed women with preterm premature rupture of membranes. *Journal of Maternal Fetal & Neonatal Medicine.* 22:1-8.

HARLEY, K.G., HUEN, K., AGUILAR SCHALL, R., HOLLAND, N.T., BRADMAN, A., BARR, D.B., ESKENAZI, B. (2011). Association of organophosphate pesticide exposure and paraoxonase with birth outcome in Mexican-American women. *PLoS One*. 6(8):e23923.

HARMON, Q.E., ENGEL, S.M., OLSHAN, A.F., MORAN, T., STUEBE, A.M., LUO, J., WU, M.C., AVERY, C.L. (2013). Association of polymorphisms in natural killer cell-related genes with preterm birth. *American Journal of Epidemiology.* 178(8):1208–18.

HARPHAM, M.E., ALGERT, C.S., ROBERTS, C.L., FORD, J.B., SHAND, A.W. (2017). Cervical cerclage placed before 14 weeks gestation in women with one previous midtrimester loss: A population-based cohort study. *The Australian & New Zealand Journal of Obstetrics & Gynaecology.* 1-6

HARPER, M,. ZHENG, S.L., THOM, E., KLEBANOFF, M.A., THORP, J. Jr., SOROKIN, Y., VARNER, M.W., IAMS, J.D., DINSMOOR, M., MERCER, B.M., ROUSE, D.J. RAMIN, S.M., ANDERSON, G.D. EUNICE KENNEDY SHRIVER NATIONAL INSTITUTE OF CHILD HEALTH AND HUMAN DEVELOPMENT (NICHD) MATERNAL-FETAL MEDICINE UNITS NETWORK (MFMU). (2011) Cytokine gene polymorphisms and length of gestation. *Obstetrics & Gynecology.* 117:125–30.

HASSAN, S.S., ROMERO, R., PINELES, B., TARCA, A.L., MONTENEGRO, D., EREZ, O., MITTAL, P., KUSANOVIC, J.P., MAZAKI-TOVI, S., ESPINOZA, J., NHAN-CHANG, C.L., DRAGHICI, S., KIM, C.J. (2010). MicroRNA expression profiling of the human uterine cervix after term labor and delivery. *American Journal Obstetrics & Gynecology.* 202(1):80.e1-8.

HASSAN, S.S., ROMERO, R., VIDYADHARI, D., FUSEY, S., BAXTER, J.K., KHANDELWAL, M., VIJAYARAGHAVAN, J., TRIVEDI, Y., SOMA-PILLAY, P., SAMBAREY, P., DAYAL, A., POTAPOV, V., O'BRIEN, J., ASTAKHOV, V., YUZKO, O., KINZLER, W., DATTEL, B., SEHDEV, H., MAZHEIKA, L., MANCHULENKO, D., GERVASI, M.T., SULLIVAN, L., CONDE-AGUDELO, A., PHILLIPS, J.A., CREASY, G.W. (2011). Vaginal progesterone reduces the rate of preterm birth in women with a sonographic short cervix: a

multicenter, randomized, double-blind, placebo-controlled trial. *Ultrasound in Obstetrics & Gynecology*. 38 (1):18-31

HAVELOCK, J.C., KELLER, P., MULEBA, N., MAYHEW, B.A. CASEY, B.M., RAINEY, W.E., WORD, R.A. (2005). Human myometrial gene expression before and during parturition. *Biology of Reproduction.* 72(3):707-19.

HEAZALL, A.E. BERNATAVICIUS, G., WARRANDER, L., BROWN, M.C., DUNN, W.B. (2012) A metabolomic approach identifies differences in maternal serum in third trimester pregnancies that end in poor perinatal outcome. *Reproductive Science.* 19(8):863-75.

HEINZMANN, A., MAILAPARAMBIL, B., MINGIRULLI, N., KRUEGER, M. (2009). Association of interleukin-13/-4 and toll-like receptor 10 with preterm births. *Neonatology*. 96:175–81.

HENG, Y. J., PENNELL, C. E., CHUA, H. N., PERKINS, J. E., & LYE, S. J. (2014). Whole blood gene expression profile associated with spontaneous preterm birth in women with threatened preterm labor. *PloS one*, *9*(5), e96901.

HENG, Y.J., PENNELL, C.E., MCDONALD, S.W., VINTURACHE, A.E., XU, J., LEE, M.W., BRIOLLAIS, L., LYON, A.W., SLATER, D.M., BOCKING, A.D., DE KONING, L., OLSON, D.M., DOLAN, S.M., TOUGH, S.C., LYE, S.J. (2016). Maternal Whole Blood Gene Expression at 18 and 28 Weeks of Gestation Associated with Spontaneous Preterm Birth in Asymptomatic Women. *PLoS One.* 11(6):e0155191.

HEZELGRAVE, N.L., WATSON, H.A., RIDOUT, A. DIAB, F., SEED, P., CHIN-SMITH, E., TRIBE, R.M., SHENNAN, A. (2016) Rationale and design of SuPPoRT: a multi-centre randomised controlled trial to compare three treatments: cervical cerclage, cervical pessary and vaginal progesterone, for the prevention of preterm birth in women who develop a short cervix. *BMC Pregnancy Childbirth.* 16**,** 358.

HO, T.K. (1995). Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC pp. 278–282.

LLEGAARD, M.V., GROVE, J., THORSEN, P., WANG, X., MANDRUP, S., CHRISTIANSEN, M., NORGAARD-PEDERSEN, B., WOJDEMANN, K.R., TABOR, A., ATTERMANN, J., HOUGAARD, D.M. (2008). Polymorphisms in the tumor necrosis factor alpha and interleukin 1- beta promoters with possible gene regulatory functions increase the risk of preterm birth. *Acta Obstetricia et Gynecologica Scandinavica.* 87(12):1285–90.

HOLZINGER, E. A., DUDEK, S.M., FRASE, A.T., PENDERGRASS, S.A., RITCHIE, M.D. (2014). ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics.* 30(5): 698–705.

HONG, X., HAO, K., JI, H., PENG, S., SHERWOOD, B., DI NARZO, A.,TSAI, H.J., LIU, X., BURD, I. WANG, G., JI, Y., CARUSO, D., MAO., G., BARTELL, T.R., ZHANG Z., PEARSON, C., HEFFNER, L., CERDA, S., BEATY, T.H., FALLIN, M. D., LEE-PARRITZ, A., ZUCKERMAN, B., WEEKS, D.E., WANG, X. (2017). Genome-wide approach identifies a novel gene-maternal pre-pregnancy BMI interaction on preterm birth. *Nature Communications.* 8:15608.
HORI, H. 2014. Methylated nucleosides in tRNA and tRNA methyltransferases. *Frontiers in Genetics.* 5:144

HOWIE, B., FUCHSBERGER, C., STEPHENS, M., MARCHINI, J., ABECASIS, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics.* 44(8):955-9.

HOWSON, C.P., KINNEY, M., LAWN, J.E., editor. MARCH OF DIMES, PMNCH, SAVE THE CHILDREN, WORLD HEALTH ORGANISATION. (2012). Born Too Soon: The Global Action Report on Preterm Birth. Geneva: World Health Organisation.

HROMADNIKOVA, I., KOTLABOVA, K., IVANKOVA, K., KROFTA, L. (2017). Expression profile of C19MC microRNAs in placental tissue of patients with preterm prelabor rupture of membranes and spontaneous preterm birth. *Molecular medicine reports*. 16(4):3849-62.

INTERNATIONAL HAPMAP CONSORTIUM. (2003). The International HapMap Project. *Nature.* 426(6968):789-96.

IVANOV, A.I., ROMANOVSKY, A.A. (2006). Putative dual role of ephrin-EPH receptor interactions in inflammation. *International Union of Biochemistry and Molecular Biology Life*. 58; 389-394.

IWANAGA, R., SUGITA, N., HIRANO, E., SASAHARA, J., KIKUCHI, A., TANAKA, K., YOSHIE, H. (2011). FcγRIIB polymorphisms, periodontitis and preterm birth in Japanese pregnant women. *Journal of Periodontal Research*. 46:292–302.

JACKSON, J.E. (2003). A user's guide to principal components. Hoboken, NJ: Wiley-Interscience.

JAIN, A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters.* 31; 651-666.

JAFARZADEH, L., DANESH, A., SADEGHI, M., HEYBATI, F., HASHEMZADEH, M. (2013) Analysis of Relationship between Tumor Necrosis Factor Alpha Gene (G308A Polymorphism) with Preterm Labor. *International Journal of Preventative Medicine*. 4:896–901.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. (2013). An Introduction to Statistical Learning: with Applications in R. *Springer Texts in Statistics*. Springer Science. New York.

JANITZA, S., CELIK, E. & BOULESTEIX, A. (2015). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification, 12*, 885-915.

JEFFCOAT, M.K., JEFFCOAT, R.L., TANNA, N., PARRY, S.H. (2014). Association of a common genetic factor, PTGER3, with outcome of periodontal therapy and preterm birth. *Journal of Periodontology*. 85:446–54.

JIA, M., PÜTZ, B., MÜLLER-MYHSOK (2019). Revisiting the RDC – further usages and extensions. Manuscript in preparation.

JONES, N.M., HOLZMAN, C., FRIDERICI, K.H., JERNIGAN, K., CHUNG, H., WIRTH, J., FISHER, R. (2010). Interplay of cytokine polymorphisms and bacterial vaginosis in the etiology of preterm delivery. *Journal of Reproductive Immunology.* 87:82–9.

JONES, N.M., HOLZMAN, C., TIAN, Y., WITKIN, S.S.., GENC, M., FRIDERICI, K., FISHER, R., SEZEN, D., BABULA, O., JERNIGAN, K.A., CHUNG, H., WIRTH, J. (2012). Innate immune system gene polymorphisms in maternal and child genotype and risk of preterm delivery. *Journal of Maternal Fetal and Neonatal Medicine*. 25(3):240–7.

JULIA, C., CZERNICHOW, S., CHARNAUX, N., AHLUWALIA, N., ANDREEVA, V., TOUVIER, M., GALAN, P., FEZEU, L. (2014). Relationships between adipokines, biomarkers of endothelial function and inflammation and risk of type 2 diabetes. *Diabetes Research and Clinical Practice.* 105(2):231-8.

JUNG, S.H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics*. 21:3097–3104.

JUNG, S.H. & YOUNG, S.S. (2012) Power and sample size calculation for microarray studies. *J Biopharm Stat*. 22(1):30–42.

KALINKA, J., BITNER, A. (2009). Selected cytokine gene polymorphisms and the risk of preterm delivery in the population of Polish women. *Ginekology Polska*. 80:111–7.

KAMATH-RAYNE, B.D. SMITH, H.C., MUGLIA, L.J., MORROW, A.L. (2014). Amniotic fluid: the use of high dimensional biology to understand fetal well-being. *Reproductive Science.* 21(1):6-19.

KANEHISA, M., GOTO, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research.* 28(1):27-30.

KARJALAINEN, M.K., HUUSKO, J.M., TUOHIMAA, A., LUUKKONEN, A., HAATAJA, R., HALLMAN, M. (2012). A study of collectin genes in spontaneous preterm birth reveals an association with a common surfactant protein D gene polymorphism. *Pediatric Research.* 71:93–9.

KARJALAINEN, M.K., HUUSKO, J.M., ULVILA, J., SOTKASIIRA, J., LUUKKONEN, A., TERAMO, K., PLUNKETT, J., ANTTILA, V., PALOTIE, A., HAATAJA, R., MUGLIA, L.J., HALLMAN, M. (2012). A potential novel spontaneous preterm birth gene, AR, identified by linkage and association analysis of X chromosomal markers. PLoS One. 7(12):e51378.

KARJALAINEN, M.K., OJANIEMI M, HAAPALAINEN AM, MAHLMAN M, SALMINEN A, HUUSKO JM, MÄÄTTÄ TA, KAUKOLA T, ANTTONEN J, ULVILA J, HAATAJA R, TERAMO K, KINGSMORE SF, PALOTIE A, MUGLIA LJ, RÄMET M, HALLMAN M. (2015). CXCR3 Polymorphism and Expression Associate with Spontaneous Preterm Birth. *Journal Immunology*. 195(5):2187-98.

KARODY, V.R., LE, M., NELSON, S., MESKIN, K., KLEMM, S., SIMPSON, P., HINES, R., SAMPATH, V. (2013) A TIR domain receptor-associated protein (TIRAP) variant SNP (rs8177374) confers protection against premature birth. *Journal Perinatology.* 33(5):341–6.

KAZEMIER, B.M., BUIJS, P.E., MIGNINI, L., LIMPENS, J., DE GROOT, C.J., MOL, B.W. (2014). Impact of obstetric history on the risk of spontaneous preterm birth in singleton and multiple pregnancies: a systematic review. *British Journal Obstetrics & Gynaecology.* 121(10), 1197-208.

KENYON, S.L., TAYLOR, D.J., TARNOW-MORDI, W., ORACLE COLLABORATIVE GROUP. (2001). Broad-spectrum antibiotic for spontaneous preterm labour, the ORACLE II randomized trial. ORACLE Collaborative Group. *Lancet.* 357:989-94.

KIM, J., STIRLING, K.J., COOPER, M.E., ASCOLI, M., MOMANY, A.M., MCDONALD, E.L., RYCKMAN, K.K., RHEA, L., SCHAA, K.L., COSENTINO, V., GADOW, E., SALEME, C., SHI, M., HALLMAN, M., PLUNKETT, J., TERAMO, K.A., MUGLIA, L.J., FEENSTRA, B., GELLER, F., BOYD, H.A., MELBYE, M., MARAZITA, M.L., DAGLE, J.M., MURRAY, J.C. (2013). Sequence variants in oxytocin pathway genes and preterm birth: a candidate gene association study. *BMC Medical Genetics.* 14:77.

KIM, S.H., MACINTYRE, D.A., FIRMINO DA SILVA, M., BLANKS, A.M., LEE, Y.S., THORNTON, S., BENNETT, P. R., TERZIDOU, V. (2015). Oxytocin activates NF-kappaB-mediated inflammatory pathways in human gestational tissues. *Molecular Cell Endocrinology*. 403, 64-77.

KIM, D., SHIN, H., SONG, Y.S., KIM, J.H. (2012). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *Journal of Biomedical Information.,* 45, 1191–1198.

KIM, Y.H., HWANG, H.S., KIM, Y.T., KIM, H.S., PARK, Y.W. (2008). Modulation of matrix metalloproteinase secretion by adenosine A3 receptor in preeclamptic villous explants. *Reproductive Science*. 15; 939-949.

KIN, K., MAZIARZ, J., CHAVAN, A.R., KAMAT, M., VASUDEVAN, S., BIRT, A., EMERA, D., LYNCH, V.J., OTT, T.L., PAVLICEV, M., WAGNER, G.P. (2016) The Transcriptomic Evolution of Mammalian Pregnancy: Gene Expression Innovations in Endometrial Stromal Fibroblasts. *Genome Biology and Evolution*. 8(8); 2459–2473.

KINDINGER, L.M., BENNETT, P.R., LEE, Y.S., MARCHESI, J.R., SMITH, A., CACCIATORE, S. HOLMES, E., NICHOLSON, J.K., TEOH, T.G., MACINTYRE, D.A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone treatment for preterm birth risk. *Microbiome.* 5 (1):6.

KITAMURA T. (2008). Enhancement of lymphocyte migration and cytokine production by ephrinB1 system in rheumatoid arthritis. *American Journal of Physiology-Cell Physiology*. 294(1):189–196.

KNIJNENBURG, T.A., VOCKLEY, J.G., CHAMBWE, N., GIBBS, D.L., HUMPHRIES, C., HUDDLESTON, K.C., KLEIN, E., KOTHIYAL, P., TASSEFF, R., DHANKANI, V., BODIAN, D.L., WONG, W.S.W., GLUSMAN, G., MAULDIN, D.E., MILLER, M., SLAGEL, J., ELASADY, S., ROACH, J.C., KRAMER, R., LEINONEN, K., LINTHORST, J., BAVEJA, R., BAKER, R., SOLOMON, B.D., ELEY, G., IYER, R.K., MAXWELL, G.L., BERNARD, B., SHMULEVICH, I., HOOD, L., NIEDERHUBER, J.E. 2019 Genomic and molecular characterization of preterm birth. *Proceedings of the National Academy of Sciences*. 116 (12) 5819-5827.

KOHL, S.M., KLEIN, M.S., HOCHREIN, J., OEFNER, P.J., SPANG, R., GRONWALD, W. (2012) State-of-the-art-data normalisation methods improve NMR based metabolomic analysis. *Metabolomics.* 8;146-160.

KRAMER, M., PAPAGEORGHIOU, A.T., CULHANE, J., BHUTTA, Z., GOLDENBERG, R.L., GRAVETT, M., IAMS, J.D., CONDE-AGUDELO, A., WALLER, S., BARROS, F., KNIGHT, H., VILLAR, J. (2012). Challenges in defining and classifying the preterm birth syndrome. *American Journal Obstetrics and Gynecology*. 206:108-12.

KUESSEL, L., GRIMM, C., KNÖFLER, M., HASLINGER, P., LEIPOLD, H., HEINZE, G., EGARTER, C., SCHMID, M. (2013). Common oxytocin receptor gene polymorphisms and the risk for preterm birth. *Disease Markers.* 34(1):51–6.

KUMAR, D. (2015). 'Genes, Genetics and Human Genomics' in Kumar, D and Eng, C (eds.) *Genomic Medicine: Principles and Practice (2nd edition).* Oxford University Press. 3.

KUSEC, R., LACZIKA, K., KNÖBL, P., FRIEDL, J., GREINIX, H., KAHLS, P., LINKESCH, W., SCHWARZINGER, I., MITTERBAUER, G., PURTSCHER, B., HAAS, O.A., LECHNER, K., JAEGER, U. (1994). AML1/ETO fusion mRNA can be detected in remission blood samples of all patients with t(8;21) acute myeloid leukemia after chemotherapy or autologous bone marrow transplantation. *Leukemia*. 8(5):735-9.

KYRGIOU, M., VALASOULIS, G., STASINOU, S.M., FOUNTA, C., ATHANASIOU, A., BENNETT, P PARASKEVADIS, E. (2015). Proportion of cervical excision for cervical intraepithelial neoplasia as a predictor of pregnancy outcomes. *International Journal of Gynaecology and Obstetrics*. 128(2):141-7.

KWON, H.S., SOHN, I.S., LEE, J.Y., LEE, S.J., KIM, S.N., KIM, B.J. (2009). Intercellular adhesion molecule-1 K469E polymorphism in Korean patients with spontaneous preterm delivery. *International Journal of Gynaecology and Obstetrics*. 104:37–9.

LACHELIN, G.C., MCGARRIGLE, H.H., SEED, P.T., BRILEY, A., SHENNAN, A.H., POSTON, L. (2009) Low saliva progesterone concentrations are associated with spontaneous early preterm labour (before 34 weeks of gestation) in women at increased risk of preterm delivery. *BJOG : an International Journal of Obstetrics and Gynaecology*. 116(11):1515–9.

LAIN, K.Y., CATALANO, P.M. (2007). Metabolomic changes in pregnancy. *Clinical Obstetrics and Gynecology.* 50(4):938-948.

LAPILLONNE, A. & GRIFFIN, I.J. (2013). Feeding preterm infants today for later metabolic and cardiovascular outcomes. *The Journal of Pediatrics*. 162(3 Suppl), S7-16.

LEE, B.E., PARK, H., PARK, E.A., GWAK, H., HA, E.H., PANG, M.G., KIM, Y.J. (2010). Paraoxonase 1 gene and glutathione S-transferase μ 1 gene interaction with preterm delivery in Korean women. *American Journal of Obstetrics and Gynecology*. 203(6):569. e1–7.

LEE, S.M., LEE, J., SEONG, H.S., LEE, S.E., PARK, J.S., ROMERO, R., YOON, B.H. (2009). The clinical significance of a positive Amnisure test™ in women with term labor with intact membranes. *Journal of Maternal-Fetal and Neonatal Medicine.* 22(4); 305-310.

LEVY, S., SUTTON, G., NG, P.C., FEUK, L., HALPERN, A.L., WALENZ, B.P., AXELROD, N., HUANG, J., KIRKNESS, E.F., DENISOV, G., LIN, Y., MACDONALD J R, PANG AWC, SHAGO M, STOCKWELL T B, TSIAMOURI, A., BAFNA, V., BANSAL, V., KRAVITZ, S.A., BUSAM, D.A., BEESON, K.Y., MCINTOSH, T.C., REMINGTON, K.A., ABRIL, J.F., GILL, J., BORMAN, J., ROGERS, Y.H., FRAZIER, M.E., SCHERER, S.W., STRAUSBERG, R.L., VENTER, J.C. (2007) The Diploid Genome Sequence of an Individual Human. *PLoS Biology*. 5(10):e254.

LEWINSKA, M., ZELENKO, U., MERZEL, F., GOLIC GRDADOLNIK, S., MURRAY, J.C., ROZMAN, D. (2013). Polymorphisms of CYP51A1 from cholesterol synthesis: associations with birth weight and maternal lipid levels and impact on CYP51 protein structure. *PLoS One*. 8:e82554.

LI, C.I., SAMUELS, D.C., ZHAO, Y.Y., SHYR, Y., GUO, Y. (2018) Power and sample size calculations for high-throughput sequencing-based experiments. *Brief Bioinform*. 19(6):1247–1255. doi:10.1093/bib/bbx061

LI, Y., REZNICHENKO, M., TRIBE, R. M., HESS, P. E., TAGGART, M., KIM, H., DEGNORE, J. P., GANGOPADHYAY, S., & MORGAN, K. G. (2009). Stretch activates human myometrium via ERK, caldesmon and focal adhesion signaling. *PloS one*, *4*(10), e7489.

LIANG, M., WANG, X., LI, J., YANG, F., FANG, Z., WANG, L., HU, Y., CHEN, D. (2010). Association of combined maternal-fetal TNF-alpha gene G308A genotypes with preterm delivery: a gene-gene interaction study. Journal of Biomedical *Biotechnology.* e396184.

LIEW, C.C., MA, J., TANG, H.C., ZHENG, R., DEMPSEY, A.A. (2006). The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. *Journal of Laboratory & Clinical Medicine.* 147:126–32.
LIN, W.J., HSUEH, H.M., CHEN, J.J. (2010). Power and sample size estimation in microarray studies. *BMC Bioinformatics*. 11:48.

LIU, X., WANG, G., HONG, X., TSAI, H.J., LIU, R., ZHANG, S., WANG, H., PEARSON, C., ORTIZ, K., WANG, D., HIRSCH, E., ZUCKERMAN, B., WANG, X. (2012). Associations between gene polymorphisms in fatty acid metabolism pathway and preterm delivery in a US urban black population. *Human Genetics.* 131:341–51.

LIU, Y., CHEN, L.Y., SOKOLOWSKA, M., EBERLEIN, M., ALSAATY, S., MARTINEZ-ANTON, A., LOGUN, C., QI, H.Y., SHELHAMER, J.H. (2014). The fish oil ingredient, docosahexaenoic acid, activates cytosolic phospholipase A(2) via GPR120 receptor to produce prostaglandin E(2) and plays an anti-inflammatory role in macrophages. *Immunology.* 143:81-95.

LIZEWSKA, B., TEUL, J., KUC, P., LEMANCEWICZ, A., CHARKIEWICZ, K., GOSCIK, J., KACEROVSKY, M., MENON, R., MILTYK, W., LAUDANSKI, P. (2018). Maternal Plasma Metabolomic Profiles in Spontaneous Preterm Birth: Preliminary Results. *Mediators of Inflammation*. eCollection:9362820.

LOCKWOOD, C.J., SENYEI, A.E., DISCHE, M.R., CASAL, D., SHAH, K.D., THUNG, S.N., JONES, L., DELIGDISCH, L., GARITEM, T.J. (1991). Fetal fibronectin in cervical and vaginal secretions as a predictor of preterm delivery. *The New England Journal of Medicine.* 325 (10), 669-74.

LOPEZ-PAZ, D., HENNIG, P., SCHÖLKOPF, B. (2013). The Randomized Dependence Coefficient. Available at https://arxiv.org/abs/1304.7717. Accessed on 20 Sept 2018.

LUO, Y.J., WEN, X.Z., DING, P., HE, Y.H., XIE, C.B., LIU, T., LIN, J.M, YUAN, S.X., GUO, X.L., JIA, D.Q., CHEN, L.H., HUANG, B.Z., CHEN, W.Q. (2012). Interaction between

maternal passive smoking during pregnancy and CYP1A1 and GSTs polymorphisms on spontaneous preterm delivery. *PLoS One.* 7(11): e49155.

MACINTYRE, D.A., SYKES, L., TEOH, T.G. & BENNETT, P.R. (2012). Prevention of preterm labour via the modulation of inflammatory pathways. *The Journal of Maternal-Fetal & Neonatal Medicine*. 25 Suppl 1:17-20.

MACINTYRE, D.A., CHANDIRAMANI, M., LEE, Y.S., KINDINGER, L., SMITH, A., ANGELOPOULOS, N., LEHNE, B., ARULKUMARAN, S., BROWN, R., TEOH, T.G., HOLMES, E., NICOHOLSON, J.K., MARCHESI, J.R., BENNETT, P.R. (2015). The vaginal microbiome during pregnancy and the postpartum period in a European population. *Scientific Reports*. 5, 8988.

MAITRE, L., FTHENOU, E., ATHERSUCH, T., COEN, M., TOLEDANO, M.B., HOLMES, E., KOGEVINAS, M., CHATZI, L., KEUN, H.C. (2014). Urinary metabolic profiles in early pregnancy are associated with preterm birth and fetal growth restriction in the Rhea mother-child cohort study. *BMC Medicine.* 12:110.

MAKIEVA, S., DUBICKE, A., RINALDI, S.F., FRANSSON, E., EKMAN-ORDEBERG, G., NORMAN, J.E. (2017). The preterm cervix reveals a transcriptomic signature in the presence of premature prelabor rupture of membranes. *American Journal of Obstetrics & Gynecology.* 216(6):602.

MANGHAM, L.J., PETROU, S., DOYLE, L.W., DRAPER, E.S., MARLOW, N. (2009). The cost of preterm birth throughout childhood in England and Wales. *Pediatrics*. 123(2), e312-27.

MACQUEEN, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1.* University of California Press. pp.281–297. Retrieved 01-09-2019.

MANN, P.C., COOPER, M.E., RYCKMAN, K.K., COMAS, B., GILI, J., CRUMLEY, S., BRHEAM, E.N., BYERS, H.M., PIESTER, T., SCHAEFER, A., CHRISTINE, P.J., LAWRENCE, A., SCHAA, K.L., KELSEY, K.J., BERENDS, S.K., MOMANY, A.M., GADOW, E., COSENTINO, V., CASTILLA, E.E., LÓPEZ CAMELO, J., SALEME, C., DAY, L.J., ENGLAND, S.K., MARAZITA, M.L., DAGLE, J.M., MURRAY, J.C. (2013). Polymorphisms in the fetal progesterone receptor and a calcium-activated potassium channel isoform are associated with preterm birth in an Argentinian population. *Journal Perinatology*. 33(5):336–40.

MANUCK, T.A., LAI, Y., MEIS, P.J., DOMBROWSKI, M.P., SIBAI, B., SPONG, C.Y., ROUSE, D.J., DURNWALD, C.P., CARITIS, S.N., WAPNER, R.J., MERCER, B.M., RAMIN, S.M. (2011). Progesterone receptor polymorphisms and clinical response to 17-alpha-hydroxyprogesterone caproate. *American Journal Obstetrics and Gynecology*. 205(2):135.e1-9.

MAREES, A. T., DE KLUIVER, H., STRINGER, S., VORSPAN, F., CURIS, E., MARIE-CLAIRE, C., & DERKS, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, *27*(2), e1608.

MARLOW, N., BENNETT, C., DRAPER, E.S., HENNESSY, E.M., MORGAN, A.S., COSTELOE, K.L. (2014). Perinatal outcomes for extremely preterm babies in relation to place of birth in England: the EPICure 2 study. *Archives of Disease in Childhood Fetal and Neonatal edition.* 99 (3), F181-8.

MARTINA, N.A., KIM, E., CHITKARA, U., WATHEN, N.C., CHARD, T., & GIUDICE, L.C. (1997). Gestational age-dependent expression of insulin-like growth factor-binding protein-1 (IGFBP-1) phosphoisoforms in human extraembryonic cavities, maternal serum, and decidua suggests decidua as the primary source of IGFBP-1 in these fluids during early pregnancy. *The Journal of Clinical Endocrinology and Metabolism*. 82(6):1894-8.

MATSUURA, H. & HAKOMORI, S. (1985) The oncofetal domain of fibronectin defined by monoclonal antibody FDC-6: it's presence in fibronectin from fetal and tumor tissues and its absence in those from normal adult tissues and plasma. *Proceedings of the National Acadamy of Science USA*. 82:6517-21.

MATTICK, J.S. (2003). The human genome and the future of medicine. *Medical Journal of Australia.* 179 (4): 212-6.

MAYOR-LYNN, K., TOLOUBEYDOKHTI, T., CRUZ, A.C., CHEGINI, N. (2011). Expression profile of microRNAs and mRNAs in human placentas from pregnancies complicated by preeclampsia and preterm labor. *Reproductive Science*. 18(1):46-56.

MCCORMICK, M.C. (1985). The contribution of low birth weight to infant mortality and childhood morbidity. *The New England Journal of Medicine*. 312 (2), 82-90.

MCDONALD, J.H. (2014). Handbook of Biological Statistics; third edition. Baltimore, Maryland: Sparky House Publishing. Pg 20.

MCDONALD, C.R., TRAN, V., KAIN, K.C. (2015). Complement Activation in Placental Malaria. *Frontiers in Microbiology.* 6, 1460.

MCDUFFIE, R.S., SHERMAN, M.P., GIBBS, R.S. (1992). Amniotic fluid tumor necrosis factor-alpha and interleukin-1 in a rabbit model of bacterially induced preterm pregnancy loss. *American Journal of Obstetrics and Gynecology*. 167(6);1583-8.

MCELROY, J.J., GUTMAN, C.E., SHAFFER, C.M., BUSCH, T.D., PUTTONEN, H., TERAMO, K., MURRAY, J.C. HALLMAN, M., MUGLIA, J.C. (2013). Maternal coding variants in complement receptor 1 and spontaneous idiopathic preterm birth. *Human Genetics*. 132(8):935–42.

MCINTYRE, D.D., LEVENO, K.J. (2008). Neonatal mortality and morbidity rates in late preterm births compared with births at term. *Obstetrics and Gynaecology*. 111: 35-41.

MCLEAN, M & SMITH, R. (1999). Corticotropin-releasing Hormone in Human Pregnancy and Parturition. *Trends in Endocrinology & Metabolism.* 10(5), 174-8.

MEDLEY, N., POLJAK, B., MAMMARELLA., S. ALFIREVIC., Z. (2018a). Clinical guidelines for the management of preterm birth: a systematic review. *BJOG: an International Journal of Obstetrics and Gynaecology.* 125:1361–1369.

MEDLEY, N., VOGEL, J., CARE, A., ALFIREVIC, Z. (2018b). Interventions during pregnancy to prevent preterm birth: an overview of Cochrane systematic reviews. *Cochrane Database of Systematic Reviews.* 11:CD012505.

MEIS, P.J., KLEBANOFF, M., THOM, E., DOMBROWSKI, M.P., SIBAI, B., MOAWAD, A.H., SPONG, C. Y., HAUTH, J. C., MIODOVNIK, M., VARNER, M.W., LEVENO, K. J., CARITIS, S., N.IAMS, J. D.WAPNER, R. J., CONWAY, D.O'SULLIVAN, M. J., CARPENTER, M., MERCER, B., RAMIN, S.M., THORP, J.M., PEACEMAN, A.M., GABBE, S. (2003). Prevention of recurrent preterm delivery by 17 alpha-hydroxyprogesterone caproate. *The New England Journal of Medicine.* 348 (24), 2379-85.

MENON, R., JONES, J., GUNST, P.R., KACEROVSKY, M., FORTUNATO, S.J., SAADE, G.R., BASRAON, S. (2014). Amniotic fluid metabolomic analysis in spontaneous preterm birth. *Reproductive Sciences.* 21(6):791-803.

METABOLOMICS SOCIETY. (2019). *Metabolomics.* [online]. Metabolomics Society, Inc. Aug 27th. URL: http://metabolomicssociety.org/metabolomics

MELCHOR, J.C., KHALIL, A., WING., D., SCHLEUSSNER., E & SURBECK., D. (2018). Prediction of preterm delivery in symptomatic women using PAMG-1, fetal fibronectin and phIGFBP-1 tests: systematic review and meta-analysis. *Ultrasound in Obstetrics & Gynecology.* 52: 442-451.

MEYER, M., SELLAM, J., FELLAHI, S., KOTTI, S., BASTARD, J.P., MEYER, O., LIOTÉ, F., SIMON, T., CAPEAU, J., BERENBAUM, F. (2013). Serum level of adiponectin is a surrogate independent biomarker of radiographic disease progression in early rheumatoid arthritis: results from the ESPOIR cohort. *Arthritis Research & Therapy*. 15(6):R210.

MIDDLETON, P., GOMERSALL, J.C., GOULD, J.F., SHEPHERD, E., OLSEN, S.F., MAKRIDES, M. (2018). Omega-3 fatty acid addition during pregnancy. *Cochrane Database of Systematic Reviews.* 11:CD003402.

MIGALE, R., MACINTYRE, D.A., CACCIATORE, S., LEE, Y.S., HAGBERG, H., HERBERT, B.R., JOHNSON, M.R., PEEBLES, D., WADDINGTON, S.N., BENNETT, P.R. (2016).

Modeling hormonal and inflammatory contributions to preterm and term labor using uterine temporal transcriptomics. *BMC Medicine.* 14(1):86.

MITTAL, P., ROMERO, R., TARCA, A.L., GONZALEZ, J., DRAGHICI, S., XU, Y., DONG, Z., NHAN-CHANG, C.L., CHAIWORAPONGSA, T., LYE, S., KUSANOVIC, J.P., LIPOVICH, L., MAZAKI-TOVI, S., HASSAN, S.S., MESIANO, S., KIM, C.J. (2010). Characterization of the myometrial transcriptome and biological pathways of spontaneous human labor at term. *Journal of Perinatal Medicine*. 38(6):617-43.

MODI, B.P., TEVES, M.E., PEARSON, L.N., PARIKH, H.I., HAYMOND-THORNBURG, H., TUCKER, J.L., CHAEMSAITHONG, P., GOMEZ-LOPEZ, N., YORK, T.P., ROMERO, R., STRAUSS, J.F. (2017). Mutations in fetal genes involved in innate immunity and host defence against microbes increase risk of preterm premature rupture of membranes (PPROM). *Molecular Genetics & Genomic Medicine.* 5(6):720-729.

MONANGI, N.K., BROCKWAY, H.M., HOUSE, M., ZHANG, G., MUGLIA, L.J. (2015). The genetics of preterm birth: progress and promise. *Seminars in Perinatology* 39:574-583.

MOURA, E., MATTAR, R., DE SOUZA, E., TORLONI, M.R., GONÇALVES-PRIMO, A., DAHER, S. (2009). Inflammatory cytokine gene polymorphisms and spontaneous preterm birth. *Journal of Reproductive Immunology.* 80:115–21.

MOZURKEWICH, E.L., LUKE, B., AVNI, M., WOLF, F.M. (2000). Working conditions and adverse pregnancy outcome: a meta-analysis. *Obstetrics & Gynecology*. 95(4):623-35.

MUSILOVA, I., KACEROVSKY, M., STEPAN, M., BESTVINA, T., PLISKOVA, L., ZEDNIKOVA, B., JACOBSSON B. (2017). Maternal serum C-reactive protein concentration and intra-amniotic inflammation in women with preterm prelabor rupture of membranes. *PLoS One.* 12(8).

MUSTAFA, M.D., BANERJEE, B.D., AHMED, R.S., TRIPATHI, A.K., GULERIA, K. (2013) Gene-environment interaction in preterm delivery with special reference to organochlorine pesticides. *Molecular Human Reproduction*. 19(1):35–42.

MYKING, S. BOYD, H.A., MYHRE, R., FEENSTRA, B., JUGESSUR, A., DEVOLD PAY, A.S., ØSTENSEN, I.H.G., MORKEN, N.H., BUSCH, T., RYCKMAN, K.K., GELLER, F., MAGNUS, P., GJESSING, H.K., MELBYE, M., JACOBSSON, B., MURRAY, J.C. (2013). X-Chromosomal Maternal and Fetal SNPs and the Risk of Spontaneous Preterm Delivery in a Danish/Norwegian Genome-Wide Association Study. *PLoS One.* 8(4): e61781.

MYKING, S., MYHRE, R., GJESSING, H.K., MORKEN, N.H., SENGPIEL, V., WILLIAMS, S.M., RYCKMAN, K.K., MAGNUS, P., JACOBSSON, B. (2011). Candidate gene analysis of spontaneous preterm delivery: new insights from re-analysis of a case-control

study using case-parent triads and control-mother dyads. *BMC Medicine Genetics.* 12:174.

NAGALAKSHMI, U., WANG, Z., WAERN, K., SHOU, C., RAHA, D. GERSTEIN, M., SNYDER, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science.* 6(320):1344-1349.

NATIONAL CENTRE FOR BIOTECHNOLOGY INFORMATION [Internet]. (2018) Bethesda (MD): National Library of Medicine (US), National Institute of Health. [cited 2018 Sep 2]. Available from: https://www.ncbi.nlm.nih.gov/gene/10154

NHS ENGLAND. (2019) Saving Babies' Lives Care Bundle Version 2. [online] Available from: https://www.england.nhs.uk/wp-content/uploads/2019/07/saving-babies-lives-care-bundle-version-two-v5.pdf

NICHOLSON, J. K.; LINDON, J. C.; HOLMES, E. (1999). "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, *29*, 1181−1189.

NIKOLOVA, T., UOTILA, J., NIKOLOVA, N., BOLOTSKIKH, V.M., BORISOVA, V.Y., DI RENZO, G.C. (2018). Prediction of spontaneous preterm delivery in women presenting with premature labor: a comparison of placenta alpha microglobulin-1, phosphorylated insulin-like growth factor binding protein-1, and cervical length. *American Journal of Obstetrics and Gynaecology.* 219 (610):e1-9.

NORMAN, J.E., MARLOW, N., MESSOW, C.M., SHENNAN, A., BENNETT, P.R., THORNTON, S., ROBSON, S. C., MCCONNACHIE, A., PETROU, S., SEBIRE, N. J., LAVENDER, T., WHYTE, S., NORRIE, J. OPPTIMUM STUDY GROUP. (2016). Vaginal progesterone prophylaxis for preterm birth (the OPPTIMUM study): a multicentre, randomised, double-blind trial. *Lancet.* 387, 2106-16.

OFFICE FOR NATIONAL STATISTICS. (2017). Statistical Bulletin: Births in England and Wales: 2017. Accessed at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2017 (22 Aug 2018)

OH, D.Y., TALUKDAR, S., BAEE, J., IMAMURA, T., MORINAGA, H., FAN, W., LI P., LU W.J., WATKINS S.M., OLEFSKY, J.M. (2010). GPR120 is an omega-3 fatty acid receptor mediating potent anti-inflammatory and insulin-sensitizing effects. *Cell.* 142:687-98.

OLIVEIRA, T.A., DA CUNHA, D.R., POLICASTRO, A., TRAINA, É., GOMES, M.T., CORDIOLI, E. (2011). The progesterone receptor gene polymorphism as factor of risk for the preterm delivery. *Revista Brasileira de Ginecologia e Obstetrícia*. 33:271–75.

ORCZYK-PAWILOWICZ, M., JAWIEN, E., DEJA, S., HIRNLE, L., ZABEK, A., MLYNARZ, P. (2016). Metabolomics of Human Amniotic Fluid and Maternal Plasma during Normal Pregnancy. *PLoS One.* 11(4):e0152740

PACAGNELLA, R.C., MOL, B.W., BOROVAC-PINHEIRO, A. Renato Passini Jr., Nomura, M.L., Andrade, K.C., Ellovitch, N., Fernandes, K.G., Bortoletto, T.G., Pereira, C.M., Miele, M.J., França, M.S., Cecatti, J.G., P5 Working Group. (2019) A randomized controlled trial on the use of pessary plus progesterone to prevent preterm birth in women with short cervical length (P5 trial). *BMC Pregnancy Childbirth* 19**,** 442.

PAPPAS, A., CHAIWORAPONGSA, T., ROMERO, R., KORZENIEWSKI, S.J., CORTEZ, J.C., BHATTI, G., GOMEZ-LOPEZ, N., HASSAN, S.S., SHANKARAN, S., TARCA, A.L. (2015). Transcriptomics of maternal and fetal membranes can discriminate between gestational age matched preterm neonates with and without cognitive impairment diagnosed at 18-24 months. *PLoS One.* 10(3):e0118573.

PARTHIBAN, P & MAHENDRA, J. (2015). Toll-Like Receptors: A Key Marker for Periodontal Disease and Preterm Birth - A Contemporary Review. *Journal of clinical and diagnostic research*. 9(9),e14-7.

PATEL, K., WILLIAMS, S., GUIRGUIS, G., GITTENS-WILLIAMS, L., APUZZIO, J. (2017) Genital tract GBS and rate of histologic chorioamnionitis in patients with preterm premature rupture of membrane. *Journal of Maternal-Fetal & Neonatal Medicine*. 1-4.

PEREYRA, S., SOSA, C., BERTONI, B. & SAPIRO, R. (2019). Transcriptomic analysis of fetal membranes reveals pathways involved in preterm birth. *BMC Medical Genomics.* 12:53.

PEREZA, N., PLEŠA, I., PETERLIN, A., JAN, Z., TUL, N., KAPOVIC, M., OSTOJIĆ, S., PETERLIN, B.
(2014). Functional polymorphisms of matrix metalloproteinases 1 and 9 genes in women with spontaneous preterm birth. *Disease Markers.* 171036.

PETRICEVIC, L., DOMIG, K.J., NIERSCHER, F.J., SANDHOFER, M.J., FIDESSER, M., KRONDORFER, I. HUSSLEIN, P., KNEIFEL, W., KISS, H. (2014). Characterisation of the vaginal Lactobacillus microbiota associated with preterm delivery. *Scientific Reports*. 4:5136.

PETRUNIN, D.D., GRYAZNOVA, I.M., PETRUNINA YU, A.,TATARINOV YU, S. (1976). Immunochemical identification of the human placenta organospecific α_2 globulin and its concentration in the amniotic fluid. *Byulleten Eksperimentalnoi Biologii i Meditsiny. 82 (7): 83-84.*

PHARANDE, P. MOHAMED, A. L. BAJUK, B. LUI, K. BOLISETTY, S. (2017). Preterm infant outcomes in relation to the gestational age of onset and duration of

prelabour rupture of membranes: a retrospective cohort study. *BMJ Paediatrics Open.* 1, 1-8.

PINTO, J., BARROS, A.S., DOMINGUES, M.R., GOODFELLOW, B.J., GALHANO, E., PITA C, ALMEIDA MDO, C., CARREIRA, I.M., GIL, A.M. (2015). Following healthy pregnancy by NMR metabolomics of plasma and correlation to urine. *Journal of Proteome Research*. 14(2):1263-74.

PLUNKETT, J., DONIGER, S., ORABONA, G., MORGAN, T., HAATAJA, R., HALLMAN, M., PUTTONEN, H., MENON, R., KUCZYNSKI, E., NORWITZ, E., SNEGOVSKIKH, V., PALOTIE, A., PELTONEN, L., FELLMAN, V., DEFRANCO, E.A., CHAUDHARI, B.P., MCGREGOR, T.L., MCELROY, J.J., OETJENS, M.T., TERAMO, K., BORECKI, I., FAY, J., MUGLIA, L. (2011). An evolutionary genomic approach to identify genes involved in human birth timing. *PLoS Genetics*. 7(4):e1001365.

PORTER, T.F., FRASER, A.M., HUNTER, C.Y., WARD, R.H., VARNER, M.W. (1997). The risk of preterm birth across generations. *Obstetrics & Gynecology*. 90 (1):63-7.

PRUIM, R.J., WELCH, R.P., SANNA, S., TESLOVICH, T.M., CHINES, P.S., GLIEDT, T.P., BOEHNKE, M., ABECASIS, G.R., WILLER, C.J. (2010). LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics.* 26(18): 2336-2337.

PLUNKETT, J., FEITOSA, M.F., TRUSGNICH, M., WANGLER, M.F., PALOMAR, L., KISTKA, Z.A., DEFRANCO, E.A., SHEN, T.T., STORMO, E., PUTTONEN, H., HALLMAN, M., HAATAJA, R., LUUKKONEN, A., FELLMAN, V., PELTONEN, L., PALOTIE, A., DAW E.W., ANN, P., TERAMO, K., BORECKI, I., MUGLIA, L.J. (2009). Mother's genome or maternally inherited genes acting in the fetus influence gestational age in familial preterm birth. *Human Heredity.* 68:209–219

POLETTINI, J., DUTTA, E.H., BEHNIA, F., SAADE, G.R., TORLONI, M.R., MENON, R. (2015). Aging of intrauterine tissues in spontaneous preterm birth and preterm premature rupture of the membranes: A systematic review of the literature. *Placenta.* 36(9):969-73.

PU, J. & ZENG, W.Y. (2007) Gene polymorphism of tumor necrosis factor-alpha promoter region in -308 site and premature births in Chinese Han populations. *Sichuan Da Xue Xue Bao Yi Xue Ban.* 38:984–6.(Abstract)

RADOCHOVA, V., KACEROVSKA MUSILOVA, I., STEPAN, M., VESCICIK, P., SLEZAK, R., JACOBSSON, B., KACEROVSKY, M. (2018). Periodontal disease and intra-amniotic complications in women with preterm prelabor rupture of membranes. *Journal of Maternal-Fetal & Neonatal Medicine*. 31(21):2852-2861.

RAVEL, J., GAJER, P., ABDO, Z., SCHNEIDER, G.M., KOENIG S.S.K., MCCULLE, S.L., KARLEBACH, S., GORLE, R., RUSSELL, J., TACKET, C.O., BROTMAN, R.M., DAVIS, C.C., AULT K., PERALTA, L., FORNEY, L.J. (2011). Vaginal microbiome of reproductive-age

women. *Proceedings of the National Academy of Sciences U S A*. 108 (Suppl. 1):4680-4687.

RAYMAN, M.P., BODE, P., REDMAN, C.W. (2003). Low selenium status is associated with the occurrence of the pregnancy disease pre-eclampsia in women from the United Kingdom. *American Journal of Obstetrics and Gynecology*. 189; 1343-9.

RAYMAN, M.P., WIJNEN, H., VADER, H., KOOISTRA, L., POP, V. (2011). Maternal selenium status during early gestation and risk for preterm birth. *Canadian Medical Association Journal.* 183:549-555.

REICH, D.E., LANDER, E.S. (2001). On the allelic spectrum of human disease. *Trends in Genetics.* 17: 502-510.

REID, G., YOUNES, J.A., VAN DER MEI, H.C., GLOOR, G.B., KNIGHT, R., BUSSCHER, H.J. (2011). Microbiota restoration: natural and supplemented recovery of human microbial communities. *Nature Reviews Microbiology.* 9(1):27–38.

REY, G., SKOWRONEK, F., ALCIATURI, J., ALONSO, J., BERTONI, B., SAPIRO, R. (2008). Toll receptor 4 Asp299Gly polymorphism and its association with preterm birth and premature rupture of membranes in a South American population. *Molecular Human Reproduction*. 14:555–9.

RINALDI, S.F., MAKIEVA, S., SAUNDERS, P.T., ROSSI, A.G., NORMAN, J.E. (2017). Immune cell and transcriptomic analysis of the human decidua in term and preterm parturition. *Molecular Human Reproduction.* 23 (10):708-724.

RITCHIE, M., HOLZINGER, E., LI, R., PENDERGRASS, S., KIM, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genomics.* 85-97.

ROBERTS, C.L., WAGLAND, P., TORVALDSEN, S., BOWEN, J.R., BENTLEY, J.P., MORRIS, J.M. (2017). Childhood outcomes following preterm prelabor rupture of the membranes (PPROM): a population-based record linkage cohort study. *Journal of Perinatology.* 37; 1230-1235.

ROCHEFORT, S. (2005). Metabolomics Reviewed: A New "Omics" Platform Technology for Systems Biology and Implications for Natural Products Research. *Journal of Natural Products.*68(12);1813-1820.

ROGGERO, P., GIANNI, M.L., GARBARINO, F. & MOSCA, F. (2013). Consequences of prematurity on adult morbidities. *European Journal of Internal Medicine*. 24 (7), 624-6.

ROHART, F., GAUTIER, B., SINGH, A., LÊ CAO, K.A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology* 13(11): e1005752.

ROMERO, R., FRIEL, L.A., VELEZ EDWARDS, D.R., KUSANOVIC, J.P., HASSAN, S.S., MAZAKI-TOVI, S., VAISBUCH, E., KIM, C.J., EREZ, O., CHAIWORAPONGSA, T., PEARCE, B.D., BARTLETT, J., SALISBURY, B.A., ANANT, M.K., VOVIS, G.F., LEE, M.S., GOMEZ, R., BEHNKE, E., OYARZUN, E., TROMP, G., WILLIAMS, S.M., MENON, R. (2010a). A genetic association study of maternal and fetal candidate genes that predispose to preterm prelabor rupture of membranes (PROM). *American Journal of Obstetrics and Gynecology.* 203(4):361.e1-e361.

ROMERO, R., VELEZ EDWARDS, D.R., KUSANOVIC, J.P., HASSAN, S.S., MAZAKI-TOVI, S., VAISBUCH, E., KIM, C.J., CHAIWORAPONGSA, T., PEARCE, B.D., FRIEL, L.A., BARTLETT, J., ANANT, M.K., SALISBURY, B.A., VOVIS, G.F., LEE, M.S., GOMEZ, R., BEHNKE, E., OYARZUN, E., TROMP, G., WILLIAMS, S.M., MENON, R. (2010b). Identification of fetal and maternal single nucleotide polymorphisms in candidate genes that predispose to spontaneous preterm labor with intact membranes. *American Journal of Obstetrics and Gynecology.* 202:431.e1-34.

ROMERO, R., MAZAKI-TOVI, S., VAISBUCH, E., KUSANOVIC, J.P., CHAIWORAPONGSA, T., GOMEZ, R., NIEN, J.K., YOON, B.H., MAZOR, M., LUO, J., BANKS, D., RYALS, J., BEECHER, C. (2010c). Metabolomics in premature labor: a novel approach to identify patients at risk for preterm delivery. *The Journal of Maternal-Fetal & Neonatal Medicine.* 23(12):1344-59.

ROMERO, R., NICOLAIDES, K., CONDE-AGUDELO, A., TABOR, A., O'BRIEN, J.M., CETINGOZ, E., DA FONSECA, E., CREASY, G. W., KLEIN, K., RODE, L., SOMA-PILLAY, P., FUSEY, S., CAM, C., ALFIREVIC, Z., HASSAN, S. S. (2012). Vaginal progesterone in women with an asymptomatic sonographic short cervix in the midtrimester decreases preterm delivery and neonatal morbidity: a systematic review and metaanalysis of individual patient data. *American Journal of Obstetrics and Gynecology.* 206(2), e1-19.

ROMERO, R., HASSAN, S.S., GAJER, P., TARCA, A.L., FADROSH, D.W., NIKITA, L., GALUPPI, M., LAMONT, R.F., CHAEMSAITHONG, P., MIRANDA, J., CHAIWORAPONGSA, T., RAVEL, J. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome.* 2:10.

ROMERO, R., CONDE-AGUDELO, A., DA FONSECA, E., O'BRIEN, J.M., CETINGOZ, E., CREASY, G.W., HASSAN, S.S., NICOLAIDES, K.H. (2018). Vaginal progesterone for preventing preterm birth and adverse perinatal outcomes in singleton gestations with a short cervix: a meta-analysis of individual patient data. *American Journal of Obstetrics & Gynecology*. 218(2):161–180.

RYCKMAN KK, MORKEN NH, WHITE MJ, VELEZ DR, MENON R, FORTUNATO SJ, MAGNUS, P., WILLIAMS, S.M., JACOBSSON, B. (2010). Maternal and fetal genetic associations of PTGER3 and PON1 with preterm birth. *PLoS One.* 5(2):e9040.

SAADE, G., BOGGESS, K.A., SULLIVAN, S.A., MARKENSON, G.R., IAMS, J.D., COONROD, D.V., PEREIRA, L.M., ESPLIN, M.S., COUSINS, L.M., LAM, G.K., HOFFMAN, M.K., SEVERINSEN, R.D., PUGMIRE, T., FLICK, J.S., FOX, A.C., LUETH, A.J., RUST, S.R., MAZZOLA, E., HSU, C., DUFFORD, M.T., BRADFORD, C.L., ICHETOVKIN, I.E., FLEISCHER, T.C., POLPITIYA, A.D., CRITCHFIELD, G.C., KEARNEY, P.E., BONIFACE, J.J., HICKOK, D.E. (2016) Development and validation of a spontaneous preterm delivery predictor in asymptomatic women. *American Journal of Obstetrics and Gynaecology.* 214(5):633.e1-633.e24.

SAK, S., BARUT, M., INCEBIYIK, A., AGACAYAK, E., KIRMIT, A., KOYUNCU, I., SAK, M. (2017). Comparison of sVCAM-1 and sICAM-1 levels in maternal serum and vaginal secretion between pregnant women with preterm prelabour ruptures of membranes and healthy pregnant women. *Journal of Maternal-Fetal and Neonatal Medicine.*1-6.

SALMINEN, A., KAARNIRANTA, K., KAUPPINEN, A. (2012). Inflammaging: disturbed interplay between autophagy and inflammasomes. *Aging.* Mar;4(3):166-75.

SALOMON, L.J., DIAZ-GARCIA, C., BERNARD, J.P., VILLE, Y. (2009). Reference range for cervical length throughout pregnancy: non-parametric LMS-based model applied to a large sample. *Ultrasound in Obstetrics & Gynecology.* 33(4):459-64.

SANDERS, A.P., BURRIS, H.H., JUST, A.C., MOTTA, V., SVENSSON, K., MERCADO-GARCIA, A., PANTIC, I., SCHWARTZ, J., TELLEZ-ROJO, M.M., WRIGHT, R.O., BACCARELLI, A. (2015). A microRNA expression in the cervix during pregnancy is associated with length of gestation. *Epigenetics,* 10(3):221-8.

SANDMAN, C.A. & DAVIS, E.P. (2012). Neurobehavioural risk is associated with gestational exposure to stress hormones. *Expert Review in Endocrinology Metabolism*. 7(4):445-459.

SATA, F., TOYA, S., YAMADA, H., SUZUKI, K., SAIJO, Y., YAMAZAKI, A. MINAKAMI, H., KISHI, R. (2009). Proinflammatory cytokine polymorphisms and the risk of preterm birth and low birthweight in a Japanese population. *Molecular Human Reproduction.* 15(2):121–30.

SCHENA, M., SHALON, D., DAVIS, R.W., BROWN, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235): 467–470.

SCHMID, M., HASLINGER, P., STARY, S., LEIPOLD, H., EGARTER, C., GRIMM, C. (2012). Interleukin-1 beta gene polymorphisms and preterm birth. European Journal of Obstetrics and Gynecology and Reproductive Biology. 165(1):33–6.

SEAL, B.S., KING, D.J., BENNETT, J.D. (1995). Characterization of Newcastle disease virus isolates by reverse transcription PCR coupled to direct nucleotide sequencing

and development of sequence database for pathotype prediction and molecular epidemiological analysis. *Journal of Clinical Microbiology.* 33(10):2624-30.

SEALFON, S.C. AND CHU, T.T. (2011). RNA and DNA microarrays. *Methods Molecular Biology.* 671:3-34.

SHARP, G.C., HUTCHINSON, J.L., HIBBERT, N., FREEMAN, T.C., SAUNDERS, P.T., NORMAN, J.E. (2016). Transcription Analysis of the Myometrium of Labouring and Non-Labouring Women. *PLoS One.* 11(5):e0155413.

SHEIKH, I.A., AHMAD, E., JAMAL, M.S., REHAN, M., ASSIDI, M., TAYUBI, I.A., ALBASRI, S.F., BAJOUH, O.S., TURKI, R.F., ABUZENADAH, A.M., DAMANHOURI, G.A., MOHD, A., AL-QAHTANI, M. (2016). Spontaneous preterm birth and single nucleotide gene polymorphisms: a recent update. *BMC Genomics.* 17(Suppl 9):759.

SHEN, R. OLSHEN AB, LADANYI M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics,* 25, 2906–2912.

SHETTY, S & COPELAND, P. (2018). The Selenium Transport Protein, Selenoprotein P, Requires Coding Sequence Determinants to Promote Efficient Selenocysteine Incorporation. *Journal of Molecular Biology.* 430(24); 5217-5232.

SHREE, R., CAUGHEY, A.B., CHANDRASEKARAN, S. (2017). Short interpregnancy interval increases the risk of preterm premature rupture of membranes and early delivery. *Journal of Maternal- Fetal and Neonatal Medicine.* 31(22):3014-3020.

SMOLINSKA, A., BLANCHET, L., BUYDENS, L.M.C., WIJMENGA, S.S. (2012). NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta.* 750(1); 82-97.

SOCIETY FOR MATERNAL FETAL MEDICINE PUBLICATIONS COMMITTEE. (2008). ACOG Committee Opinion number 419: Use of progesterone to reduce preterm birth. *Obstetrics & Gynecology.* 112, 963-5.

SOININEN, P., KANGAS, A.J., WÜRTZ, P., TUKIAINEN, T., TYNKKYNEN, T., LAATIKAINEN, R., JÄRVELIN, M.R., KÄHÖNEN, M., LEHTIMÄKI, T., VIIKARI, J., RAITAKARI, O.T., SAVOLAINEN, M.J., ALA-KORPELA, M. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst*. 134 (9):17781-5.

SOORANNA, S.R., LEE, Y., KIM, L.U., MOHAN, A.R., BENNETT, P.R., JOHNSON, M.R. (2004). Mechanical stretch activates type 2 cyclooxygenase via activator protein-1 transcription factor in human myometrial cells. *Molecular Human Reproduction*. 10(2), 109-13.

SPECHT, D.F. (1990). Probabilistic neural network.*Neural Networks,* 3; 109–118.

SPONG, C.Y. (2013). Defining “term” pregnancy: recommendations from the Defining “Term” Pregnancy Workgroup. *Journal of the American Medical Association.* 309(23):2445-6.

STANFIELD, Z., LAI, P. F., LEI, K., JOHNSON, M. R., BLANKS, A. M., ROMERO, R., CHANCE, M. R., MESIANO, S., & KOYUTÜRK, M. (2019). Myometrial Transcriptional Signatures of Human Parturition. *Frontiers in genetics*, *10*, 185.

STEFFEN, K.M., COOPER, M.E., SHI, M., CAPRAU, D., SIMHAN, H.N., DAGLE, J.M., MARAZITA, M.L., MURRAY, J.C. (2007). Maternal and fetal variation in genes of cholesterol metabolism is associated with preterm delivery. *Journal of Perinatology.* 27 (11):672–80.

STOUT, M.J., ZHOU, Y., WYLIE, K.M., TARR, P.I., MACONES, G.A., TUULI, M.G. (2017). Early pregnancy vaginal microbiome trends and preterm birth. *American Journal of Obstetrics and Gynaecology.* 217 (3), 1–18.

STONEK, F., METZENBAUER, M., HAFNER, E., PHILIPP, K., TEMPFER, C. (2008). Interleukin-10 -1082 G/A promoter polymorphism and pregnancy complications: results of a prospective cohort study in 1,616 pregnant women. *Acta Obstetrics Gynecology Scandinavia.* 87:430–43.

STRAUSS, J.F., ROMERO, R., GOMEZ-LOPEZ, N., HAYMOND-THORNBURG, H., MODI, B.P., TEVES, M.E., PEARSON, L.N., YORK, T.P., SCHENKEIN, H.A. (2018). Spontaneous preterm birth: advances toward the discovery of genetic predisposition. *American Journal of Obstetrics and Gynaecology.* 218 (3): 294-314.e2

SUBRAMANIAM, A., TAMAYO, P., MOOTHA, V.K., MUKHERJEE, S., EBERT, B.L., GILLETTE, M.A. PAULOVICH, A., POMEROY, S.L., GOLUB, T.R., LANDER, E,S., MESIROV, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Acadamy of Science USA*. 102(43):15545–50.

SUGITA, N., KOBAYASHI, T., KIKUCHI, A., SHIMADA, Y., HIRANO, E., SASAHARA, J., TANAKA, K., YOSHIE, H. (2012). Immunoregulatory gene polymorphisms in Japanese women with preterm births and periodontitis. *Journal of Reproductive Immunology.* 93(2):94–101.

SUH, Y.J., KIM, Y.J., PARK, H., PARK, E.A., HA, E.H. (2008). Oxidative stress-related gene interactions with preterm delivery in Korean women. *American Journal of Obstetrics and Gynecology*. 198:541. e1–7.

SUNG, J.H., KUK, J.Y., CHA, H.H., CHOI, S.J., OH, S.Y., ROH, C.R., KIM, J.H. (2017). Amniopatch treatment for preterm premature rupture of membranes before 23 weeks' gestation and factors associated with its success. *Taiwanese Journal of Obstetrics & Gynecology*. 56(5):599-605.

SUH, Y.J., PARK, H.J., LEE, K.A., LEE, B.E., HA, E.H., KIM, Y.J. (2013). Associations between genetic polymorphisms of beta-2 adrenergic receptor and preterm delivery in Korean women. *American Journal of Reproductive Immunology.* 69:85–91.

TAKANO, M., LU, Z., GOTO, T., FUSI, L., HIGHAM, J., FRANCIS, J., WITHEY, A., HARDT, J., CLOKE, B., STAVROPOULOU, A.V., ISHIHARA, O., LAM, E.W., UNTERMAN, T.G., BROSENS, J.J., KIM, J.J. (2007). Transcriptional cross talk between the forkhead transcription factor forkhead box O1A and the progesterone receptor coordinates cell cycle regulation and differentiation in human endometrial stromal cells. *Molecular Endocrinology.* 21(10):2334-49.

THAN, N.G., ROMERO, R., TARCA, A.L., DRAGHICI, S., EREZ, O., CHAIWORAPONGSA, T., KIM, Y.M., KIM., S.K. VAISBUCH, E., TROMP G. (2009). Mitochondrial manganese superoxide dismutase mRNA expression in human chorioamniotic membranes and its association with labor, inflammation, and infection. *The Journal of Maternal-Fetal & Neonatal Medicine.* 22(11):1000-13.

THARWAT, A., GABER, T., IBRAHIM, A., HASSANIEN, A.E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications.* 00; 1–22

THEOBALD, O. (2017). Machine Learning for Absolute Beginners (2nd Ed). Great Britain. Publisher: Amazon.

THOTA, C., MENON, R., WENTZ, M.J., FORTUNATO, S.J., BARTLETT, J., DROBEK, C.O., NAIR, S., AL-HENDY, A. (2012) A single-nucleotide polymorphism in the fetal catechol-O-methyltransferase gene is associated with spontaneous preterm birth in African Americans. *Reproductive Science* 19(2):135–42.

TIAN, Z., PALMER, N., SCHMID, P., YAO, H., GALDZICKI, M., BERGER, B., WU, E., KOHANE, I.S. (2009) A practical platform for blood biomarker study by using global gene expression profiling of peripheral whole blood. *PLoS ONE*. 4(4):e5157

TIERENEY, B. (2018) Random Machine Learning in R, Python and SQL – Part 1. Toad World Blog. Aug, 31. Available at: *https://blog.toadworld.com/2018/08/31/random-forest-machine-learning-in-r-python-and-sql-part-1*

TING, H.S., CHIN, P.S., YEO, G.S. & KWEK, K. (2007). Comparison of bedside test kits for prediction of preterm delivery: phosphorylated insulin-like growth factor binding protein-1 (pIGFBP-1) test and fetal fibronectin test. *Annals of the Academy of Medicine*. 36(6), 399-402.

TO, M.S., ALFIREVIC, Z., HEATH, V.C., CICERO, S., CACHO, A.M., WILLIAMSON, P.R., NICOLAIDES, K.H., FETAL MEDICINE FOUNDATION SECOND TRIMESTER SCREENING GROUP. (2004). Cervical cerclage for prevention of preterm delivery in women with short cervix: randomised controlled trial. *Lancet*. 363(9424), 1849-53.

TO, M.S., SKENTOU, C.A., ROYSTON, P., YU, C.K., NICOLAIDES, K.H. (2006). Prediction of patient-specific risk of early preterm delivery using maternal history and sonographic measurement of cervical length: a population-based prospective study. *Ultrasound in Obstetrics & Gynecology*. 27(4); 362-7.
TONDE, C., & ELGAMMAL, A.M. (2016). Learning Kernels for Structured Prediction using Polynomial Kernel Transformations. *ArXiv, abs/1601.01411*.

TOPRAK, E., BOZKURT, M., DINCGEZ CAKMAK, B., OZCIMEN, E.E., SILAHLI, M., ENDER YUMRU, A., ÇALIŞKAN, E. (2017). Platelet-to-lymphocyte ratio: A new inflammatory marker for the diagnosis of preterm premature rupture of membranes. *Journal of the Turkish German Gynecological Association.* 18(3):122-6.

VAN DEN BERG, R. A., HOEFSLOOT, H. C., WESTERHUIS, J. A., SMILDE, A. K., & VAN DER WERF, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, *7*, 142.

VAN GELDER, R.N., VON XASTROW, M.E., YOOL, E. DEMENT, W.C., BARCHAS, J.D., EBERWINE, J.H. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proceedings of the National Academy of Sciences USA*. 87(5):1663-1667.

VANDENBROUCKE, L., DOYEN, M., LE LOUS, M., BEUCHEE, A., LOGET, P., CARRAULT, G., PLADYS, P. (2017). Chorioamnionitis following preterm premature rupture of membranes and fetal heart rate variability. *PLoS One*. 12(9):e0184924.

VANDERHOEVEN, J. & TOLOSA, J.E. (2010). Tobacco and Preterm Birth. In: Berghella V, editor. *Preterm Birth: Prevention and Management.* Wiley-Blackwell, Oxford, UK.

VELEZ, D.R., MENON, R., THORSEN, P., JIANG, L., SIMHAN, H., MORGAN, N., FORTUNATO, S.J., WILLIAMS, S.M.(2007). Ethnic differences in interleukin 6 (IL-6) and IL6 receptor genes in spontaneous preterm birth and effects on amniotic fluid protein levels. *Annals of Human Genetics.*71(Part 5):586–600.

VELEZ, D.R., FORTUNATO, S.J., WILLIAMS, S.M., MENON, R. (2008a). Interleukin-6 (IL-6) and receptor (IL6-R) gene haplotypes associate with amniotic fluid protein concentrations in preterm birth. *Human Molecular Genetetics*. 17:1619–30.

VELEZ, D.R., MENON, R., SIMHAN, H., FORTUNATO, S., CANTER, J.A., WILLIAMS S.M. (2008b). Mitochondrial DNA variant A4917G, smoking and spontaneous preterm birth. *Mitochondrion.* 8 (2); 130-135.

VELEZ, D.R., FORTUNATO, S., THORSEN, P., LOMBARDI, S.J., WILLIAMS, S.M., MENON, R. (2009). Spontaneous preterm birth in African Americans is associated with infection and inflammatory response gene variants. American Journal Obstetrics & Gynecology. 200(2):209. e1–27.

VERSTRAELEN, H., VERHELST, R., CLAEYS, G., DE BACKER, E., TEMMERMAN, M., VANEECHOUTTE, M. (2009). Longitudinal analysis of the vaginal microflora in pregnancy suggests that L. crispatus promotes the stability of the normal vaginal microflora and that L. gasseri and/or L. iners are more conducive to the occurrence of abnormal vaginal microflora. *BMC Microbiology*. 9:116.

VILLAR, J., PAPAGEORGHIOU, A.T., KNIGHT, H.E., GRAVETT, M.G., IAMS, J., WALLER, S.A., KRAMER, M. CULHANE, J. F. BARROS, F. C. CONDE-AGUDELO, A. BHUTTA, Z. A. GOLDENBERG, R. L. (2012). The preterm birth syndrome: a prototype phenotypic classification. *American Journal of Obstetrics and Gynecology*. 206 (2), 119-23.

VIRGILIOU, C., GIKA, H.G., WITTING, M., BLETSOU, A.A., ATHANASIADIS, A., ZAFRAKAS, M., THOMAIDIS, N.S., RAIKOS, N., MAKRYDIMAS, G., THEODORIDIS, G.A. (2017). Amniotic Fluid and Maternal Serum Metabolic Signatures in the Second Trimester Associated with Preterm Delivery. *Journal of Proteome Research.* 16(2):898-910.

VORA, B., WANG, A., KOSTI, I., HUANG, H., PARANJPE, I., WOODRUFF, T.J., MACKENZIE, T., SIROTA, M. (2018). Meta-Analysis of Maternal and Fetal Transcriptomic Data Elucidates the Role of Adaptive and Innate Immunity in Preterm Birth. *Frontiers in Immunology.* 9:993.

VOUSDEN, N., HEZELGRAVE, N., CARTER, J., SEED, P.T., SHENNAN, A.H. (2015). Prior ultrasound-indicated cerclage: how should we manage the next pregnancy? *European Journal of Obstetrics, Gynecology, and Reproductive Biology*. 188:129-32.

VUNTA, H., DAVIS, F., PALEMPALLI, U.D., BHAT, D., ARNER, R.J., THOMPSON, J.T., PETERSON, D.G., REDDY, C.C., PRABHU, K.S. (2007) The anti-inflammatory effects of selenium are mediated through 15-deoxy-12,14-prostaglandin J2 in macrophages. *Journal of Biological Chemistry*. 282(25): 17964-73.

WAGGONER, D., WAIN, K. E., DUBUC, A. M., CONLIN, L., HICKEY, S. E., LAMB, A. N., MARTIN, C. L., MORTON, C. C., RASMUSSEN, K., SCHUETTE, J. L., SCHWARTZ, S., MILLER, D. T., & ACMG PROFESSIONAL PRACTICE AND GUIDELINES COMMITTEE (2018). Yield of additional genetic testing after chromosomal microarray for diagnosis of neurodevelopmental disability and congenital anomalies: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine.* 20(10), 1105–1113.

WALSH, M.P. (1994). Calmodulin and the regulation of smooth muscle contraction. *Molecular and Cellular Biochemistry.* 135(1): 21–41.

WANG, Y., LUO, H., CHE, G., LI, Y., GAO, J., YANG, Q., ZHOU, B., GAO, L., WANG, T., LIANG, Y., ZHANG, L. (2018). Placental protein 14 as a potential biomarker for diagnosis of preterm premature rupture of membranes. *Molecular Medicine Reports.* 18, 113-122.

WANG, Z., GERSTEIN, M., SNYDER, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics.* 10(1):57-63.

WATANABE, K., TASKESEN, E., BOCHOVEN A.V., POSTHUMA, D. (2017). Functional mapping and annotation of genetic associations with FUMA was used to identify the enriched genes and associated systems. *Nature Communications.* 8; 1826

WATERS, K. (2010). Serotonin in the sudden infant death syndrome. *Drug News and Perspectives.* Nov;23(9):537-48.

WATSON J.D. & CRICK F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*. 171(4356): 737–38.

WATSON, H.A., SEED, P.T., CARTER, J., HEZELGRAVE, N.L., KUHRT, K., TRIBE, R.M., SHENNAN, A.H. (2020) Development and validation of the predictive models for the QUiPP App v.2: a tool for predicting preterm birth in high-risk asymptomatic women. *Ultrasound Obstetrics & Gynecology.* 55 (3), 348-356
WALDORF, K.M.A., SINGH, N., MOHAN, A. R., YOUNG, R. C., NGO, L., DAS, A., TSAI, J., BANSAAL, A., PAOLELLA, L., HERBERT, B.R., SOORANNA, S.R., GOUGH, M.G., ASTLEY, C., VOGEL, K., BALDESSARI, A.E., BAMMLER, T.K., MACDONALD, J., GRAVETT, M.G., RAJAGOPAL, L., JOHNSON, M.R. (2015). Uterine overdistention induces preterm labor mediated by inflammation: observations in pregnant women and nonhuman primates. *American Journal of Obstetrics and Gynecology*, 213(6), 830.e1–830.e19.

WANG, Y., YANG, X., ZHENG, Y., WU, Z.H., ZHANG, X.A, LI, Q.P. HE, X.Y., WANG, C.Z., FENG, Z.C. (2013) The SEPS1 G-105A polymorphism is associated with risk of spontaneous preterm birth in a Chinese population. *PLoS One.* 8(6):E65657.

WEI, W., WANG, H., JI, S. (2017). Paradoxes of the EphB1 receptor in malignant brain tumors. *Cancer Cell International*, 17, 21.

WILSON, J.M.G. & JUNGNER, G. (1968). Principles and practices of screening for disease. Public Health Papers No. 34. World Health Organization; Geneva, Switzerland. Available from: http://whqlibdoc.who.int/php/WHO_PHP_34.pdf

WISHART DS, FEUNANG, Y.D., MARCU, A., GUO, A.C., LIANG, K., VÁZQUEZ-FRESNO, R., SAJED, T, JOHNSON, D., LI, C., KARU, N., SAYEEDA, Z., LO, E., ASSEMPOUR, N., BERJANSKII, M., SINGHAL, S., ARNDT, D., LIANG, Y., BADRAN, H., GRANT, J., SERRA-CAYUELA, A., LIU, Y., MANDAL, R., NEVEU, V., PON, A., KNOX, C., WILSON, M., MANACH, C., SCALBERT, A. (2018). HMDB 4.0 — The Human Metabolome Database for 2018. *Nucleic Acids Research*. 46(D1):D608-17.

WORLD HEALTH ORGANSATION (WHO) (2015) Country Statistics and Global Health Estimates by WHO and UN partners [cited 2016 01/03/2016]. Available from: http://www.who.int/gho/countries/gbr/pdf.

WORTHEY, E., MAYER, A., SYVERSON, G. HELBLING, D., BONACCI, B.B., DECKER, B., SERPE, J.M., DASU, T., TSCHANNEN, M.R., VEITH, R.L., BASEHORE, M.J., BROECKEL, U., TOMITA-MITCHELL, A., ARCA, M.J., CASPER, J.T., MARGOLIS, D.A., BICK, D.P., HESSNER, M.J., ROUTES, J.M., VERBSKY, J.W., JACOB, H.J., DIMMOCK, D.P. (2011) Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine* 13**,** 255–262

XI, X., LI, T., HUANG, Y., SUN, J., ZHU, Y., YANG, Y., LU, Z. (2017). RNA Biomarkers: Frontier of Precision Medicine for Cancer. *Noncoding RNA.* 3(1): pii: E9.

XIA, J. & WISHART, D.S. (2011). Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nature Protocols.* 6:743-760.

XIA, J. & WISHART, D.S. (2016). Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. *Current Protocols in Bioinformatics*. 55:14.10.1-14.10.91.

XU, C., LONG, A., FANG, X., WOOD, S.L., SLATER, D.M., NI, X., OLSON, D.M. (2013). Effects of PGF2alpha on the expression of uterine activation proteins in pregnant human myometrial cells from upper and lower segment. *The Journal of Clinical Endocrinology and Metabolism.* 98(7), 2975-83.

YEE, W.H., SORAISHAM, A.S., SHAH, V.S., AZIZ, K., YOON, W., LEE, S.K., CANADIAN NEONATAL NETWORK. (2012). Incidence and timing of presentation of necrotizing enterocolitis in preterm infants. *Pediatrics*. 129(2), e298-304.

YILMAZ, Y., VERDI, H., TANERI, A., YAZICI, A.C., ECEVIT, A.N., KARAKAŞ, N.M., TARCAN, A., HABERAL, A., OZBEK, N., ATAC, F.B. (2012) Maternal- fetal proinflammatory cytokine gene polymorphism and preterm birth. *DNA Cell Biology.* 31(1):92–7.

ZHANG, G., FEENSTRA, B., BACELIS, J., LIU, X., MUGLIA, L.M., JUODAKIS, J., MILLER, D.E., LITTERMAN, N., JIANG, P.P., RUSSELL, L., HINDS, D.A., HU, Y., WEIRAUCH, M.T., CHEN, X., CHAVAN, A.R., WAGNER, G.P., PAVLIČEV, M., NNAMANI, M.C., MAZIARZ, J., KARJALAINEN, M.K., RÄMET, M., SENGPIEL, V., GELLER, F., BOYD, H.A., PALOTIE, A., MOMANY, A., BEDELL, B., RYCKMAN, K.K., HUUSKO, J.M., FORNEY, C.R., KOTTYAN, L.C., HALLMAN, M., TERAMO, K., NOHR, E.A., DAVEY SMITH, G., MELBYE, M., JACOBSSON, B., MUGLIA, L.J. (2017). Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. *New England Journal of Medicine.* 377 (12):1156-1167.

ZHANG, L.X., SUN, Y., ZHAO, H., ZHU, N., SUN, X.D., JIN, X., ZOU, A.M., MI, Y., XU, J.R. (2017). A Bayesian Stepwise Discriminant Model for Predicting Risk Factors of Preterm Premature Rupture of Membranes: A Case-control Study. *Chinese Medical Journal.* 130(20):2416-22.

ZHANG, X., KRAMER, M.S. (2009). Variations in mortality and morbidity by gestational age among infants born at term. *Journal of Pediatrics.* 154: 358-62.

Appendices

Appendix A: Ethical Approvals

NRES Committee North West - Liverpool Central
3rd Floor
Barlow House
4 Minshull Street
Manchester
M1 3DZ

Telephone: 0161 625 7818

04 November 2011

Dr Andrew Sharp
Clinical Lecturer Obstetrics and Gynaecology
University of Liverpool
University Dept.
Liverpool Women's Hospital
Crown St.
L8 7SS

Dear Dr Sharp

Study title: **The development of novel biomarkers for the prediction of preterm labour in a high risk population**
REC reference: **11/NW/0720**

The Research Ethics Committee reviewed the above application at the meeting held on 02 November 2011. Thank you for attending to discuss the study.

Ethical opinion

The Chair welcomed you to the REC and thanked you for attending to discuss the study. You agreed to the presence of the observer for the discussion of the application.

The Committee advised that the Consent Form needed to be revised to allow for consent to send the samples outside of the UK. You agreed to do so and told the Committee that this may later change and the samples may be analysed in the UK. The Committee asked that the Participant Information Sheet be revised to include the fact that samples may be sent to Finland and to include the amount of blood that will be taken. You agreed to add the amount of blood in mls and teaspoon size.

The Committee asked at what point the linked anonymised samples would be de-linked. You stated that all patients are on the Viewpoint database and you will use the number from that initially. You also stated that you can generate a random number post recruitment and analysis. The Committee accepted this and pointed out that the link should only be maintained as long as necessary after delivery and analysis, especially as the samples are to be gifted. You agreed to break the link and keep only demographic information after the study and anonymise with a number.

You showed the Committee the glove to be used and told the members that it is not sterile but is the same as the probe the patients will have. You have included a low vaginal swab to cover all eventualities. You told the Committee that it is not likely that an infection would be introduced by the gloves. They are available in the UK and you will not include high risk patients in the study.

The members of the Committee present gave a favourable ethical opinion of the above

research on the basis described in the application form, protocol and supporting documentation, **subject to the conditions specified below.**

Ethical review of research sites

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" below).

Conditions of the favourable opinion

The favourable opinion is subject to the following conditions being met prior to the start of the study.

Management permission or approval must be obtained from each host organisation prior to the start of the study at the site concerned.

Management permission ("R&D approval") should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements.

Guidance on applying for NHS permission for research is available in the Integrated Research Application System or at http://www.rdforum.nhs.uk.

Where a NHS organisation's role in the study is limited to identifying and referring potential participants to research sites ("participant identification centre"), guidance should be sought from the R&D office on the information it requires to give permission for this activity.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of approvals from host organisations

Other conditions specified by the REC

a. The Committee would like to see the Participant Information Sheet revised to
   i) include in "What will happen to any samples I give?" Finland in the sentence beginning "The development of these tests"
   ii) Include the amount of blood which will be taken
b. The Committee would like to see the Consent Form revised to include a further point "I agree to my samples being sent outside of the UK for analysis"

It is responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

You should notify the REC in writing once all conditions have been met (except for site approvals from host organisations) and provide copies of any revised documentation with updated version numbers. Confirmation should also be provided to host organisations together with relevant documentation

Approved documents

The documents reviewed and approved at the meeting were:

Document	*Version*	*Date*
Covering Letter		03 October 2011
Investigator CV		
Letter from Sponsor		03 October 2011
Other: self assessment sheet for vaginal monitoring		
Participant Consent Form	1	12 September 2011
Participant Information Sheet	1	12 September 2011
Protocol	1	12 September 2011
REC application	3.3	07 October 2011

Membership of the Committee

The members of the Ethics Committee who were present at the meeting are listed on the attached sheet.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees (July 2001) and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review

Reporting requirements

The attached document "After ethical review – guidance for researchers" gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study

The NRES website also provides guidance on these topics, which is updated in the light of changes in reporting requirements or procedures.

Feedback

You are invited to give your view of the service that you have received from the National Research Ethics Service and the application procedure. If you wish to make your views known please use the feedback form available on the website.

Further information is available at National Research Ethics Service website > After Review

11/NW/0720	Please quote this number on all correspondence

With the Committee's best wishes for the success of this project

Yours sincerely

Professor Sobhan Vinjamuri
Chair

Email: carol.ebenezer@northwest.nhs.uk

Enclosures: *List of names and professions of members who were present at the meeting and those who submitted written comments*
"After ethical review – guidance for researchers"

Copy to: *Mrs Gillian Vernon*
Lindsay Carter

Appendix B: Ethical Amendment

1

NHS
Health Research Authority
National Research Ethics Service

NRES Committee North West - Liverpool Central
3rd Floor
Barlow House
4 Minshull Street
Manchester
M1 3DZ

Tel: 0161 625 7818
Fax: 0161 625 7299

20 May 2014

Dr Angharad Care
Obstetrics & Gynaecology Specialist Trainee
Liverpool Women's NHS Foundation Trust
Crown Street
Liverpool
L8 7SS

Dear Dr Care

Study title:	**The development of novel biomarkers for the prediction of preterm labour in a high risk population**
REC reference:	**11/NW/0720**
Amendment number:	**1**
Amendment date:	**02 May 2014**
IRAS project ID:	**86655**

To collect RNA for sampling, an additional tube for RNA analysis is required containing a specific medium. An additional 2mls of blood will be taken.
Change of Chief investigator from Dr Andrew Sharp to Dr Angharad Care

The above amendment was reviewed by the Sub-Committee in correspondence.

Ethical opinion

The Committee had no ethical issues with this amendment.

The members of the Committee taking part in the review gave a favourable ethical opinion of the amendment on the basis described in the notice of amendment form and supporting documentation.

Approved documents

The documents reviewed and approved at the meeting were:

Document	*Version*	*Date*
Covering letter on headed paper		02 May 2014
Notice of Substantial Amendment (non-CTIMP)	1	02 May 2014
Participant consent form	2.0	24 April 2014

2

Participant information sheet (PIS)	2.0	24 April 2014
Research protocol or project proposal	2.0	24 April 2014
Investigator CV Care		24 April 2014

Membership of the Committee

The members of the Committee who took part in the review are listed on the attached sheet.

R&D approval

All investigators and research collaborators in the NHS should notify the R&D office for the relevant NHS care organisation of this amendment and check whether it affects R&D approval of the research.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

We are pleased to welcome researchers and R & D staff at our NRES committee members' training days – see details at http://www.hra.nhs.uk/hra-training/

11/NW/0720:	**Please quote this number on all correspondence**

Yours sincerely

pp.

Mrs Julie Brake
Chair

E-mail:

Enclosures: *List of names and professions of members who took part in the review*

Copy to: *Ms Gillian Vernon, Liverpool Women's NHS Foundation Trust*

Appendix C: Co-Sponsorship

UNIVERSITY OF LIVERPOOL

Mrs Lindsay Carter
Research Coordinator
Legal, Risk & Compliance

Faculty Support Office
University of Liverpool
1st Floor
Duncan Building
Daulby Street
Liverpool
L69 3GA

Tel 0151 706 4523
Fax 0151 706 5668
Lindsay.Carter@liv.ac.uk

Faculty of Health and Life Sciences Ref: UoL000798

Dr Andrew Sharp
Department of Women's and Children's Health
Institute of Translational Medicine

Friday, 28 October 2011

Dear Dr Sharp

I am pleased to confirm that the University is prepared to act as Co-Sponsor with the Liverpool Women's NHS Foundation Trust under the Department of Health's Research Governance Framework for Health and Social Care (2005) for your study entitled "The Development of Novel Biomarkers for Predicting Preterm Birth". This approval for co-sponsorship is subject to the following:

1. The University expects you, as Chief Investigator, to conduct the study in full compliance with the requirements of the Framework so that it is able to meet its obligations as Co-Sponsor.

2. In addition to sponsorship, your study will require NHS ethical approval through the National Research Ethics Service (NRES). If you have not already done so, in order to apply for this please use the Integrated Research Application System (IRAS) at https://www.myresearchproject.org.uk/Home.aspx. Please contact me on 0151 706 4523 or at sponsor@liverpool.ac.uk) for further advice.

3. As the Chief Investigator, the University expects you to comply, where appropriate, with the University's policy on the use and / or storage of human tissues, details of which may be found at www.liverpool.ac.uk/humantissues.

4. An appropriate agreement should be completed with the Co-Sponsor which details the allocation of responsibilities between the Co-Sponsors. Please contact me for further advice.

   In addition to the above agreement, if you wish to conduct any part of the study in a site outside the UK, or the study requires the participation of a site other than one belonging to the Co- Sponsor, or you wish to sub-contract any part of the study to a third party (not including the Co-Sponsor) please contact me to ensure that appropriate contractual arrangements are in place.

5. University professional indemnity and study's insurances will apply to the study as appropriate. This is on the assumption that no part of the study will take place outside of the UK. Such cover will extend to cover for non-negligent harm.

6. A copy of the equivalent confirmation of co-sponsorship from the Co-Sponsor should be sent to me within 60 days of the date of this letter.

I trust that this statement will enable you to proceed with your research but if you have any queries please contact me on 0151 706 4523 (email sponsor@liverpool.ac.uk).

Yours sincerely,

Mrs Lindsay Carter
Research Coordinator, Faculty of Health and Life Sciences

Cc Head of Institute, Institute of Translational Medicine

Liverpool Women's NHS
NHS Foundation Trust

Crown Street
Liverpool
L8 7SS

Tel: 0151 708 9988
www.lwh.nhs.uk

Our ref: LWH0905
Date: 3rd October 2011

Dr. Andrew Sharp
University Department
Liverpool Women's NHS Foundation Trust
Crown Street
Liverpool
L8 7SS

Direct dial: 0151 702 4346
Direct fax: 0151 702 4299
Email: gillian.vernon@lwh.nhs.uk

Dear Dr. Sharp

Re: Developing novel biomarkers for the prediction of preterm labour

This letter is to confirm that Liverpool Women's NHS Foundation Trust (LWFT) intend to co-sponsor the proposed research detailed above. LWFT will take on sponsorship responsibilities as set out in the Research Governance Framework for Health & Social Care, second edition 2005, section 3.8. As sponsor the LWFT recognises its responsibility as set out in the attached appendix.

All other sponsor responsibilities will be undertaken by the University of Liverpool.

Kind Regards,

Mrs Gillian Vernon
R&D Manager

Liverpool Women's

Appendix D: Patient Information Leaflet

Liverpool Women's NHS
NHS Foundation Trust
Crown Street
Liverpool
L8 7SS
Tel: 0151 708 9988
www.lwh.nhs.uk

PTL Clinic - Patient information
Version:2.2
Date: 22.01.2016
Project ID number:11/NW/0720

The Development of Novel Biomarkers for Preterm Labour

Dr Angharad Care – Clinical Research Fellow, Liverpool Women's Hospital
Professor Zarko Alfirevic – Professor in Fetal & Maternal Medicine, Liverpool Women's Hospital

We are inviting many of the women who attend our unit to take part in a research study. Before you decide whether or not to take part it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. One of our team will go through the information sheet with you and answer any questions you may have.

Why are we doing the study?
The aim of this study is to develop earlier and safer ways of detecting problems in pregnancy. Research has shown that it may be possible to use a blood tests, vaginal swabs, urine or stool taken from the mother early in pregnancy to gather information which may help us work out why preterm birth happens and to predict the chances of preterm labour. We will do this by assessing a number of 'biomarkers' within these samples including different proteins, different types of bacteria and DNA. We will be taking samples at approximately 16 and 20 weeks. We want to identify why some women fail to respond to certain treatments for preterm labour, so we can work out which treatments are best for which women.

There is also evidence that women with less acidity (a higher pH) or certain types of bacteria in the vagina may be more at risk of preterm delivery. We wish to evaluate the ability of vaginal pH and types of bacteria to predict preterm birth or a short cervix. We will perform a vaginal swabs for pH and bacterial species check when you come to clinic at 16 weeks and 20 weeks of your pregnancy. This is not a swab for sexually transmitted diseases, if this is something you want this should be done at a local family planning or sexual health clinic.

There is a link between bacteria, infection and preterm birth but it is not clear how the bacteria cause preterm birth or how bacteria gets into the womb. Certain "pro-biotics" are being developed for good gut health and help with digestive problems. In this project we are also analysing bacteria in urine (wee) and stool (poo) on top of vaginal bacteria. If there are "bad" bacteria that are linked with preterm birth we may be able to design a "probiotic" to replace the bacteria in the tummy that may be causing preterm birth. We are also collecting stool samples which can be done at 16 and 20 weeks of pregnancy at home with our specially designed kits and posted to our lab.

It might be that a combination of these "biomarkers" might allow us to predict preterm birth, so our statisticians will be putting all this data together to see if combinations of this information that we find out about you can predict if women are protected from preterm birth or if women are at risk.

In the future this data may also help us target specific treatments to women to prevent preterm birth if we think we can work out why preterm birth may be happening.

Before we can offer new tests routinely it is important to ensure they work well and are accurate. To do this we need the help of women who are at high risk of preterm labour, either because of a previous preterm birth or cervical surgery.

Why have I been chosen?
We are inviting all women over 18 years of age attending the preterm labour clinic to take part in the study.

Do I have to take part?
It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and asked to sign a consent form. If you decide to take part you are free to withdraw at any time without giving a reason. If you do not feel able to take part it will not in any way affect the care you or your family receives.

What will happen to me if I take part?
If you agree to take part we will ask you to donate additional blood samples; 24ml (5 teaspoons) of blood will be taken at 16 weeks and 20 weeks at your PTL clinic appointments. We will arrange for you to have vaginal swab at the time of your internal scan to screen for bacterial vaginosis and provide you with treatment if required. As an optional part of the study, we will ask you provide a urine sample and stool sample which can be done either at clinic or at home with our specially designed kits and posted to the lab. We will not be testing for STI's, genetic disease or paternity on any of these samples.

Will my taking part in this study be kept confidential?
Yes. We will follow ethical and legal practice and all information will be handled in confidence. Any information you give us will only be used by the research team in the course of the research to develop these new tests. Any samples and data stored will be stored securely. They will be coded, and no personal data (name and address) will be available to the researchers. However, if DNA analysis provides clinically relevant information we will inform your medical doctor

What are the possible benefits of taking part?
The results of this research will not be available in the course of your pregnancy and will not directly benefit you. We hope that the results of the study overall will enable us to improve antenatal care provided to women by developing safer prenatal tests that will help us detect pregnancies at risk of preterm birth. We will ensure that your doctor is informed of any progress that means these new tests could be available for you in future pregnancies.

What are the possible disadvantages and risks of taking part?
Providing samples for research will add on additional time to your clinic appointments. An appointment may last for up to 45 mins. Blood taking can be uncomfortable and this will be carried out by someone who is skilled in venepuncture (taking blood). Some people may experience bruising at the site which will resolve over a few days. Ultrasound has an excellent safety record and will not harm your baby.

What will happen if I don't want to continue in the study?
You are free to withdraw at anytime. If you withdraw from the study we will not access any further samples and will destroy any of your samples that were collected for the study.

What will happen to any samples I give?
Samples will be collected specifically for this study. The samples will be coded and no personal data (name and address) will be stored with the sample. The development of these tests will be done in laboratories in the UK, Finland and the Germany. We also ask whether you would be willing to gift samples to be used for other ethically approved research studies into pregnancy problems. Your samples will only be used in research studies designed to develop these new methods for the early detection of pregnancy complications.

What will happen to the results of the research study?
The results from our project will be published as research papers in medical journals. No data will be published that will allow individuals to be identified.

Where can I get further information or discuss any problems?
Please contact a member of the fetal centre on 0151 702 4608 to discuss any questions or worries about the study, or if you have any complaints. If your concerns are not resolved, please contact Patient Advisory Liaison Services (PALS) on 0151 702 4353, if you have any concerns regarding the care you have received, or as an initial point of contact if you have a complaint. You can also visit PALS by asking at the hospital reception.

Who is organising and funding the research?
This research is organised by the Department of Women's and Children's Health, University of Liverpool.

Who has reviewed the study?
All research in the NHS is looked at by independent group of people, called a Research Ethics Committee, to protect your interests. This study has been reviewed and given favourable opinion by NRES Committee North West - Liverpool Central **NRES** Committee.

Thank you for taking the time to read this information leaflet.

Appendix E: Consent Forms

THE UNIVERSITY of LIVERPOOL

Liverpool Women's NHS
NHS Foundation Trust

Study Number:
Patient Identification Number for this trial:

CONSENT FORM

Title of Project: **Developing novel biomarkers for the prediction of preterm labour**

Name of Researcher:
Dr Angharad Care, Clinical Research Fellow, University of Liverpool/Liverpool Women's Hospital
Dr Andrew Sharp, Clinical Lecturer, University of Liverpool/Liverpool Women's Hospital
Prof Zarko Alfirevic, Professor of Obstetrics, University of Liverpool/Liverpool Women's Hospital

Please initial box

1. I confirm that I have read and understand the information sheet dated … ☐

(version ……) for the above study and have had the opportunity to ask questions.

2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason, without my medical care or legal rights being affected. ☐

3. I give permission for my medical notes and data collected during this research to be looked at by responsible individuals from the University of Liverpool & Liverpool Women's Hospital or from regulatory authorities where it is relevant to my taking part in research. ☐

4. I agree for DNA/RNA testing on my blood samples. ☐

5. I give permission for the researcher to contact my medical practitioner in the event of clinically significant findings from this research. ☐

6. I give permission for my samples to be sent outside of the UK ☐

7. Once we carry out the study on the samples you kindly donate, if there is any surplus sample it will be stored in the Liverpool Women's Research Tissue Bank, of which the University of Liverpool is the custodian. ☐

Name of Patient	Date	Signature
Name of Person taking consent (if different from researcher)	Date	Signature
Researcher	Date	Signature

1 for patient; 1 for researcher; 1 to be kept with hospital notes

Appendix F: Standard Operating Procedure for Sample Collection

The development of novel biomarkers for the prediction of preterm labour in a high risk population

Standard Operating Procedure for Sample Identification, Collection, Processing and Storage.

The purpose of this document is to provide instruction on correct sample identification, collection, processing and storage for the precision medicine for preterm birth prevention project.

Sample Identification – High Risk (150)

Women who are eligible for recruitment must have:

- Singleton pregnancy
- History of previous preterm delivery >16 weeks and <34 weeks / history PPROM >16 weeks - <34 weeks
- Able to provide informed consent

Sample Identification – Low Risk (250)

- Singleton pregnancy
- Previous uncomplicated term birth
    - (Exclude if e.g. gestational diabetes, pre-eclampsia, abruption, IUGR, fetal anomaly, medical condition necessitating iatrogenic preterm delivery)

Exclude women who have:

Uterine anomaly
Fetal congenital anomaly
History of "care-giver" indicated (iatrogenic) preterm birth

Visit 1 – 16 weeks (sample A)

Patient should have samples collected **16+0-16+6 weeks of gestation.** (For practical purposes samples will be accepted between 15^{+0}-18^{+0})

Samples required:

- Blood
- Vaginal Swabs
- Urine (optional)
- Stool (optional)

Visit 2 – 20 weeks (sample B)

Patients should give samples between **20+0-20+6** weeks of gestation (for practical purposes samples will be accepted between 19^{+0}-22^{+0}).

Samples required:

- Blood
- Vaginal Swabs
- Urine (optional)
- Stool (optional)

SAMPLING

1 . Maternal Blood

Materials needed:
Transport Container
Coolbox with ice
BD Vacutainer® Eclipse™ blood collection needles
6ml BD vacutainer® tubes containing clot activator for biomarkers (Red tube)
6ml BD vacutainer® K_2EDTA for maternal genome (Lavender small)
10mL BD vacutainer® K_2EDTA for plasma (Lavender large)
2ml PAXgene RNA tube for RNA (PAXgene, red top)
Sharps bin
Alcowipe
Cotton wool
Small plaster
Tourniquet
Trigene spray
Virkon 2% solution
Waste
Pipette tips
7ml bijou
Pastettes
1ml cryotubes (approx. 9)

1. Place a small polystyrene box with crushed ice or cool block into a transport container.

2. Venepuncture will be performed with BD Vacutainer® Eclipse™ blood collection needles. 24ml maternal blood will be obtained at 16 (sample A) and 20 (sample B) week visits by staff trained in venepuncture at the preterm labour clinic. Bottles required for the study are listed above.

3. Skin should be prepared with an alcowipe prior to blood sampling. Cotton wool can be used to apply pressure once the needle is withdrawn post sampling, prior to putting a plaster over the puncture site if required.

4. Immediately after blood has been taken gently invert all tubes 3-5 times to mix with any reagent in the vacutainer.

5. Label all specimens with study number, letter to indicate sample visit (e.g. 1A), "PTB study", date and time of sample collection.

   e.g. 1A PTB study
   1/1/2015 14:00

6. 10mL and 6 mL BD vacutainer® K_2EDTA (largeand small lavender top tube) should be stored on crushed ice immediately following labelling and arrive at lab within 90 minutes of collection, but as soon as possible.

7. 6ml BD vacutainer® tubes (red top) and 2ml PAXgene RNA tube (clear/red) should be stored at room temperature until transfer to the lab.

8. On arrival to laboratory 0618, log the samples in the Sample Reception Log (black file)

9. The proprietary reagent in the PAXgene Blood RNA tubes stabilizes intracellular RNA in collected blood specimens for 3 days at room temperature (15–25°C). Therefore from time of arrival to lab they are stored at room temperature for 2 hours standing.

10. Turn the centrifuge on. Set the centrifuge speed to 3000rpm, the time to 10mins, temperature to 4°C.

11. Place 6ml BD vacutainer® with clot activator (Red tube) tubes and 10mL BD vacutainer® K_2EDTA (lavender large) in the bucket inserts in the centrifuge ensuring samples are balanced by both volume and position. Using water filled blood bottles next to the centrifuge on left hand side if necessary for balance. The 6mL BD vacutainer® K_2EDTA for maternal genome (Lavender small) and 2ml PAXgene RNA tube should **NOT** be centrifuged.

12. Place clear lid over rotor buckets in use. Close the lid and press start.

13. Ensure rotor buckets and centrifuge basin are cleaned with ethanol and dried after use. Switch off the centrifuge and leave lid open. Complete user log.

14. Turn on safety hood.

15. Spray work surface with 1:100 Trigene spray and wipe with paper towels before starting.

16. Prepare all equipment required. i) Partly fill one waste container (approximately one eighth full) with Virkon 2% solution. ii) polystyrene cold block (located to left of hood) iii) Pastettes iv) Sharps bin (already in hood), cooled cryotube holder (located top drawer LEC freezer in Prep lab), 1000ml Gilson pipette and racked (sterile autoclaved) blue tips.

17. Carefully transfer centrifuged blood to hood. Remove lid from serum (red top) blood tube and aspirate serum layer into 7ml bijou using a pastette. Eject any residue from pastette into waste container containing Virkon. Transfer 1ml aliquot into 1ml cryotubes using Gilson pipette (approx. 2-3 x 1ml cryotube per sample), any remaining serum into 500µl aliquots. Any residual serum will be stored if a minimum of 250µl. Fill blood tube with Virkon and replace the lid securely prior to disposal in sharps bin.

18. Remove lid from 10ml BD vacutainer® K_2EDTA. Using fresh pastette aspirate plasma layer into 7ml labelled bijou. Pipette plasma in 1ml aliquots into 1ml cryovials.

19. Using a fresh pastette, aspirate 2ml sample of packed red blood cells with buffy coat into 1ml cryotube .

20. Eject any residue from pastette into waste container containing Virkon. Fill blood tube with Virkon and replace the lid securely prior to disposal in sharps bin.

21. Remove cold block containing samples from hood. Ensure all samples correctly labelled with unique study number, sample A, sample B or sample C, date of collection and substance (plasma, serum, or RBC) with sample number (e.g. serum 1, serum 2 etc).

22. Remove equipment once finished. Clean hood surface with Trigene and switch off hood. Allow waste products to stand for 24hours for disinfection prior to disposal.

23. Store labelled cryotubes in freezer at -80°C and log location of cryotubes under "PTL Biomarker study" (study 12) in freezer log.

24. The PAX gene blood RNA tube will be moved to the -20 freezer for 24hours and then moved to the -80 freezer until RNA extraction.

25. The 6ml vacutainer® K_2EDTA is stored in the freezer in the original tube labelled with date of collection, unique study number, sample A or B and study name. The box is located in freezer 4. Shelf 3. Enter into freezer log (study 12).

26. Upload sample and demographic information to MACRO samples database.

Appendix G: Standard Operating Procedure for Quantifying DNA using PicoGreen Reagent Kit

STANDARD OPERATING PROCEDURE
QUANTIFYING DNA USING PICOGREEN™ REAGENT KIT

Consumables

Description	Supplier	Catalogue Number
Quant-iT™ PicoGreen® dsDNA Assay Kit (2000 assays) (10 x 100 µL)	**Invitrogen**	**P11496**

Reagents
1xTE buffer (10mM Tris-HCl, 1mM EDTA, pH7.5) – must be nucleic acid free.

Preparing the Assay Buffer

Prepare a 1X TE working solution by diluting the concentrated buffer 20-fold with sterile, distilled, DNase-free water.

For a complete 96 well plate add 2.5ml of 20X TE to 47.5ml of DNase-free water.

Preparing the DNA Standard Curve

Dilute the lambda DNA standard (100µg/ml), provided in the Quant-iT™ 50-fold in 1X TE to make a 2 µg/ml working solution.

To prepare sufficient for the standard curve, pipette 5µl of the 100 µg/ml dsDNA standard and 245µl of 1X TE into a sterile eppendorf tube. Briefly mix by pipetting up and down.

To create a five-point standard curve from 1 ng/ml to 1 µg/ml, dilute the 2 µg/ml DNA stock solution in sterile eppendorf tubes as shown in the table below. Briefly mix the standards by pipetting up and down.

Volume of 1X TE (µl)	Volume of 2 µg/ml DNA stock (µl)	[DNA Standard]
0	220	2 µg/ml
228	22	200 ng/ml
247.8	2.2	20 ng/ml
999	1	2 ng/ml
220	0	0 ng/ml

Unused 1 ng/ml standard and 2 µg/ml DNA stock can be stored with the Quant-iT™ PicoGreen® kit for further assays.

Add 100µl of each DNA standard to the wells of a 96-well plate according to the following wells:

[DNA Standard] (ng/ml)	Wells
2 µg/ml	A1 + A2
200 ng/ml	B1 + B2
20 ng/ml	C1 + C2
2 ng/ml	D1 + D2
0 ng/ml	E1 + E2

Sample Analysis

For samples measuring under 200 ng/µl by nanodrop, add 1ul to 99 ul 1xTE (1/100 dilution).

For samples measuring over 200 ng/µl by nanodrop, add 1 to 9 ul 1xTE (1/10 dilution) and add 1ul diluted solution to 99 ul 1xTE (total 1/1000 dilution).

Add 1xTE followed by your DNA samples to the wells of a 96-well plate according to your plate plan.

Preparing the Reagent

On the day of the experiment, allow the Quant-iT™ PicoGreen® reagent to warm to room temperature before opening the vial.

In a plastic container, prepare a diluted working solution of the Quant-iT™ PicoGreen® reagent by making a 200-fold dilution of the concentrated solution in 1X TE.

For a 96-well plate, prepare enough diluted solution to assay 100 samples in a 100 µl final volume per well, add 50µl of Quant-iT™ PicoGreen® dsDNA reagent to 9.95ml of 1X TE.

Protect the working solution from light by covering with foil or placing in the dark. Use this solution within a few hours of its preparation.

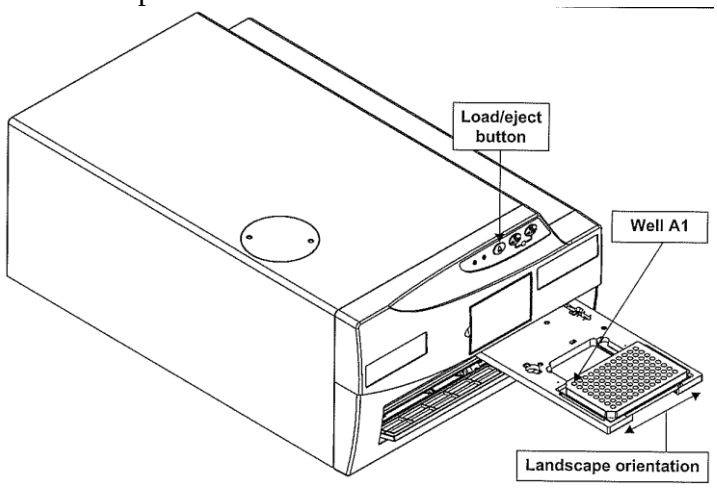
In reduced lighting add 100µl of the aqueous working solution of PicoGreen® reagent to each sample and DNA standard.

Incubate the plate for 2 to 5 minutes at room temperature, protected from light.

Measure the sample fluorescence using the Beckman Coulter DTX880, Warehouse Building; Room A212.

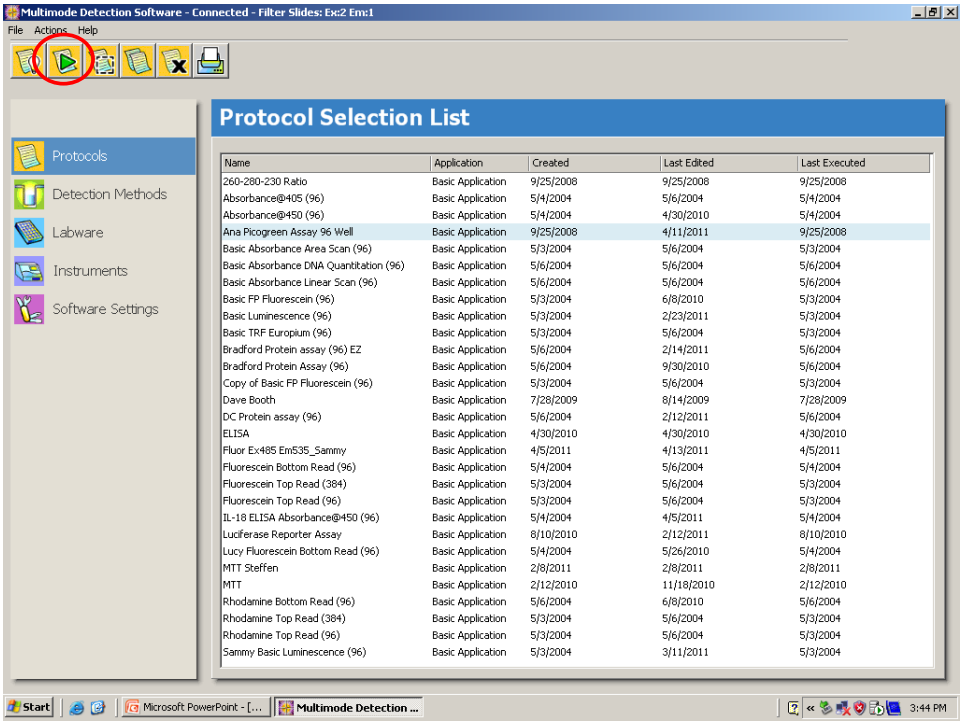
Using the Beckman Coulter DTX880 Multimode Detector

Turn on the machine using the switch at the back. Press the "load/eject" button the front of the machine and place the assay plate on the tray that emerges ensuring that well A1 is in the left-rear position as shown below:

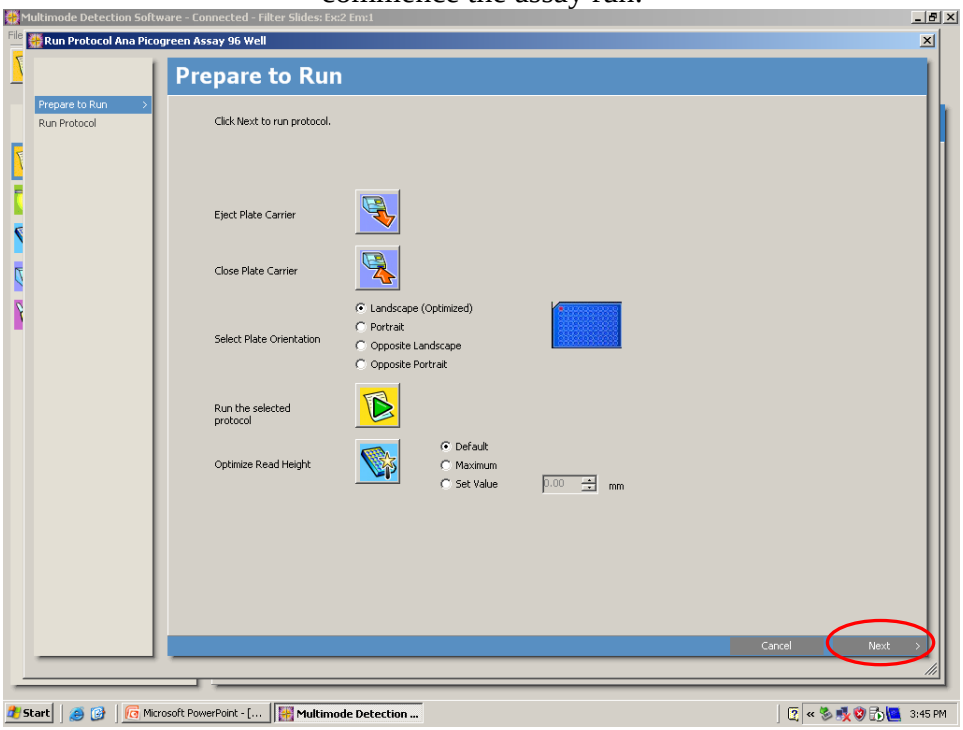


Press the "load/eject" button to send the tray back in. On the desktop of the PC (shown below), select the "Multimode Detection" software as highlighted.

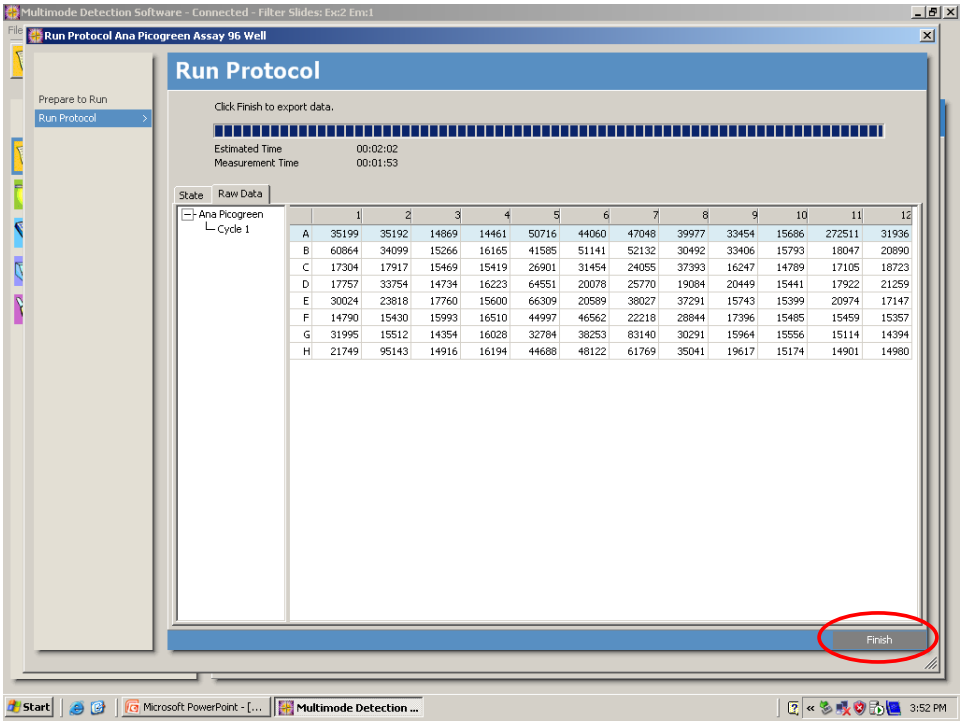
Select "Protocols" and highlight "Ana Picogreen assay 96 Well". Click on the "Run Protocol" icon (highlighted below):



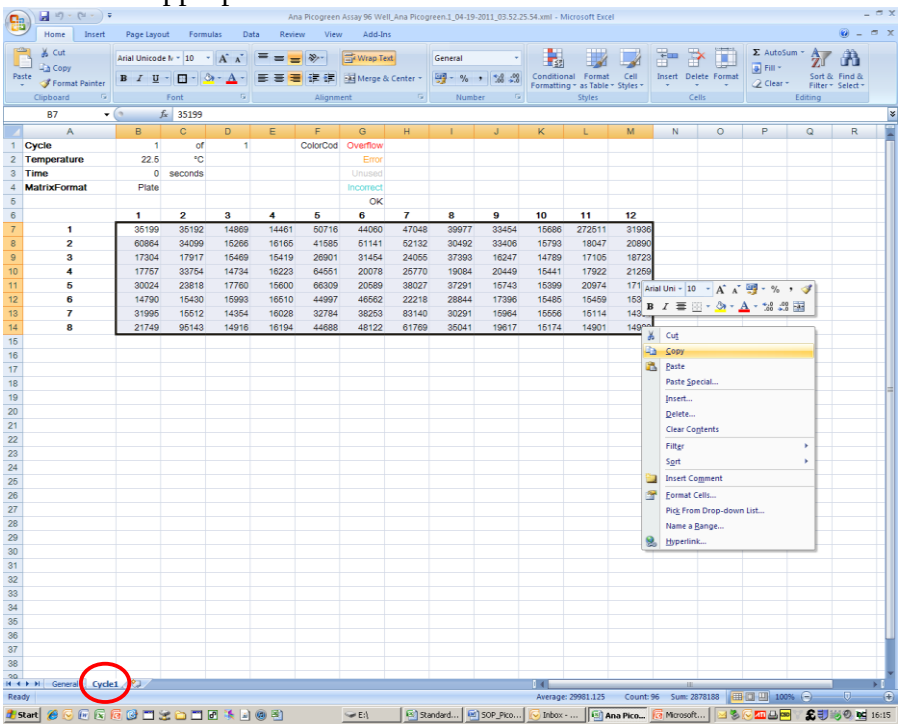
On the “Prepare to Run” screen (below) click “Next” (highlighted below) to commence the assay run:



After approximately 3 minutes the run will complete and the following screen will appear:

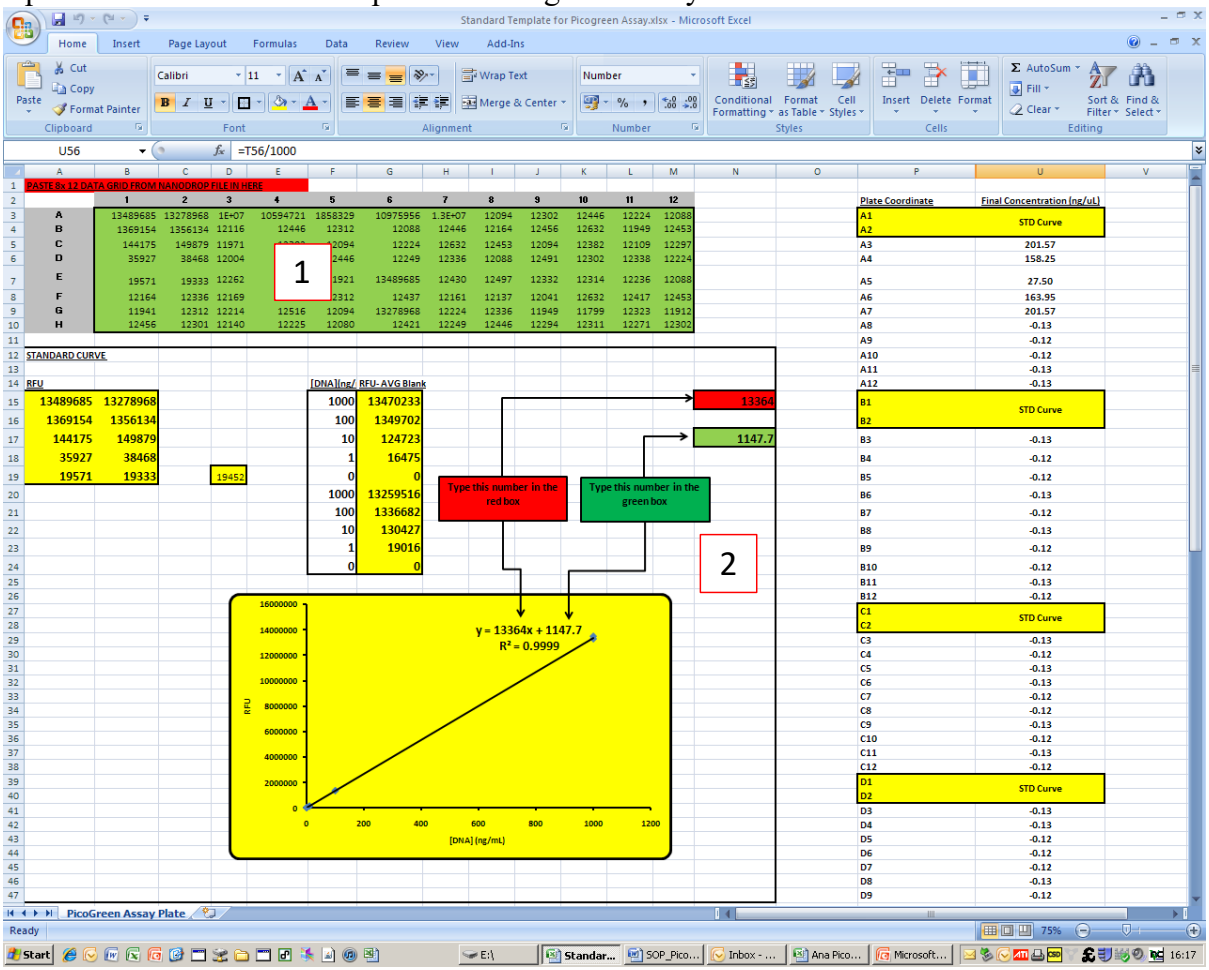


Click finish (highlighted) and an Excel spreadsheet of your data will be opened. Save this file to an appropriate folder.



In the “Cycle 1” tab (highlighted), select and copy the 8x12 grid of numbers representing your raw data as shown above.

Open the file “Standard Template for Picogreen Assay.xlsx”.



Paste the raw data into the green box in the top left (1). As instructed, type the figures from the trend line equation into the Red and Green boxes (2). Your unknown sample concentrations will appear in “Column U” with the associated Well coordinate in “Column P”

Appendix H: GWAS QC protocol

From UK Biobank Axiom array chip raw data is saved in PLINK file formats. Originally these are stored in .ped and .map PLINK files, but they may be transferred in binary PLINK file format (.bed, .bim and .fam files).

GWAS QC protocol

[Key: yellow highlight = file names; can be changed
Blue highlight = subfolder the files are saved in (after this full protocol has been carried out)
Green highlight = name of covariate in analysis; can be changed]

Using binary files (ptb.bed, ptb.bim and ptb.fam). Add sex information to original fam file (i.e. 2 for females) to all samples in fam file (open up in excel, keep FID, IID and replace column E with 2 for all samples, then remove columns C, D and F). Save at text file, then use following command to create an updated fam file:

```
plink --noweb --bfile ptb --update-sex ptb_gender.txt --make-bed --out ptb_gender
```

Resulting files are saved in subfolder: 1_ptb_gender

1) Remove duplicate

No duplicate samples present.

2) Gender check

```
plink --noweb --bfile ptb_gender  --check-sex --out ptb_gender_sex_check
```

Generates .sexcheck file.

```
grep PROBLEM ptb_gender_sex_check.sexcheck > fail_sex_check_ptb.txt
```

Produce a text file of samples that have failed sex check - first table of patients to exclude, include FID and IID (GWAS write-up).

Resulting files are saved in subfolder: 2_ptb_gender_sex_check

3) Missing data

```
plink --noweb  --bfile ptb_gender --missing  --out ptb_missing
```

This produces imiss and lmiss files.

Resulting files are saved in subfolder: 3_ptb_missing

4) Remove non-autosomes

Generate a list of nonautosome SNPs by filtering original study bim file for chromosomes: 23, 24, 25 and 26. Make a text file of these SNPs/rs IDs (ptbnonautosomes.txt) to be removed.

```
plink --noweb --bfile ptb_gender  --exclude ptbnonautosomes.txt  --make-bed  --out ptb_gender_ex
```

Resulting files are saved in subfolder: 4_ptb_gender_ex

5) Heterozygosity

```
plink --noweb --bfile ptb_gender_ex  --het --out ptb_gender_het
```

Use script to produce imiss vs het plot - can adapt to make it for 98% call rate. To use R scripts to use command: r_submit (script filename).
File **imiss_script.R** contains:

```
# IMISS-vs-Het plot for all SNPs, 95% Call rate:
imiss=read.table("ptb_missing.imiss",h=T)
imiss$logF_MISS = log10(imiss[,6])
het=read.table("ptb_gender_het.het",h=T)
het$meanHet = (het$N.NM.-het$O.HOM.)/het$N.NM.
colors  <- densCols(het$meanHet,imiss$logF_MISS)
png("ptb_imiss-vs-het.png")
plot(het$meanHet,imiss$logF_MISS,col=colors,xlim=c(0,0.3),ylim=c(-
3,0),pch=20,xlab="Heterozygosity rate",ylab="Proportion of missing
genotypes",main="PTB pilot imiss-het",axes=F)
axis(1,at=c(0,0.05,0.10,0.15,0.2,0.25,0.3),tick=T)
axis(2,at=c(-3,-2,-1,0),labels=c(0.001,0.01,0.1,1))
abline(v=mean(het$meanHet)-(3*sd(het$meanHet)),col="GREEN",lty=2)
abline(v=mean(het$meanHet)+(3*sd(het$meanHet)),col="GREEN",lty=2)
abline(v=mean(het$meanHet)-(5*sd(het$meanHet)),col="RED",lty=2)
abline(v=mean(het$meanHet)+(5*sd(het$meanHet)),col="RED",lty=2)
abline(h=-1.301030, col="RED", lty=2)
dev.off()
```

If there are any patient outliers then this has to be recorded in a table for write-up.

Resulting files are saved in subfolder: 5_ptb_gender_het

6) IBD

```
plink --noweb --bfile ptb_gender_ex --indep-pairwise 50 5 0.2 --out ptb_gender_ex_thin
```

This command will produce .prune.in and .prune.out files.

```
plink --noweb --bfile ptb_gender_ex --extract ptb_gender_ex_thin.prune.in --genome --out ptb_gender_ex_thin.genome
```

This second command will produce a .genome file.

Use IBD script to produced histogram. Second table of patients to exclude for GWAS write-up (need to check F_MISS values from imiss file to decide which samples to remove form the pairs that are reported. Include PI_HAT and F_MISS values in the table).

File **ptb_ibd_plot.r** contains:

```
## Histogram of IBD in R:

GEN=read.table("ptb_gender_ex_thin.genome.genome",header=T,as.is=T)
png("ptb_ibd_histogram.png")
hist(GEN$PI_HAT,ylim=c(0,100),xlim=c(0,1.0),breaks=100,main="PTB: IBD
Estimation",xlab="Estimated mean pairwise IBD",ylab="Frequency")
dev.off()
```

To output a file with outliers use:
awk '$10 >=0.1875 {print}' ptb_gender_ex_thin.genome.genome > ptb_ibd_outliers

Resulting files are saved in subfolder: 6_ptb_gender_ex_thin.prune

7) Merge with HapMap3

Open ptb_gender_ex bim file. Keep all rs IDs, (delete top two rows and other columns). Create text file and save as snplist_ptb.txt. Use **HapMap3 bed, bim and fam** files (provided).

```
plink --noweb --bfile Hapmap3 --extract snplist_ptb.txt --make-bed --out Hapmap3_ptb
```

Open Hapmap3_ptb bim file in Excel and remove all columns but keep the rs IDs. Save as text file, snplist_Hapmap3.txt.

```
plink --noweb --bfile ptb_gender_ex --extract snplist_Hapmap3.txt --make-bed --out ptb_Hapmap3
```

```
plink --noweb --bfile Hapmap3_ptb --bmerge ptb_Hapmap3.bed ptb_Hapmap3.bim ptb_Hapmap3.fam --make-bed --out Hapmap3_ptb_merged1
```

If a missnp file has been created that means the strands needs flipping. Therefore use:

```
plink --noweb --bfile Hapmap3_ptb --flip Hapmap3_ptb_merged1.missnp --make-bed --out Hapmap3_ptb_flipped
```

If this works, then continue with merge (like in the previous step but with updated name of bfile and output only).

```
plink --noweb --bfile Hapmap3_ptb_flipped --bmerge ptb_Hapmap3.bed ptb_Hapmap3.bim ptb_Hapmap3.fam --make-bed --out Hapmap3_ptb_merged2
```

If that still does not work, exclude the missnp file from the previous flip attempt.

```
plink --noweb --bfile Hapmap3_ptb_flipped --exclude Hapmap3_ptb_merged2.missnp --make-bed --out Hapmap3_ptb_flipped_excl
```

Exclude .missnp from the ptb_Hapmap3 file.

```
plink --noweb --bfile ptb_Hapmap3 --exclude Hapmap3_ptb_merged2.missnp --make-bed --out ptb_Hapmap3_excl
```

Then use --bmerge to merge the previous output files with Hapmap3_ptb_flipped_excl. Both files should now have an equal number of SNPs.

```
plink --noweb --bfile Hapmap3_ptb_flipped_excl --bmerge ptb_Hapmap3_excl.bed ptb_Hapmap3_excl.bim ptb_Hapmap3_excl.fam --make-bed --out Hapmap3_ptb_merged_final_excl
```

Resulting files are saved in subfolder: 7_ptb_merged

8) Ethnicity

```
plink --noweb --bfile Hapmap3_ptb_merged_final_excl --indep-pairwise 50 5 0.2 --out Hapmap3_ptb_merged_thin_excl
```

```
plink --noweb --bfile Hapmap3_ptb_merged_final_excl --extract Hapmap3_ptb_merged_thin_excl.prune.in --genome --out Hapmap3_ptb_merged_thin_MDS_excl
```

```
plink --noweb --bfile Hapmap3_ptb_merged_final_excl --cluster --mds-plot 10 --out Hapmap3_ptb_merged_MDS10_excl --read-genome Hapmap3_ptb_merged_thin_MDS_excl.genome
```

Process output .mds file:

- Open up MDS10.mds file in excel to add ethnicities for patient samples. Use file **Hapmap3_populationID_502_CEU_CHB_JPT_YRI.txt** for Hapmap ethnicity information.
- Sort file by IID column, Z to A. Check this has been done correctly, otherwise the plot will not be correct. Add in columns: Population and Data label.
- Excel file should now contain: FID, IID, Population, Data label, Sol, C1, C2 to C10
- Population column:
    - order the Hapmap ethnicities in the same order (IID column, Z to A) as the output file
    - copy and paste the ethnicities into the population column from Hapmap
    - only the study samples should not have ethnicity inputs (i.e. leave blank)
- Data label column: show Eunice the plot without data labels, then label only the outliers identified
- Use script to produce a PCA plot or SPSS - check the graph is correct, if not, repeat sort column step
- The outliers identified should be used to produce the third table of patients to exclude (GWAS write-up).

SPSS: File > open > data (select file). Chart builder > scatterplot > drag on: y-axis (C1), x-axis (C2), colours (population) > point ID labels > select the column called data labels

Ethnicity outliers

Use mds file used to plot ethnicity graph. Sort C1 column smallest to largest. Highlight values that deviate from the general ethnicity trend e.g. patient 22 drops from -0.044 (C1) to -0.02 (C2), this would be an ethnic outlier. Also select outliers that are found between ethnicities (when ordered) and have values that deviate. [For ptb study need to select everything that is not CEU]. Repeat this step with C2 column using a different colour to highlight. Some values will be borderline, select in a third colour.

Plot a graph in Excel, by selecting the C1 and C2 column to see if the outliers are present or not in the ethnicity plot.

Samples to Exclude

Open the .het file in Excel and calculate the **het score** to determine the outliers due to the heterozygosity (samples outside the 3SD and 5SD thresholds on the het-imiss graph from earlier).

Het_score = (N(NM)-O(HOM))/N(NM)

Sort by the het scores column smallest to largest. Plot this as a graph.
Identify the lowest values (also compare back to imiss vs het plot to see which samples they are)

Create a text file of all samples to be excluded based on sex check, ethnicity, heterozygosity and IBD: excludesexhetIBDethnicity.txt. Only include FID and IID columns. This will be removed in the next step.
For ptb do not exclude any of the treated or excluded for clinical reasons patients yet. Do this after final QC step.

Resulting files are saved in subfolder: 8_ptb_ethnicity

9) Remove samples outliers

Exclude outliers from the original fam file which includes the nonautosomes, but has no duplicate samples.

```
plink --noweb --bfile ptb_gender --make-bed --out ptb_sampleqc --remove excludesexhetIBDethnicity.txt
```

Resulting files are saved in subfolder: 9_ptb_sampleqc

10) SNP QC

For ptb study use 95% threshold. (Using 98% (0.02) removes more SNPs). Set up a text file with FID, IID and phenotype (ptb column) for case-controls (1 = control; 2 = case (female): -9 = missing/exclusion). Name file as ptb_pheno.txt.

```
plink --noweb --bfile ptb_sampleqc --pheno ptb_pheno.txt --pheno-name ptb --geno 0.05 --hwe 0.000001 --maf 0.01 --make-bed --out ptb_finalqc
```

Resulting files are saved in subfolder: 10_ptb_finalqc

Statistical analysis

Use ANOVA for continuous data. Use Chi-square for binary data. Generated values will be used for results analysis.
Select variables to investigate. In Excel, format the data.

- 3 columns: IID, outcome (binary code, 0 or 1) and third column for variable.
- Code outcomes as 0 = term (normal outcome) and 1 = PTB (for PPROM and SPON, as outcome interest).

- Leave samples with no outcomes blank for outcomes and variable column (these samples will be excluded from phenotypes later).
- Independent variable (factor/x variable) is outcome. Dependent variable (y variable) is the variable under investigation, e.g. BMI.
- Run ANOVA [open data in SPSS > Analyze > compare means > one way ANOVA]
- Or run Chi-square [open data > Analyze > Descriptive Statistics > Crosstabs > (rows = outcome; column = cervix) > statistics > Chi-square and Phi and Cramer's V > continue > cells > Select Observed, Row, Column and Total > ok]
- For chi-square, report the **Pearson Chi-Square** value and the Asymp. Sig value in the same row (this tells us the significance).
-

Preparation for results analysis

Open ptb_finalqc.fam:

- Delete column 3 to 6
- Add column headers: FID, IID, C1, C2, cervix
- Add in cervix column/phenotype information (-9 for exclusion [for ptb all treated and excluded samples] or missing values), write cervix values (for binary variables use 0 or 1) for other samples. This file will be used for analysis.
- Enter C1 and C2 values from previous mds file (should have less samples in fam file as samples were excluded from qc)
- Save as text file ptb_covar_cervix.txt

Results analysis

Use ptb_pheno.txt produced in step **10)SNP QC**

```
plink --noweb --bfile ptb_finalqc --logistic --pheno ptb_pheno.txt --pheno-name ptb --covar ptb_covar_cervix.txt --covar-name C1,C2,Cervix --out ptb_cervix --hide-covar
```

This produces an assoc.logistic file.

Run command without cervix and rename output file.

```
plink --noweb --bfile ptb_finalqc --logistic --pheno ptb_pheno.txt --pheno-name ptb --covar ptb_covar_cervix.txt --covar-name C1,C2 --out ptb_PCA--hide-covar
```

This produces another assoc.logistic file.

Haploview

https://www.broadinstitute.org/haploview/downloads

- Plink file format
- Upload .logistic output file
- Select integrated Map info box
- Click 'ok'
- A table will appear in a new window. Click 'plot'.
- Another window will appear. In this window add name for the plot (study name).

- Select x-axis as chromosome. Y-axis as P (for p value). Y-axis scale should be –log10.
- Enter suggested threshold (1x10^-5) as 5
- Enter significance threshold (5x10^-8) as 7.3
- Right-click on Manhattan plot and save as png.

Identify any significant SNPs or SNPs with high p values form Manhattan plot. Open the ptb_cervix.assoc.logistic file in Excel and order by p value. Filter p-values <= 5x10^-4 (less than or equal to 0.0005) to view rsID of SNPs of interest. Order by chromosome number, this will tell you how many SNPs per chromosome are top SNPs. Chromosomes with more than 1 top SNP are of most interesting to investigate further.

Regional plots can be produced in order to further investigate specific SNP calls. Go back to original table window, select chromosome and specify marker (can type SNP ID rs). Alternatively, can use Locus Zoom.

Locus zoom

http://locuszoom.org/

- Select plot your data > single data > set for: PLINK data > upload ptb_cervix.assoc.logistic file
- Copy & paste rs ID of SNP of interest into SNP reference name.
- For genome build ensure the option selected is hg19/1000 Genomes Nov 2014 EUR
- Legend location: right.
- Click plot data.
- A pdf of the graph will automatically be downloaded from the server.
- Repeat for all relevant SNPs.

Resulting files are saved in subfolder: 11_ptb_analysis_graphs (consists of folders: locus_zoom_plots and PTB_study_GWAS_analysis, with graphs produced from haploview)

GWAS Pre-Imputation QC

1) Download latest perl script from Will Rayner to QC your PLINK data against the HRC reference panel: http://www.well.ox.ac.uk/~wrayner/tools/#Checking (**HRC-1000G-check-bim.pl**)

2) Generate a freq file of final qc binary files using this command:

```
plink --noweb --bfile ptb_finalqc --freq --out ptb_feq
```

Resulting files are saved in subfolder: 12_ptb_frq

3) Check your genetic variants against the HRC panel using the following command. Latest HRC panel is located in the shared bioinf1 directory (use this path name in the reference panel part of the following command): /ph-users/shared/Eunice/HRC/HRC.r1-1.GRCh37.wgs.mac5.sites.tab

Make sure you have a frequency file generated prior to executing this command:
perl HRC-1000G-check-bim.pl -b <bim file> -f <Frequency file> -r <Reference panel> -h

gwas_perl_command.sh file contains the following command:

perl HRC-1000G-check-bim.pl -b ptb_finalqc.bim -f ptb_feq.frq -r /ph-users/shared/Eunice/HRC/HRC.r1-1.GRCh37.wgs.mac5.sites.tab -h

```
script_submit gwas_perl_command.sh
```

The script will produce several .txt files.
Resulting files are saved in subfolder: 13_ptb_perl

4) Perform SNP QC using the PLINK2 script generated from the perl script run.

```
script_submit Run-plink.sh
```

This script will produce binary files per chromosome (e.g. ptb_finalqc-updated-chr1.bed).
Resulting files are saved in subfolder: 14_ptb_finalqc_updated

5) For each chromosome, convert binary PLINK into VCF format:

For chromosomes 1-22, use the "--recode vcf-iid --out" command in PLINK2

```
plink2_submit --bfile ptb_finalqc-updated-chr1 --recode vcf-iid --out ptb_chr1
```

For chr2-22 use commands:

```
plink2_submit --bfile ptb_finalqc-updated-chr2 --recode vcf-iid --out ptb_chr2
plink2_submit --bfile ptb_finalqc-updated-chr3 --recode vcf-iid --out ptb_chr3
plink2_submit --bfile ptb_finalqc-updated-chr4 --recode vcf-iid --out ptb_chr4
plink2_submit --bfile ptb_finalqc-updated-chr5 --recode vcf-iid --out ptb_chr5
plink2_submit --bfile ptb_finalqc-updated-chr6 --recode vcf-iid --out ptb_chr6
plink2_submit --bfile ptb_finalqc-updated-chr7 --recode vcf-iid --out ptb_chr7
plink2_submit --bfile ptb_finalqc-updated-chr8 --recode vcf-iid --out ptb_chr8
plink2_submit --bfile ptb_finalqc-updated-chr9 --recode vcf-iid --out ptb_chr9
plink2_submit --bfile ptb_finalqc-updated-chr10 --recode vcf-iid --out ptb_chr10
plink2_submit --bfile ptb_finalqc-updated-chr11 --recode vcf-iid --out ptb_chr11
plink2_submit --bfile ptb_finalqc-updated-chr12 --recode vcf-iid --out ptb_chr12
plink2_submit --bfile ptb_finalqc-updated-chr13 --recode vcf-iid --out ptb_chr13
plink2_submit --bfile ptb_finalqc-updated-chr14 --recode vcf-iid --out ptb_chr14
plink2_submit --bfile ptb_finalqc-updated-chr15 --recode vcf-iid --out ptb_chr15
plink2_submit --bfile ptb_finalqc-updated-chr16 --recode vcf-iid --out ptb_chr16
plink2_submit --bfile ptb_finalqc-updated-chr17 --recode vcf-iid --out ptb_chr17
plink2_submit --bfile ptb_finalqc-updated-chr18 --recode vcf-iid --out ptb_chr18
plink2_submit --bfile ptb_finalqc-updated-chr19 --recode vcf-iid --out ptb_chr19
plink2_submit --bfile ptb_finalqc-updated-chr20 --recode vcf-iid --out ptb_chr20
plink2_submit --bfile ptb_finalqc-updated-chr21 --recode vcf-iid --out ptb_chr21
plink2_submit --bfile ptb_finalqc-updated-chr22 --recode vcf-iid --out ptb_chr22
```

For chromosome 23, use the "--set-hh-missing --recode vcf-iid --out" command in PLINK2

```
plink2_submit --bfile ptb_finalqc-updated-chr23 --set-hh-missing --recode vcf-iid --out ptb_chr23
```

Important note! HRC does not recognise chr23 for imputation, make sure you change chr23 to chrX within your VCF file.

```
sed ‘s/^23/X/g’ ptb_chr23.vcf > ptb_chrX.vcf
```

Resulting files are saved in subfolder: 15_ptb_vcf (this folder contains the .log files for the commands run, but due to the next zip commands being run, the .vcf files have been converted to vcf.gz and are not in this subfolder as .vcf files).

6) Zip each VCF file using bgzip. Bioinf1 bgzip location: /users/apps/htslib/htslib-1.3.2/bgzip

```
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr1.vcf
```

```
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr2.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr3.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr4.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr5.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr6.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr7.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr8.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr9.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr10.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr11.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr12.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr13.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr14.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr15.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr16.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr17.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr18.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr19.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr20.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr21.vcf
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chr22.vcf
```

Command for ChrX:

```
/users/apps/htslib/htslib-1.3.2/bgzip ptb_chrX.vcf
```

Files produced with extension .vcf.gz
Resulting files are saved in subfolder: 16_ptb_vcf_gz

7) Tabix each bgzipped VCF file. Bioinf1 Tabix location: /users/apps/htslib/htslib-1.3.2/tabixb -p vcf

```
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr1.vcf.gz
```

```
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr2.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr3.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr4.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr5.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr6.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr7.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr8.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr9.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr10.vcf.gz
```

```
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr11.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr12.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr13.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr14.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr15.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr16.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr17.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr18.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr19.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr20.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr21.vcf.gz
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chr22.vcf.gz
```

Command for ChrX:

```
/users/apps/htslib/htslib-1.3.2/tabix -p vcf ptb_chrX.vcf.gz
```

Files produced with extension .vcf.gz.tbi
Resulting files are saved in subfolder: 17_ptb_vcf_gz_tbi

8) Upload your VCF files to the Michigan Imputation Server for imputation:
https://imputationserver.sph.umich.edu/index.html

Eunice logged into her account for this upload. Imputed files downloaded by Eunice.
Select the following options:

- Select latest HRC reference panel
- Use Eagle to phase chromosomes 1-22
- Use ShapeIT to phase chromosome X
- Specify population

Imputation can take around 2 weeks when server is moderately busy. Files will be available to download from the website, but has to be done within 2-3 days. These will be zip files per uploaded chromosome.
Resulting files are saved in subfolder: 18_ptb_imputation

GWAS post-imputation protocol

Preparation to unzip files from Imputation server

1) Download from https://sourceforge.net/projects/p7zip/?source=typ_redirect

Copy and paste file onto cluster where zipped, results files are.
Use command to unzip:

```
tar xvjf p7zip_16.02_src_all.tar.bz2
```

A file called p7zip_16.02 will be created in the cluster.
Use command:

```
cd p7zip_16.02
```

Open file for instructions: /ph-users/juhi/PTB_study/p7zip_16.02/README

2) From this weblink use the command: `make all3`
http://www.linuxfromscratch.org/blfs/view/svn/general/p7zip.html

Then use command:

```
make DEST_HOME=/usr \
        DEST_MAN=/usr/share/man \
        DEST_SHARE_DOC=/usr/share/doc/p7zip-16.02 install
```

File 7z should now be located in /ph-users/juhi/PTB_study/p7zip_16.02/bin

Password (generated by Imputation server) for chromosomes 1-22: SRObVBj8vTl7uW
Password (generated by Imputation server) for chromosome X: MBe3PfqEccwPb6

3) To extract files use command:

```
/ph-users/juhi/PTB_study/p7zip_16.02/bin/7z x -pSRObVBj8vTl7uW chr_1.zip
```

This will give 3 resulting files in the directory for that chr. Repeat this command for all chr including chr22.
For Chr X, change working directing to /ph-users/juhi/PTB_study/chrX
Then use command:

```
/ph-users/juhi/PTB_study/p7zip_16.02/bin/7z x -pMBe3PfqEccwPb6
chr_X.no.auto_female.zip
```

Resulting files are saved in subfolder: 19_ptb_imputation_files

Post-imputation QC steps

1) View .info.gz files without unzipping the files (they are too large to unzip). Use command:

```
less chr1.info.gz
```

Press q to quit.
To filter SNPs with score less than 0.3, use command:

```
gzip -dc chr1.info.gz | awk '$7<0.3 {print$1, $7}' > chr1_rsq_snps_toexclude.txt
```

Repeat this for all chromosomes. (Change working directory for chrX before running this command).

```
gzip -dc chrX.no.auto_female.info.gz | awk '$7<0.3 {print$1, $7}' >
chrX_rsq_snps_toexclude.txt
```

Resulting files are saved in subfolder: 20_ptb_rsq_snps

2) Convert vcf to plink files using plink2. Use command:

```
plink2_submit --vcf chr1.dose.vcf.gz --out chr1_plink
```

Repeat for all chromosomes.
Change directory for chr X and use:

```
plink2_submit --vcf chrX.no.auto_female.dose.vcf.gz --out chrX_plink
```

Resulting files are saved in subfolder: 21_ptb_plink

3) Generate frequency files. Use:

```
plink --noweb --bfile chr1_plink --freq --out chr1_freq
```

Repeat for all chromosomes.
Change directory for chr X and use:
plink --noweb --bfile chrX_plink --freq --out chrX_freq [problem reading BIM file line 1] No file produced for chrX
Resulting files are saved in subfolder: 22_ptb_frequency

4) Extract SNPS with maf= 0. Use:

```
awk '$5==0 {print$2, $5}' chr1_freq.frq > chr1_maf_toexclude.txt
```

Repeat for chr2-22.
Resulting files are saved in subfolder: 23_ptb_maf_toexclude

5) Combine SNP ID columns from rsq txt file and maf txt file. This is to add one list to another list in a new file (duplicates are not removed by this command). Use command:

```
awk '{print $1}' chr1_rsq_snps_toexclude.txt <(awk '{print $1}'
chr1_maf_toexclude.txt)  > chr1_excludesnps.txt
```

Repeat for chr2-22.
Resulting files are saved in subfolder: 24_ptb_excludesnps

6) To exclude snps from original vcf file. Use this command in a script not login node!! The file produced will be fairly large, so will take a while to run. Use command:

```
script_submit exc_snp_chr1.sh
```

Script contains the following command for chr1:

```
/ph-users/shared/Eunice/vcftools_0.1.13/bin/vcftools --gzvcf chr1.dose.vcf.gz
--exclude chr1_excludesnps.txt --recode --stdout
```

Use the following command in the login node to compress the stdout file (this is a temporary file that will get overwritten by next command) before running the next command!

```
/users/apps/htslib/htslib-1.3.2/bgzip -c stdout > chr1_qc_bgzip.vcf.gz
```

For chr2-22, use the --exclude command above in separate .sh script per chromosome. For the bgzip command type directly into login node.
Resulting files are saved in subfolder: 25_ptb_qc_bgzip

7) Use this command in the login node (separately per chr) for bgzip output:

```
/users/apps/htslib/htslib-1.3.2/tabix -p vcf chr1_qc.vcf.gz
```

Resulting files are saved in subfolder: 26_ptb_qc_tbi

SNP test v2.5

Use clinical variable with most significance as covariate for analysis. (Used continuous cervical length for 82 patients, excluding medical intervention.)

1) To prepare file for SNP test analysis (for this data will be **using frequentist association test**), make an excel sheet with the relevant headings and coding:

http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html

Save excel file in cluster and convert to space separated file. The file headers should look like this:

```
ID_1 ID_2 missing c1 c2 cvl ptb
0 0 0 C C C B
A01_1B PTB_1B 0 99.9532139 99.952438 32 0
A01_P06_97B PTB_97B 0 99.9538758 99.9520761 27 1
```

Convert using:

```
expand --tabs =1        .txt  >    .txt
```

Then use:

```
dos2unix samplefile_snptest.sample
```

Resulting files are saved in subfolder: 27_ptb_sample

2) Run command for **Frequentist Association Tests** (in a separate script for each chromosome):

```
/usr/local/bin/snptest_v2.5 -data chr1.dose.vcf.gz PTB_samplefile.sample -
genotype_field GT -o chr1_ptb_frequentist_gen_snptest_cervixlength.out -
frequentist 1 -method expected -pheno ptb -cov_names c1 c2 cvl -missing_code NA
-lower_sample_limit 20 -hwe -log chr1_snptest.log
```

Output file e.g. chr1_ptb_frequentist_gen_snptest_cervixlength.out

Resulting files are saved in subfolder: 28_ptb_freqtest

QC steps after SNPtest analysis

1) Delete first few lines of chr files. Then rename this to chr001.txt (must have extra zero).

For chr 1 file:

```
sed '/^#/d' chr1_ptb_frequentist_gen_snptest_cervixlength.out > chr001.txt
```

Delete all header lines from chr2 to chr 22 files and rename similarly.

For chr2 file onwards:

```
sed '/^#/d' chr2_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' >
chr002.txt
```

Repeat for chr3 to 22:

```
sed '/^#/d' chr3_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' >
chr003.txt
sed '/^#/d' chr4_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' >
chr004.txt
sed '/^#/d' chr5_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' >
chr005.txt
```

```
sed '/^#/d' chr6_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr006.txt
sed '/^#/d' chr7_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr007.txt
sed '/^#/d' chr8_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr008.txt
sed '/^#/d' chr9_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr009.txt
sed '/^#/d' chr10_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr010.txt
sed '/^#/d' chr11_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr011.txt
sed '/^#/d' chr12_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr012.txt
sed '/^#/d' chr13_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr013.txt
sed '/^#/d' chr14_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr014.txt
sed '/^#/d' chr15_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr015.txt
sed '/^#/d' chr16_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr016.txt
sed '/^#/d' chr17_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr017.txt
sed '/^#/d' chr18_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr018.txt
sed '/^#/d' chr19_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr019.txt
sed '/^#/d' chr20_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr020.txt
sed '/^#/d' chr21_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr021.txt
sed '/^#/d' chr22_ptb_frequentist_gen_snptest_cervixlength.out | sed '/^alternate_ids/d' > chr022.txt
```

To concatenate all chr files use the following command cat chr0*.txt > chr1_22_finalsnp.txt submit as script:

```
script_submit cat_snps.sh
```

Resulting files are saved in subfolder: 29_ptb_concat

2) To extract columns, submit as script:

```
script_submit extractcol.sh
```

Shell script **extractcol.sh** contains command:

```
awk '{print $3, $2, $4, $45}' chr1_22_finalsnp.txt > chr1_22_snpfile.txt
```

To rename headers, submit as script:

```
script_submit rename_header.sh
```

Shell script **rename_header.sh** contains command:

```
sed -e 's/chromosome/CHR/' chr1_22_snpfile.txt | sed 's/rsid/SNP/' | sed
's/position/BP/' | sed 's/frequentist_add_pvalue/P/'  >
chr1_22_snpfile_rename.txt
```

To remove NA pvalues, submit as script:

```
script_submit remove_na.sh
```

Shell script **remove_na.sh** contains command:

```
sed '/NA/d' chr1_22_snpfile_rename.txt > chr1_22_snpfile_update.txt
```

Resulting files are saved in subfolder: 30_ptb_update

3) Use R script to produce Manhattan plot (specify input filename and output filename in script):

```
r_submit manhattan_qqman_ptb.r
```

Script **manhattan_qqman_ptb.r** contains:

```
## This is a R script for generating your manhattan plot! Required
headers in your results file are CHR SNP BP P.
## Results file can be space or tab-limited.
library(qqman)
results <-read.table("chr1_22_snpfile_update.txt",header=T) ##Specify
filename containing CHR SNP BP P information.
png("ptb_manplot.png", width=1500, height=800, res=120) ##Specify
output filename.
manhattan(results)
dev.off()
```

To modify Manhattan plot e.g. change colours etc, the following linkas have suggestions on the required commands:
http://www.gettinggeneticsdone.com/2014/05/qqman-r-package-for-qq-and-manhattan-plots-for-gwas-results.html
http://www.gettinggeneticsdone.com/2011/04/annotated-manhattan-plots-and-qq-plots.html
Resulting files are saved in subfolder: 31_ptb_manhattan

Identify SNPs from Manhattan plot

Based on Manhattan plot, select chromosome of interest. Use following command to extract all SNPs from the specified chr. For example, for chr3 use the following commend in a script (**snps_chr3.sh**):

```
awk '$1==03 {print $1, $2, $3, $4}' chr1_22_snpfile_update.txt > snps_chr3.txt
```

Open up the text file in excel and sort by pvalue (filter by <=5x10^-6 or 0.000005; then sort smallest to largest) to identify the tophits per chromosome.
To get rsID for one SNP:

```
awk '($1 == "3") && ($2 == "134815715") { print $1, $2, $3 }' /ph-
users/shared/Eunice/HRC/HRC.r1-1.GRCh37.wgs.mac5.sites.tab > chr3_tophit.txt
```

Calculate plus and minus 500kb (500000b) of the snp position and search for the rsIDs in this region per hit. **Note** these upper and lower bp values may not correspond to actual chr positions/rsIDs/p-values, they are only being used to give a range of positions that exit.
To get rsIDs for range of SNPs around region of interest/top hit:

```
awk '($1 == "3") && ($2 <= "135315715" && $2 >= "134315715" ) { print $1, $2, $3 }' /ph-users/shared/Eunice/HRC/HRC.r1-1.GRCh37.wgs.mac5.sites.tab > chr3_tophit.txt
```

```
awk '($1 == "20") && ($2 <= "50449572" && $2 >= "49449572" ) { print $1, $2, $3 }' /ph-users/shared/Eunice/HRC/HRC.r1-1.GRCh37.wgs.mac5.sites.tab > chr20_tophit.txt
```

To get p values for SNPs within range of interest (for single digit chr number put 0 in front, to match snpfile):

```
awk '($1 == "03") && ($3 <= "135315715" && $3 >= "134315715" ) { print $2, $3, $4 }' chr1_22_snpfile_update.txt > chr3_pval_tophit.txt
```

```
awk '($1 == "20") && ($3 <= "50449572" && $3 >= "49449572")  { print $2, $3, $4 }' chr1_22_snpfile_update.txt > chr20_pval_tophit.txt
```

Resulting files are saved in subfolder: 32_ptb_topsnps

Prep locus zoom file

In Excel, open up the file containing p values for the region of interest (chr3_pval _tophit.txt) then paste in rsIDs column (from chr3 _tophit.txt). Highlight duplicated cells of the two columns containing the positions. Sort by colour (make sure you highlight the position column and the rsID column, so that the order is not lost). Delete the unique values (shift cells up).
Highlight duplicate values again but with a different colour e.g. green to pick out any repeated rsIDs (order as before, then delete highlighted cells). Delete the second duplicates (keep the ones that have rsIDs) as appropriate, find where the remaining rsID belong in the order (ctr+f) and paste in. Then check order of positions matches across the rows.
Then arrange columns as: chr pos rsID p and save file as txt file with ext locuszoomfile.txt

Locus Zoom

Submit rsID of the top hit identified. Use tab as the delimiter (also default). Upload locuszoomfile.txt and set marker column as rsID and p-value column as p. Flanking region is automatically set at 400kb. Genome build is automatically hg19/1000 Genomes Nov 2014 EUR (latest).

Resulting files are saved in subfolder: 33_ptb_locuszoom

Appendix I: R script used for pooled data analysis

```
library(meta)
library(rmeta)
setwd("Documents/Max Planck 27.07.2018/")
muglia <- read.table("pre_top10000.tab",h=T)
head(muglia)

lp <-
read.table("28_ptb_freqtest/chr3_ptb_frequentist_gen_snptest_cervixlength.out",h=T
)
lp2 <- lp[,c(1:6,29,45,47:48)]

comb <-merge(muglia,lp2,by.x = c("chr","pos"),by.y = c("chromosome","position"))
comb <- comb[with(comb,order(p)),]
head(comb)

snp <- comb[1:10,]
# snp <- comb[which(comb$snp %in%
c("rs201450565","rs200745338","rs3849531")),]
snp

snp$eff <- log(snp$eff)
snp$A2 <- sub("/.*","",snp$alleles) # effect
snp$A1 <- sub(".*/","",snp$alleles) # reference
snp$FLIP <- NA
snp[which(snp$A1==snp$alleleA & snp$A2==snp$alleleB),]$FLIP <- F
snp[which(snp$A1==snp$alleleB & snp$A2==snp$alleleA),]$FLIP <- T
snp[which(snp$FLIP),]$frequentist_add_beta_1 <- -
snp[which(snp$FLIP),]$frequentist_add_beta_1

snp$P_1s <- snp$frequentist_add_pvalue/2
snp[which(sign(snp$frequentist_add_beta_1)!=sign(snp$eff)),]$P_1s <- 1-
snp[which(sign(snp$frequentist_add_beta_1)!=sign(snp$eff)),]$P_1s

i <- 1
for(i in 1:nrow(snp))

  metaobj <-
metagen(TE=c(snp[i,]$eff,snp[i,]$frequentist_add_beta_1),seTE=c(snp[i,]$se,snp[i,]
$frequentist_add_se_1))
  meta <- metaobj
  meta <-
data.frame(t(c(meta$TE.fixed,meta$seTE.fixed,meta$pval.fixed,meta$lower.fixed,m
eta$upper.fixed)))
  colnames(meta) <-
c("META_BETA","META_SE","META_P","META_CI_l","META_CI_u")
  meta <- cbind(snp[i,],meta)
  if(i==1) res <- meta else res <- rbind(res,meta)
```

```
res$META_OR <- exp(res$META_BETA)
res$META_CI_l <- exp(res$META_CI_l)
res$META_CI_u <- exp(res$META_CI_u)

res$M_OR <- exp(res$eff)
res$M_CI_l <- res$M_OR-1.96*res$se
res$M_CI_u <- res$M_OR+1.96*res$se

res$LP_OR <- exp(res$frequentist_add_beta_1)
res$LP_CI_l <- res$LP_OR-1.96*res$frequentist_add_se_1
res$LP_CI_u <- res$LP_OR+1.96*res$frequentist_add_se_1

head(res)

write.table(res,"meta_muglia.txt",c=T,r=F,qu=F)
write.csv(res,"meta_muglia.csv",row.names=F)

# forest(metaobj)
for(i in 1:nrow(res))
  mmean <- c(NA,res[i,]$M_OR,res[i,]$LP_OR,res[i,]$META_OR)
  mcl <- c(NA,res[i,]$M_CI_l,res[i,]$LP_CI_l,res[i,]$META_CI_l)
  mcu <- c(NA,res[i,]$M_CI_u,res[i,]$LP_CI_u,res[i,]$META_CI_u)
  mtext <-
data.frame(cbind(c("Muglia","Liverpool","Meta"),c(res[i,]$p,res[i,]$frequentist_add
_pvalue,res[i,]$META_P)),stringsAsFactors=F)
  mtext$X2 <- format(as.numeric(mtext$X2),sci=T,dig=2)
  mtext <- rbind(c("Name","p-value"),mtext)
  msum <- c(T,F,F,T)
  sn <- paste0(as.character(res[i,]$snp),"_",res[i,]$A2)
  png(paste0(sn,".png"))

forestplot(mtext,mmean,mcl,mcu,is.summary=msum,xlab=sn,zero=1,col=meta.color
s(summary="#1F3D76",lines="grey30",box="grey30",zero="grey80"))
  dev.off()
```

Appendix J: Procedure for Manual RNA extraction

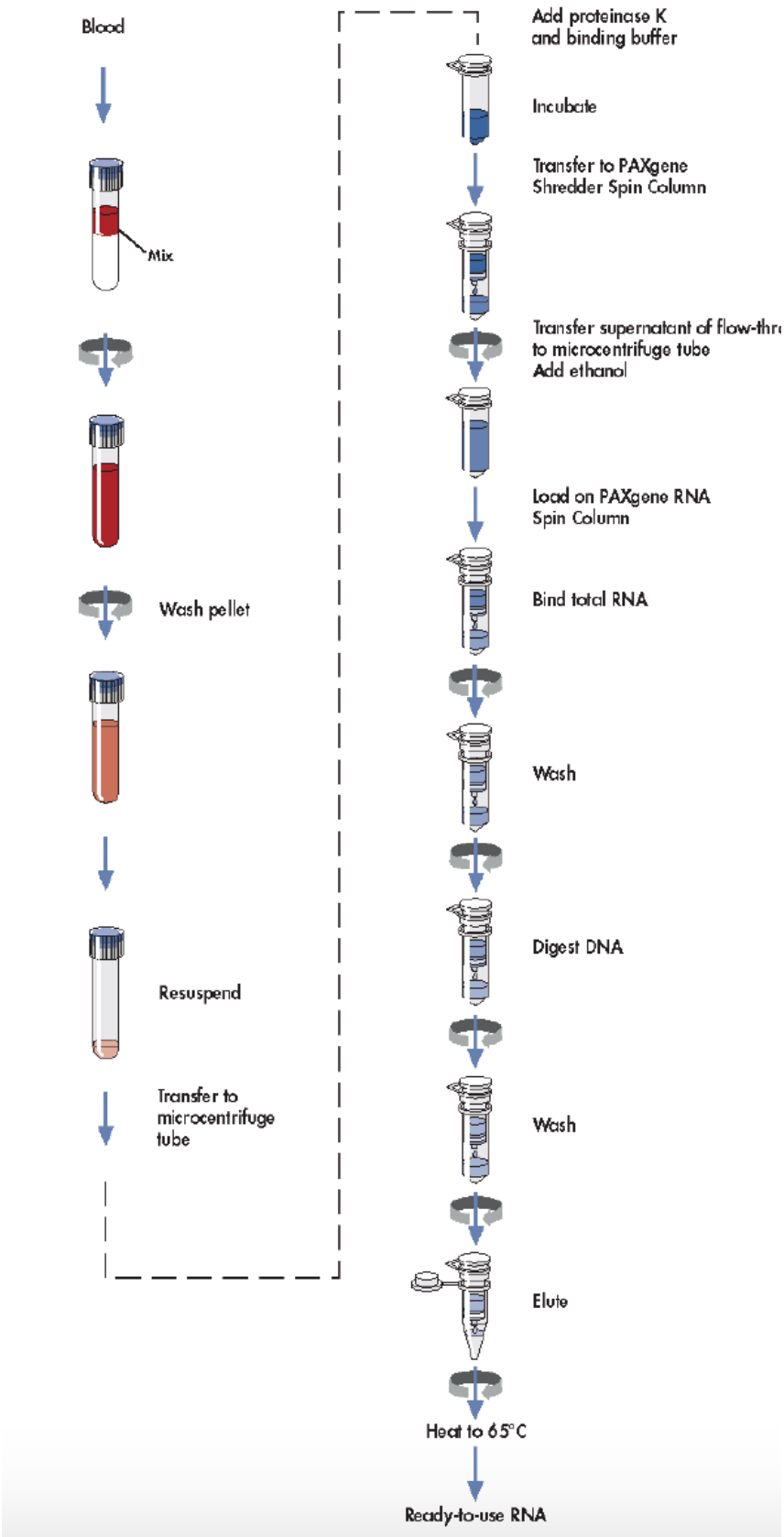


Image from the *PAXGene Blood RNA Kit Handbook Version 2*

Procedure

1. **Centrifuge the PAXgene Blood RNA Tube for 10 minutes at 3000–5000 x *g* using a swing-out rotor.**

Note: Ensure that the blood sample has been incubated in the PAXgene
Blood RNA Tube for a minimum of 2 hours at room temperature (15-25° C), in order to achieve complete lysis of blood cells.

Note: The rotor must contain tube adapters for round-bottom tubes. If othertypes of tube adapter are used, the tubes may break during centrifugation.

2. Remove the supernatant by decanting or pipetting. Add 4 ml RNase-free water to the pellet, and close the tube using a fresh secondary BD Hemogard closure (supplied with the kit).
If the supernatant is decanted, take care not to disturb the pellet, and dry the rim of the tube with a clean paper towel.

3. Vortex until the pellet is visibly dissolved, and centrifuge for 10 minutes at 3000–5000 x *g* using a swing-out rotor. Remove and discard the entire supernatant.
Small debris remaining in the supernatant after vortexing but before
centrifugation will not affect the procedure.

Note: Incomplete removal of the supernatant will inhibit lysis and dilute the lysate, and therefore affect the conditions for binding RNA to the PAXgene membrane.

4. Add 350 µl Buffer BR1, and vortex until the pellet is visibly dissolved.
5. **Pipet the sample into a 1.5 ml microcentrifuge tube. Add 300 µl Buffer
BR2 and 40 µl proteinase K. Mix by vortexing for 5 seconds, and incubate for 10 minutes at 55°C using a shaker–incubator at 400–1400 rpm. After incubation, set the temperature of the shaker–
incubator to 65°C (for step 20).**

Note: Do not mix Buffer BR2 and proteinase K together before adding them to the sample.

6. Pipet the lysate directly into a PAXgene Shredder spin column (lilac)
placed in a 2 ml processing tube, and centrifuge for 3 minutes at maximum speed (but not to exceed 20,000 x *g*).

Note: Carefully pipet the lysate into the spin column and visually check that the lysate is completely transferred to the spin column.
To prevent damage to columns and tubes, do not exceed 20,000 x *g*.
Note: Some samples may flow through the PAXgene Shredder spin column
without centrifugation. This is due to low viscosity of some samples and should not be taken as an indication of product failure.

7. Carefully transfer the entire supernatant of the flow-through fraction
to a fresh 1.5 ml microcentrifuge tube without disturbing the pellet in
the processing tube.

8. Add 350 µl ethanol (96–100%, purity grade p.a.). Mix by vortexing,
and centrifuge briefly (1–2 seconds at 500–1000 x *g*) to remove drops from the inside of the tube lid.

Note: The length of the centrifugation must not exceed 1-2 seconds, as this may result in pelleting of nucleic acids and reduced yields of total RNA.

9. Pipet 700 µl sample into the PAXgene RNA spin column (red) placed
in a 2 ml processing tube, and centrifuge for 1 minutes at 8000–20,000 x *g*. Place the spin column in a new 2 ml processing tube, and discard the old processing tube containing flow-through.

10. Pipet the remaining sample into the PAXgene RNA spin column, and centrifuge for 1 minutes at 8000–20,000 x *g*. Place the spin column in a new 2 ml processing tube, and discard the old processing tube containing flow-through.

Note: Carefully pipet the sample into the spin column and visually check that the sample is completely transferred to the spin column.

11. Pipet 350 µl Buffer BR3 into the PAXgene RNA spin column. Centrifuge for 1 minute at 8000–20,000 x *g*. Place the spin column in a new 2 ml processing tube, and discard the old processing tube containing flow-through.

12. Add 10 µl DNase I stock solution to 70 µl Buffer RDD in a 1.5 ml microcentrifuge tube. Mix by gently flicking the tube, and centrifuge briefly to collect residual liquid from the sides of the tube.

If processing, for example, 10 samples, add 100 µl DNase I stock solution to 700 µl Buffer RDD. Use the 1.5 ml microcentrifuge tubes supplied with the kit.

Note: DNase I is especially sensitive to physical denaturation. Mixing should only be carried out by gently flicking the tube. Do not vortex.

13. Pipet the DNase I incubation mix (80 µl) directly onto the PAXgene RNA spin column membrane, and place on the benchtop (20–30°C) for 15 minutes.

Note: Ensure that the DNase I incubation mix is placed directly onto the membrane. DNase digestion will be incomplete if part of the mix is applied to and remains on the walls or the O-ring of the spin column.

14. Pipet 350 µl Buffer BR3 into the PAXgene RNA spin column, and centrifuge for 1 minute at 8000–20,000 x *g*. Place the spin column in a new 2 ml processing tube, and discard the old processing tube containing flow- through.

15. Pipet 500 µl Buffer BR4 into the PAXgene RNA spin column, and centrifuge for 1 minute at 8000–20,000 x *g*. Place the spin column in a new 2 ml processing tube, and discard the old processing tube containing flow-through.

16. Add another 500 µl Buffer BR4 to the PAXgene RNA spin column. Centrifuge for 3 minutes at 8000–20,000 x *g*.

17. Discard the processing tube containing the flow-through, and place the PAXgene RNA spin column in a new 2 ml processing tube. Centrifuge for 1 minute at 8000–20,000 x *g*.

18. Discard the processing tube containing the flow-through. Place the PAXgene RNA spin column in a 1.5 ml microcentrifuge tube, and pipet 40 µl Buffer BR5 directly onto the PAXgene RNA spin column membrane. Centrifuge for 1 minute at 8000–20,000 x *g* to elute the RNA.

It is important to wet the entire membrane with Buffer BR5 in order to achieve maximum elution efficiency.

19. Repeat the elution step (step 18) as described, using 40 µl Buffer BR5 and the same microcentrifuge tube.

Incubate the eluate for 5 minutes at 65°C in the shaker–incubator (from step 5) without shaking. After incubation, chill immediately on ice.

This incubation at 65° C denatures the RNA for downstream applications.
Do not exceed the incubation time or temperature.

21. If the RNA samples will not be used immediately, store at –20°C or –70°C.

Appendix K: Samples included in transcriptomic analysis with QC results.

RNA quality and quantity extracted per samples with corresponding classification group. Samples used in analysis highlighted in green. No sample highlighted in red.

Sample	Quantity (ng)	260/280	RIN	Sample	Quantity (ng)	260/280	RIN	Phenotype	Include/ Exclude
N/S	N/S	N/S	N/S	P71B	2045	2.11	8.80	Term Rx	EXCLUDE
N/S	N/S	N/S	N/S	P72B	7404	2.16	7.80	Term Rx	EXCLUDE
P73A	6690	2.12	9	N/S	N/S	N/S	N/S	PPROM	EXCLUDE
P74A	5316	2.1	8.70	P74B	7939	2.15	8.90	TERM CONTROL	INCLUDE
P75A	8909	2.12	8.40	P75B	7215	2.09	8.90	sPTB	INCLUDE
P76A	10927	2.14	8.20	P76B	8681	2.12	8.70	TERM CONTROL	INCLUDE
P77A	6194	2.17	8.60	P77B	2785	2.18	8.70	TERM CONTROL	INCLUDE
P78A	4803	2.15	8.70	P78B	1771	2.10	8	PPROM	INCLUDE
P79A	5074	2.09	8.40	P79B	735	1.91	8.50	TERM CONTROL	INCLUDE
P80A	4073	2.18	7.90	P80B	3532	2.15	8.50	Iatrogenic PTB	EXCLUDE
P81A	2272	2.17	8.10	N/S	N/S	N/S	N/S	PPROM + chorio	EXCLUDE
P82A	2408	2.16	8.30	P82B	2145	2.16	8.70	PPROM	INCLUDE
P83A	3382	2.24	8.20	P83B	4095	2.07	8.10	sPTB	INCLUDE
P84A	825	2.14	8.40	P84B	7629	2.13	8.20	LATE sPTB	INCLUDE
P85A	6527	2.16	8.50	P85B	5700	2.11	8.70	sPTB	INCLUDE
N/S	N/S	N/S	N/S	P86B	3495	2.07	8.30	TERM CONTROL	EXCLUDE
P87A	6543	2.12	8.50	P87B	4990	2.11	8.30	TERM CONTROL	INCLUDE
P88A	6939	2.13	7.90	P88B	3101	2.09	8.10	sPTB	INCLUDE
P89A	7905	2.12	8.10	N/S	N/S	N/S	N/S	TERM CONTROL	EXCLUDE
P90A	5130	2.15	8.50	P90B	6128	2.1	9.10	LATE sPTB	INCLUDE
N/S	N/S	N/S	N/S	P91B	2708	2.11	9.20	TERM CONTROL	EXCLUDE
P92A	7125	2.16	8.70	P92B	3068	2.07	8.10	LATE sPTB	INCLUDE
P93A	600	2.34	9	N/S	N/S	N/S	N/S	Term Rx	EXCLUDE
P94A	2625	2.08	8.50	P94B	0.0	0.87	N/A	Term Rx	EXCLUDE
P95A	5175	2.14	9.10	P95B	4628	2.11	8.40	TERM CONTROL	INCLUDE
N/S	N/S	N/S	N/S	P96B	4005	2.14	8.80	TERM CONTROL	EXCLUDE
N/S	N/S	N/S	N/S	P97B	2310	2.17	8.70	LATE sPTB	EXCLUDE
P98A	2572	2.17	8.10	P98B	2183	2.19	8.10	TERM CONTROL	INCLUDE
P99A	4278	2.18	8	P99B	0.0	N/A	N/A	TERM CONTROL	EXCLUDE
P100A	2198	2.23	8	P100B	0.0	-0.32	N/A	TERM CONTROL	EXCLUDE
N/S	N/S	N/S	N/S	P101B	0.0	-1.67	N/A	TERM CONTROL	EXCLUDE
P102A	0.0	-3.64	N/A	P102B	0.1	0.68	N/A	TERM CONTROL	EXCLUDE
P103A	0.3	2.32	N/A	P103B	0.1	0.77	N/A	TERM CONTROL	EXCLUDE

P104A	0.0	-0.12	N/A	P104B	0.1	0.75	N/A	TERM CONTROL	EXCLUDE
P105A	0.0	-2.86	N/A	P105B	0.2	1.68	N/A	TERM CONTROL	EXCLUDE
P106A	0.0	0.53	N/A	P106B	38	2.55	8.90	Term Rx	EXCLUDE
N/S	N/S	NS	N/S	P107B	0.2	1.36	N/A	Term Rx	EXCLUDE
P108A	0.0	N/A	N/A	P108B	0.1	2.37	N/A	LATE sPTB	EXCLUDE
P109A	0.1	1.49	N/A	P109B	0.1	1.09	N/A	TERM CONTROL	EXCLUDE
P110A	0.1	1.32	N/A	P110B	0.4	1.32	N/A	LATE sPTB	EXCLUDE
P111A	1095	2.23	8.10	P111B	582	2.23	8.10	TERM CONTROL	INCLUDE
P112A	0.2	1.35	N/A	P112B	1268.5	2.20	8.20	TERM CONTROL	EXCLUDE
P113A	2865	2.11	8.70	P113B	794	2.59	8.70	TERM CONTROL	INCLUDE
P114A	0.2	1.61	N/A	P114B	1175	2.33	8.40	LATE sPTB	EXCLUDE
P115A	1485	2.36	8.10	P115B	1582	2.14	8.00	TERM CONTROL	INCLUDE
P116A	1226	1.96	8.30	P116B	4466	2.07	8.50	TERM CONTROL	INCLUDE
P117A	715.5	2.56	2.40	P117B	1648	1.85	7.90	TERM CONTROL	EXCLUDE
P118A	2204	1.9	8.30	P118B	2485	1.88	7.90	TERM CONTROL	INCLUDE
P119A	1410	1.81	8	P119B	2325	1.96	7.70	TERM CONTROL	INCLUDE
P120A	1050	1.86	8	P120B	3278	1.99	8.50	LATE sPTB	INCLUDE
P121A	1769	1.93	8.20	P121B	2953	2	8.60	LATE sPTB	INCLUDE
P122A	3528	2.11	8	P122B	2484	2.02	8	TERM CONTROL	INCLUDE
P123A	389	1.81	7.50	N/S	N/S	N/S	N/S	sPTB	EXCLUDE
P124A	1138	2.17	7.70	P124B	2743	2.03	7.90	TERM CONTROL	INCLUDE
P125A	460	1.74	7.90	P125B	2355	2.01	8	Term Rx	EXCLUDE
P126A	1548	2.04	8	P126B	1723	1.99	8.20	PPROM - poly	EXCLUDE
P127A	1552	2.08	7.70	P127B	5169	2.1	8.10	Term Rx	EXCLUDE
P128A	5092	2.1	8.70	N/S	N/S	N/S	N/S	PPROM – genetic abn	EXCLUDE
P129A	4185	2.09	7.80	P129B	1508	2.02	7.90	TERM CONTROL	INCLUDE

Appendix L: R script used for random forest analysis

```
ranger(dependent.variable.name = "phenotype", data = trans_merged[,-
1], importance = "permutation",num.trees=10000) -> ptb.rf
```

Appendix M: List of metabolites detected by NMR

Phosophocholine
2-hydroxybutyrate
Arginine
2-hydroxyvalerate
3-hydroxybutyrate
acetate
acetoacetate
alanine
choline
citrate
creatine
creatinine
desaminotyrosine
Gln
glucarate
myoinositol
glucose
glutamate
glycylproline
histidine
histidine
Isoleucine
isopropanol
Lactate
Leucine
lysine
mannose
mobile-lipids
n-dimethylamine
phenylalanine
proline
propylene-glycol
threonine
tyrosine
Valine
.

Appendix N: Variable Importance by Test

Random Forest:

TPM1_week16	100.000
CDH1_week16	71.121
unknown_5_16weeks	64.849
BAG1_week16	61.099
SPX_week16	60.438
BCL2L1_week16	59.805
YBX3_week16 5	2.718
unknown_37_16weeks	51.713
TSTA3_week16	45.386
glucose_44_16weeks	43.232
unknown_52_16weeks	42.900
unknown_54_16weeks	42.900
ST13_week16	41.704
threonineunknown_55_16weeks	39.075
unknown_53_16weeks	38.521
GABARAPL2_week16	36.739
ANLN_week16	36.619
GYPA_week16	34.987
glucose_36_16weeks	34.483
SIAH2_week16	33.238
desaminotyrosine_12_16weeks	32.800

phenylalanine_6_16weeks	32.423
SLC38A5_week16	32.133
unknown_98_16weeks	31.926
PLVAP_week16	29.616
glucose_39_16weeks	29.186
CTDSPL_week16	28.575
glucose_40_16weeks	28.308
proline_108_16weeks	27.702
XK_week16	27.088
unknown_125_16weeks	26.771
glucose_45_16weeks	26.639
CA1_week16	24.966
glucoseunknown_50_16weeks	24.821
GSPT1_week16	24.095
glucose_35_16weeks	23.725
X2_hydroxyvalerate_28_16weeks	23.557
BCL2L13_week16	23.475
FAM46C_week16	22.879
SGIP1_week16	22.557
EPB42_week16	20.791
GYPB_week16	20.571
glucose_60_16weeks	20.349
MOXD2P_diff	20.208
creatinine_29_20weeks	20.120

glucose_43_16weeks	20.056
glucose_59_16weeks	19.378
USP17L12_week16	18.304
unknown_129_20weeks	18.291
glucose_38_16weeks	18.288
HIST1H3H_week16	17.921
unknown_41_16weeks	17.736
MIR4255_diff	17.124
glucose_49_16weeks	16.854
CD79A_week16	16.824
ProGlu_115_20weeks	16.794
RSU1_week16	16.707
TNFRSF17_week16	16.701
myoinositol_56_16weeks	15.610
APEX1_week16	15.244
phenylalanine_8_16weeks	15.098
glycylproline_114_20weeks	14.965
glucose_58_16weeks	14.524
unknown_51_16weeks	13.995
Isoleucine_138_20weeks	13.967
glucose_47_16weeks	13.873
mobile_lipids_18_16weeks	12.600
glucose_19_16weeks	12.482
unknown_9_16weeks	12.371

glucose_66_16weeks	12.149
TUBB1_week16	12.135
lipidunknown_32_16weeks	12.079
glucaratemyoinositol_30_16weeks	11.987
glucose_64_16weeks	11.915
CHRNE_week16	11.626
glucose_61_16weeks	11.499
ABALON_week16	11.462
unknown_10_16weeks	11.352
CDH1_diff	11.204
KEL_diff	11.040
man_20_20weeks	10.902
DYRK3_week16	10.792
SNORD41_week16	10.720
DDB1_week16	10.716
OR2W3_week16	10.632
SPTB_diff	10.614
acetoacetate_111_16weeks	10.507
citrate_103_16weeks	10.473
SELENBP1_week16	10.421
ANXA2_week16	10.345
glucaratemyoinositol_30_20weeks	10.332
man_20_16weeks	10.085
unknown_129_16weeks	10.075

SVM linear:

all variables have equal importance

SVM Gaussian kernel:

all variables have equal importance

Kmeans

mobile_lipids_131_20weeks	100.000
mobile_lipids_131_16weeks	78.502
Lactate_130_20weeks	39.099
Lactate_130_16weeks	37.416
mobile_lipids_132_20weeks	23.241
mobile_lipids2_hydroxyisovalerate_145_20weeks	19.616
argininePhosophocholine_82_20weeks	18.935
mobile_lipids_132_16weeks	18.414
mobile_lipids2_hydroxyisovalerate_145_16weeks	16.054
argininePhosophocholine_82_16weeks	13.234

LDA

All variables have equal importance

Genetic Expression Programming:

SPTB_diff	100.000
SNORA11_diff	50.000

PNN:

desaminotyrosine_12_16weeks	100.000