Major sources of variation in placental DNA methylation

by

Victor Yuan

B.Sc., Concordia University, 2016

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

March 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Major sources of variation in placental DNA methylation

submitted by Victor Yuan in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Genome Science and Technology

Examining Committee:

Dr. Wendy P. Robinson, Professor, Medical Genetics, UBC

Supervisor

Dr. Alex G. Beristain, Associate Professor, Obstetrics and Gynaecology, UBC

Supervisory Committee Member

Dr. Michael Kobor, Professor, Medical Genetics, UBC

Supervisory Committee Member

Dr Paul Pavlidis, Professor, Michael Smith Laboratories, UBC

University Examiner

Dr. Cathie Garnis, Associate Professor, Surgery, UBC

University Examiner

Dr. Jeffrey Craig, Associate Professor, Paediatrics, University of Melbourne

External Examiner

Additional Supervisory Committee Members:

Dr. Sara Mostafavi, Associate Professor, Paul G. Allen School of Computer Science and Engineering, University of Washington

Supervisory Committee Member

Abstract

DNA methylation (DNAm) is an epigenetic mark that can control or reflect gene expression in a highly cell-specific manner. Placental DNAm has been studied in various contexts, however, several sources of variation remain uncharacterized. Before we are able to understand how placental DNAm contributes to important health-relevant contexts, we must develop an understanding of the underlying normal variation that occurs in the placental methylome. For example, a high proportion of variation in DNAm can vary by ethnicity and genotype. DNAm is also highly cell-specific, and the placenta is a heterogeneous tissue comprised of several distinct cell populations. However, due to difficulty of isolating cell populations, most placental DNAm research is conducted on whole placental chorionic villi tissue. Isolating placental tissue without contamination from maternal tissue, such as decidua and blood, can be challenging, especially in earlier gestation samples where sampling enough tissue is difficult.

In this thesis, I hypothesize that a large proportion of the variation in placental DNAm can attributed to ethnicity, genetic ancestry, cell composition, cell-specific effects, and the presence of maternal cells. Using high-density DNAm microarray profiling, and open access genomic data repositories, I assessed these factors in placental samples with various phenotypes. I found that ethnicity and genetic ancestry are associated with placental DNAm variation in samples containing self-reports of White/Caucasian, East Asian/Asian, and Black/African American ethnicity. Further, I found that it is possible to predict ethnicity and genetic ancestry with high accuracy and reliability from placental DNAm. Another major source of variation is cell-specific DNAm, which I characterized from placental samples of first trimester and term pregnancies. I found that trophoblast and Hofbauer cells are highly epigenetically distinct, and many placental epigenetic features are conserved in trophoblasts but not always in other placental cells. I developed a reference to estimate cell composition from placental chorionic villi DNAm. Lastly, I developed an approach to estimate maternal cells present in chorionic villi samples, using DNAm, and found several previously published placental DNAm studies to contain maternally-contaminated samples. Overall, I contributed to our understanding of placental DNAm, and have provided bioinformatic tools for future placental DNAm research.

Lay Summary

The placenta is an essential organ for a healthy pregnancy. DNA methylation (DNAm) is a molecular mark that helps the placenta control what genes are turned on and off. DNAm is an important component of placental biology, and when pregnancy complications occur, corresponding changes in DNAm can occur. However, our understanding of what controls placental DNAm itself is rudimentary. We need to understand what influences placental DNAm under normal, healthy conditions before we can understand it under other contexts like in disease. My thesis aims at understanding some of the major factors that influence placental DNAm. Throughout my research, I have also developed several analysis tools to help future placental DNAm research. Ultimately, my research has contributed to our understanding of the molecular characteristics of the placenta, which in turn helps us understand the role of placental biology during pregnancy.

Preface

Parts of this dissertation have been published and include work performed by collaborators.

Chapter 1

Figure 1 "Structure of the placenta and chorionic villi" is from Del Gobbo 2021, which is an image that was adapted from Del Gobbo 2019. This figure has been reprinted with permission from the authors and copyright holder, Springer-Verlag GmbH Germany, part of Springer Nature. Del Gobbo GF created the original and modified image.

Del Gobbo GF (2021) Genomic profiling in the placenta: toward a greater understanding of genetic variation contributing to placental insufficiency and fetal growth restriction. University of British Columbia

Del Gobbo GF, Konwar C, Robinson WP (2019) The significance of the placental genome and methylome in fetal and maternal health. Hum Genet. https://doi.org/10.1007/s00439-019-02058-w

Parts of Chapter 1 subsection "DNA methylation" has been submitted to be published as a book chapter by V Yuan and WP Robinson (2022). I wrote all parts (general features of DNA methylation, DNA methylation in development) that were included in this dissertation. All authors provided critical revisions of the final manuscript.

Parts of Chapter 1 subsection "Epigenetics and placenta cell types" have been published as a book chapter in *Human Reproductive and Prenatal Genetics*.

Robinson WP, Peñaherrera MS, Konwar C, Yuan V, Wilson SL (2019) Chapter 13 - Epigenetic Modifications in the Human Placenta. In: Leung PCK, Qiao J (eds) Human Reproductive and Prenatal Genetics. Academic Press, pp 293–311

Copyright © 2019 Elsevier Inc.I wrote the section on titled "The epigenetic profiles of placental cell types." WP Robinson wrote the Overview and introduction sections. MS Peñaherrera wrote the "Monoallelic gene inactivation" section. SL Wilson and C Konwar wrote the Placental Pathology and "Clinical applications" section. All authors reviewed and edited the final manuscript.

A version of Chapter 1 subsection "Placental DNA methylation variation across human populations" has been published in *BMC Epigenetics & Chromatin*.

Yuan V, Price EM, Del Gobbo G, Mostafavi S, Cox B, Binder AM, Michels KB, Marsit C, Robinson WP (2019) Accurate ethnicity prediction from placental DNA methylation data. Epigenetics & Chromatin 12:51. https://doi.org/10.1186/s13072-019-0296-3

 V Yuan, EM Price, G Del Gobbo, S Mostafavi, and WP Robinson contributed to the design of the study. V Yuan, EM Price, and G Del Gobbo performed all data analysis. EM Price, G Del Gobbo, S Mostafavi, Cox B, Binder AM, Michels KB, Marsit C, Robinson WP helped generate and contribute the data. EMP and WR conceived of the study. All authors approved and contributed to the writing of the final manuscript. All authors read and approved the final manuscript.

A version of Chapter 1 subsection "Epigenetics and placenta cell types" has been published in *BMC Genomics*.

Yuan V, Hui D, Yin Y, Peñaherrera MS, Beristain AG, Robinson WP (2021) Cell-specific characterization of the placental methylome. BMC Genomics 22:6. https://doi.org/10.1186/s12864-020-07186-6

 WP Robinson, AG Beristain, and MS Peñaherrera contributed to project design. V Yuan performed analysis of data, and wrote manuscript. D Hui standardized and performed tissue and cell collection supported by AGB, MSP and VY. D Hui, MS Peñaherrera, V Yuan, ran methylation arrays. Y Yin helped V Yuan developed interactive data exploration app. All authors read, edited, and approved the final manuscript.

Chapter 2

A version of chapter 2 has be published in BMC Epigenetics & Chromatin.

Yuan V, Price EM, Del Gobbo G, Mostafavi S, Cox B, Binder AM, Michels KB, Marsit C, Robinson WP (2019) Accurate ethnicity prediction from placental DNA methylation data. Epigenetics & Chromatin 12:51. https://doi.org/10.1186/s13072-019-0296-3

 V Yuan, EM Price, G Del Gobbo, S Mostafavi, and WP Robinson contributed to the design of the study. V Yuan performed all data analysis. EM Price, G Del Gobbo, S Mostafavi, Cox B, Binder AM, Michels KB, Marsit C, Robinson WP helped generate and contribute the data. EMP and WR conceived of the study. All authors approved and contributed to the writing of the final manuscript. All authors read and approved the final manuscript. Data, text, figures and tables published in this article are contained in Chapter 2.

Chapter 3

A version of chapter 3 has be published in BMC Genomics.

Yuan V, Hui D, Yin Y, Peñaherrera MS, Beristain AG, Robinson WP (2021) Cell-specific characterization of the placental methylome. BMC Genomics 22:6. https://doi.org/10.1186/s12864-020-07186-6

 WP Robinson, AG Beristain, and MS Peñaherrera contributed to project design. V Yuan performed analysis of data, and wrote manuscript. D Hui standardized and performed tissue and cell collection supported by AGB, MSP and VY. D Hui, MS Peñaherrera, V Yuan, ran methylation arrays. Y Yin helped V Yuan developed interactive data exploration app. All authors read, edited, and approved the final manuscript. The data, text, figures, and tables published in this article are included in Chapter 3.

Chapter 4

V Yuan and WP Robinson contributed equally to project design. V Yuan obtained the data, performed all analyses, and wrote the text. V Yuan and WP Robinson read, edited, and approved the final text.

The work presented herein was approved by the University of British Columbia and the Children's and Women's Hospital Research Ethics board in Vancouver, BC, Canada; certificates H04–70488, H16–02280, H13–00640).

Table of Contents

List of Tables

List of Figures

Glossary

450k	Infinium HumanMethylation450 Beadchip
850k/EPIC	Infinium HumanMethylationEPIC Beadchip
AFR	African
AIMs	Ancestry Informative Markers
ASI	Asian
BMIQ	Beta-mixture interquantile normalization
CAU	Caucasian
CBP	CREB-binding protein
CBS	CIBERSORT
CP	Constrained projection
CpG	Cytosine-guanine dinucleotide
CTB	Cytotrophoblast
CV	Chorionic villi
DMC	Differentially methylated cytosine
DMR	Differentially methylated region
DNAm	DNA methylation
DNMT	DNA methyltransferase
EA	Epigenetic age
EC	Endothelial cells
EGA	Epigenetic gestational age
EGF	Epidermal growth factor
EOPE	Early-onset preeclampsia
EPIC	Infinium HumanMethylation EPIC Beadchip
eQTL	Expression quantitative trait loci
eSTB	Enzymatically-separated syncytiotrophoblast
EVT	Extra-villous trophoblast
EWAS	Epigenome-wide association study
FACS	Fluorescence-activated cell-sorting
FDR	False-discovery rate
GA	Gestational age

GLMNET	Generalized linear model with elastic net regularization
H3K27me(2/3)	Di- or tri- methylation of lysine 27 on histone H3
H3K4me3	Trimethylation of lysine 4 on histone H3
H3K9me2	Dimethylation of lysine 9 on histone H3
HAT	Histone acetyltransferases
HB	Hofbauer cells
HDAC	Histone deacetylase complexes
HIF	Hypoxia-inducible Factor
ICM	Inner cell mass
ICR	Imprinting control region
IUGR	Intrauterine growth restriction
KNN	K-nearest neighbours
LINE-1	Long interspersed nuclear element-1
LODOCV	Leave-one-dataset-out cross validation
LOPE	Late-onset preeclampsia
MAE	Mean absolute error
MHC	Major histocompatibility complex
mQTL	Methylation quantative trait loci
noob	Normal exponential out-of-band normalization
nRBC	Nucleated red blood cells
NSC	Nearest shrunken centroids
PCA	Principal component analysis
PE	Preeclampsia
PlaNET	Placental DNAme Elastic Net Ethnicity Tool
PMD	Partially methylated domain
PTB	Preterm birth
RMSE	Root mean squared error
RPC	Robust partial correlations
SC	Stromal cells
scRNAseq	Single-cell RNA sequencing
scWGBS	Single-cell whole genome bisulfite sequencing

SNP	Single nucleotide polymorphism
STB	Syncytiotrophoblast
SVM	Support vector machines
TB	Trophoblast
TE	Trophectoderm
VEGF	Vascular endothelial growth factor

Acknowledgements

I acknowledge that this dissertation was mainly written on the traditional, ancestral, and unceded territory of the Coast Salish People, which include the Musqueam, Squamish, and Tsleil-Waututh nations.

Many thanks to my supervisor, Wendy Robinson, for her unwavering guidance and commitment, and support for my research. Thank you for also enriching my graduate experience with many opportunities, and for also nurturing a environment that I felt comfortable exploring and expressing my own ideas and interests. Thank you for your kindness; you have been a great role model.

I am very grateful to my commitee members - Alex Beristain especially for all the practical help, but also for being a wonderful human being. I have very much enjoyed our meetings and discussions, and appreciate always being available when I needed it. Thank you Sara Mostafavi and Michael Kobor for all the insightful guidance on my research and also my graduate and post-grad experience. I want to also especially thank Vincenzo Coia, for being an amazing mentor, and more importantly a big inspiration for me not only professionally but also personally too. Thank you everyone for supporting me.

Before starting graduate school, it is not uncommon to hear how difficult and lonely it can be. I have always been independent and felt like this would not be a problem for me. But I found that my graduate experience was not a lonely one at all. Thank you Robinson lab members for making my graduate experience enjoyable. Thank you Maria, for supporting my research, and for pushing me to always take a second look at my data. Thank you Magda, my first mentor, for teaching and inspiring me about placental genomics research. Thank you Sam, Chaini, and Giulia for taking care of me, while I stumbled through the first years of graduate life. Thank you Amy and Iciar, for always making time for my questions and discussions, and for also being great friends. Thank you to my mentees, students that I have had the privelege to mentor and work alongside. Thank you Desmond for helping me with my research, and for being a great friend too.

Thank you BCCHR for creating an amazing research environment, and for prioritizing the experience of trainees, through resources, workshops, research days, and other events. Thank you UBC and CIHR for funding and professional and academic support.

Lastly, thank you Yvonne for supporting me when I most needed it. Thank you Sandy for being the best dog, I wish you were there for my graduation. Thank you my parents, for of course supporting my education, and inspiring me to choose UBC and graduate life.

They say PhD is a lonely pursuit, but I disagree. None of this would have been possible without the help of others.

Dedication

To my parents

1 Introduction

1.1 Dissertation context and overview

The placenta is a key component to the healthy development of the fetus during pregnancy. At the cellular level, highly specialized placental cells regulate gene expression programs in part through epigenetics. DNA methylation (DNAm), an epigenetic modification that involves the chemical addition of a methyl group to cytosines, plays an integral role in regulating gene expression at the cellular level. Most cytosine DNAm occurs at cytosine-guanine sites (CpGs), although non-CpG methylation (i.e. at cytosines followed by thymine, adenine, or another cytosine) can play important roles in development and disease (1).

Recent developments in the past two decades have allowed the characterization of DNAm at an unprecedented level of resolution in the human genome. Specifically, advances in microarray technology combined with bisulfite conversion have allowed the measurement of DNAm on the order of hundreds of thousands of CpGs per sample. Large human population studies of DNAm using microarrays has taken off in popularity, which has in turn spurred significant developments in bioinformatics, allowing researchers to investigate and understand phenotype-relevant signals in this information-rich genomic data.

Understanding the variation in placental DNAm is important to our understanding of the molecular processes that contribute placental biology, which in turn will lead to a better understanding of the role of the placenta in pregnancy. Although there is great interest in studying DNAm, there are significant technical and biological challenges that can lead to spurious results. Understanding the contributors to DNAm variability and how to account for them in our analyses will be a key step in moving DNAm research forward. In this dissertation, I characterized some of the major contributors to variotion in placental DNAm, and have developed bioinformatic methods to predict and account for these factors in future placental DNAm research. These factors that influence placental DNAm, that I have studied in this dissertation, include ethnicity and genotype, cell-specificity and cell composition dynamics across gestation, and maternal cell contamination.

1.2 Placental function

The placenta is an essential organ that develops from the embryo to nurture its development throughout pregnancy (Figure 1.1). Sometimes described as a "vascular organ," the placenta is responsible for the establishment and regulation of maternal-fetal blood flow. Maternal blood is entered from the uterine spiral arteries into the intervillous space of the placenta. Chorionic villi, the main structural unit of the placenta, are tree-like structures populating the intervillous space that connect to the umbilical cord. Deoxygenated and nutrient-depleted fetal blood travels through arteries into the chorionic villi, and oxygenated blood returns

through veins to the fetus (2). Maternal-fetal blood exchange is essential for the proper growth of the fetus, as imbalances, such as placental insufficiency, a condition where the placenta does not deliver sufficient nutrients and oxygen to the developing fetus, can lead to complications such as fetal growth restriction, preeclampsia, and pre-term birth (3). The placenta also influences the maternal environment (cardiovascular, immune, and metabolic systems) through hormonal processes to support the proper nutrient and oxygen supply to the growing fetus (4). Placental hormones such as those from the proclactin growth hormone family and steroids, are especially important in promoting maternal physiological adaptations to increase the bioavailability of nutrients to the fetus (5). Fetomaternal tolerance is also an important pregnancy process that is in part regulated by the direct interactions between the placenta and maternal cells in the decidua. Trophoblasts, specificaly extravillous populations that travel into the decidua, express nonclassical major histocamptibility complex (MHC) molecules that modulates various populations of uterine immune cells to promote tolerogenicity of the placenta/fetus, in addition to promoting placental growth and invasion (6,7). Lastly, the placenta acts as a biophysical barrier that protects the fetus against potential pathogenic infections (8,9).

1.3 DNA methylation in the human genome

Despite all cells of the human body being nearly genetically identical, human cell types must coordinate a highly specific gene expression program to fulfill each of their own unique cellular niches. To achieve cell-specific gene expression programs, epigenetic processes function as a layer of molecular information that influences DNA interactions to regulate gene expression. Types of epigenetic marks include DNAm, histones and their post-translational modifications, non-histone chromatin proteins, higher order chromatin and chromosome organization, and non-coding RNAs. DNAm is the most widely studied epigenetic mark, owing to its relative stability and availability of cost-efficient technologies.

DNAm at gene promoters is highly anti-correlated with gene expression. CpG islands, which are genomic regions (~1000bp long) with high CpG density often found near gene promoters, are often devoid of DNAm (10). Methylation of DNA at CpG islands results in the silencing of gene expression, for example through the recruitment of repressive methyl-binding proteins, or by impairing transcription factor binding (11). One of the main roles of DNAm is to also silence transposable and viral elements, which make up Approximately 45% of the mammalian genome (12). DNAm also serves important roles in regulating genomic imprinting (13), which is where a gene's expression occurs from one allele (monoallelic) in a parent-of-origin -specific manner, and in X-inactivation, which is the silencing of one X chromosomes in females (14).

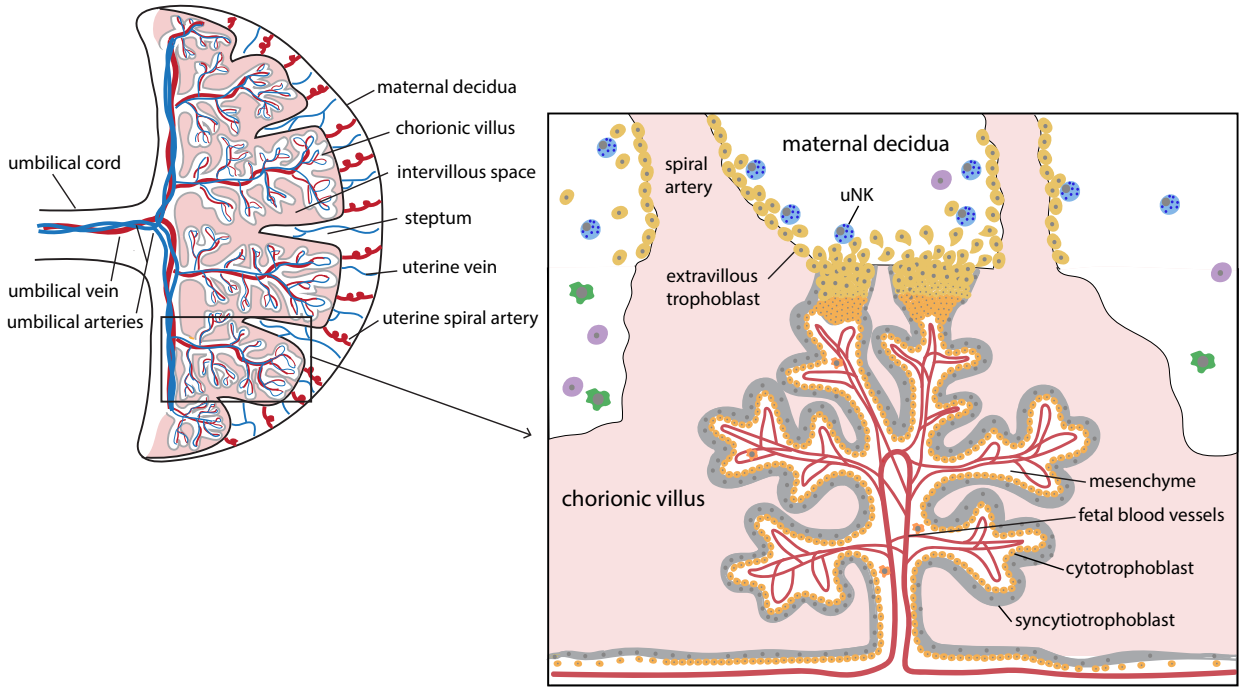

Figure 1.1: Schematic representation of the organization of the placenta (left), which is embedded in the maternal uterus. It is organized into branched chorionic villi that contain the fetal vasculature that will connect to the fetus via the umbilical vein and arteries. The villi are bathed by maternal blood contained in the intervillous space, delivered by the uterine spiral arteries. Representation of a chorionic villus (right), composed of the outer syncytiotrophoblast and inner cytotrophoblast cell layers, inner mesenchyme, and fetal blood vessels. The villus is anchored to the maternal decidua, and extravillous trophoblast cells migrate and invade the maternal decidua, where they interact with resident maternal immune cells such as the uterine natural killer cells (uNK), and remodel the uterine spiral arteries. From Del Gobbo 2021 (15), University of British Columbia, Vancouver, Canada. Original image adapted from Del Gobbo 2019 (16). Copyright 2019 by Springer-Verlag GmbH Germany, part of Springer Nature. Reprinted with permission.

Developmentally, DNAm is highly dynamic and serves as an important mark that is reprogrammed for the generation and establishment of new genomes (Figure 1.2). DNAm is first erased during the maturation of primordial germ cells, owing to actions of TET enzymes that convert 5' methyl-cytosine (5'mc) to 5' hydroxymethyl-cytosine (5'hmc), and also to passive dilution through successive replicative events coupled with the absence of normal DNAm maintenance activity (17,18). During gametogenesis, maturation of sperm and egg cells, DNAm and other epigenetic marks are re-established in a highly cell-specific manner. But, after fertilization, the asymmetric epigenetic marks acquired in the sperm and egg genomes must become equalized, and thus go through another round of reprogramming (19). With the exception of imprinted regions, for which their allele-specificity is protected from this erasure, DNAm is erased in the paternal genome through primarily active processes governed by the TET family of enzymes. This is in contrast to the maternal genome, which undergoes DNAm erasure through a primarily replicative-dependent passive dilution mechanism (20). Lastly, at the blastocyst stage, DNAm is at it's lowest point, and new epigenetic

marks can begin establishing in a highly cell-specific manner, with many cell types and tissues not reaching their final differentiated profiles until later in life.

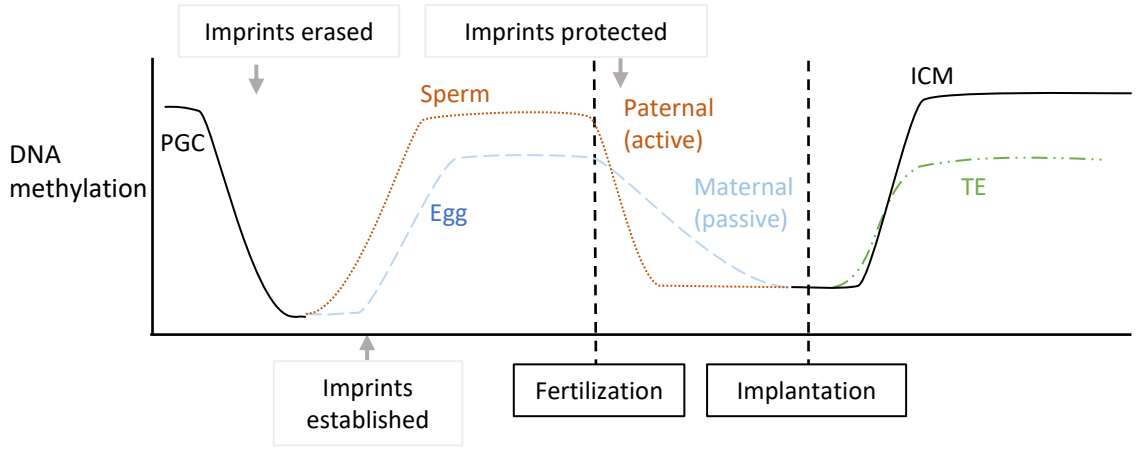

Figure 1.2: DNA methylation (DNAm) reprogramming during development. The first major genome-wide DNAm erasure occurs in primordial germ cells (PGCs), and includes erasure of prior gametic imprints. DNAm is reestablished, but to a greater extent in sperm as compared to oocytes (egg). After fertilization, the asymmetric paternal and maternal epigenomes begin reprogramming. The paternal genome undergoes active demethylation, whereas the maternal genome undergoes demethylation more gradually through passive replication-dependent mechanisms. Gamete-specific differentially methylated regions (i.e. genomic imprints) and most repetitive sequences are protected during post-fertilization epigenetic reprogramming. At the blastocyst stage, there is the first lineage-specification event of inner cell mass (ICM) and trophectoderm (TE). By the blastocyst stage, DNAm is at its lowest point. After implantation, DNAm is established in both ICM and TE, but in TE remains lower, a difference that is retained through development.

1.3.1 Measuring DNA methylation with Illumina microarrays

Although DNAm can be measured at various resolutions and technologies, ranging from targeted gene assays to whole genome bisulfite sequencing, this dissertation exclusively relies on the use of two generations of Illumina DNAm microarrays. The Infinium HumanMethylation450 and Infinium MethylationEPIC BeadChip Microarrays, commonly referred to as the 450k and 850k arrays, measure DNAm at ~485,000 and ~850,000 CpGs across the genome. Because of their reasonable coverage over element-rich genomic regions (e.g. genes, enhancers), affordability, and reliability, these microarrays are popular for population-based studies investigating disease and enivornmental effects on DNAm. Moreover, about 90% of the 450k is also covered by the newer 850k, which means that a significant amount of data is comparable from older studies (21). Particularly in at a time when data reproducibility is a high priority, the ability to compare new data and findings to past research is a highly beneficial characteristic of the Illumina DNAm microarrays.

The microarray relies on DNA hybridization technology, where DNAm at each CpG is measured by its own 50 base pair probe that is complementary such that the 3' end of the probe ends with the CpG site. To

distinguish methylated CpGs from unmethylated ones, bisulfite conversion is used to convert unmethylated Cs into uracils (U) through chemical examination, and then to thymine (T) through polymerase chain reaction (PCR). Then, bisulfite-converted DNA is hybridized in a specific manner depending on two probe types that have different underlying mechanisms to separate methylated and unmethylated -associated signal. Type I probes consists of two probe designs that are attached to two individual beads: unmethylated and methylated. Probe hybridization for type I probes depends on the methylation-dependent sequence complementarity that results from bisulfite conversion (bisulfite converted "T" vs methyl-protected "C"). Type II probes, in contrast, have only one type of bead, where either unmethylated and methylated DNA hybridizes to. Methylated and unmethylated CpGs are distinguished instead at the single base extension step, which is when one fluorescently-labelled base is added to the 3' end of the probe; either a "G" for methylated sequences (complement to C), or "A" for unmethylated sequences (complement to T). Different fluorescent dyes (red and green) therefore distinguishes methylated and unmethylated DNA only for type II probes, whereas for Type I probes the methylated and unmethylated intensities is distinguished by whether specific beads have any (red or green) fluorescense signal. The Type I probes were the primary probe type of earlier generation Illumina DNAm arrays, and perform better in CpG dense regions, which were a higher focus in earlier DNAm studies. Type II probes, however, only require 1 probe design, which means more probes can be fit per array, which was factor that enabled increased coverage for the recent generations of DNAm microarrays (450k and 850k).

A challenge introduced by having two probe types is that technical differences exist in the resulting DNAm measurements, which must be accounted for when comparing CpGs measured by different probe types (21). This is accounted for using statistical normalization techniques that address the difference in DNAm distributions between the different probe types (22,23). Additional types of technical variation is accounted for by these DNAm normalization techniques, such as the differences in methylated and unmethylated channels, green and red fluorescence, and in background signal intensities. A small proportion of the 450k and 850k probes have issues with performance. One set of probes that are commonly referred to as "cross-hybridizing" have a non-trivial amount of sequence homology to multiple parts of the genome, resulting in spurious signal. Another set of probes have single nucleotide polymorphisms (SNPs) within the probe body itself, therefore affecting the strength of hybridization directly. Efforts have been made to identify cross-hybridizing and probes with SNPs in the probe body for removal from downstream analyses .

1.4 Placental DNA methylation variation across human populations

Epigenome-wide association studies (EWAS) have shown that a substantial amount of variation in DNAm exists between human populations (24–30). Therefore, if left unaccounted for, population-associated variation can interfere with the discovery of DNAm alterations associated with disease or environment. This type of confounding, often referred to as population stratification, can be addressed by inferring population-associated variation directly from DNAm data itself (31–33), as is done in genome-wide association studies (GWAS) (34). However, unlike genetic markers, epigenetic markers are tissue-specific, and therefore a DNAm-based method developed in a specific tissue or population may not generalize well to other tissues with unique DNAm profiles, such as the placenta.

In EWAS, confounding from population stratification is most often addressed using self-reported ethnicity/race to stratify study samples across the phenotype of interest. But, defining ethnicity/race is a complex task requiring the interpretation of a combination of biological and social factors leading to several complications: (i) inconsistent definition of ethnicity/race categories between individuals/organizations (35,36); (ii) self-reporting more than one ethnicity/race (37); and (iii) missing ethnicity information altogether. Ultimately, ethnicity and race are concepts that try to classify humans but are social constructs and are therefore always subject to the viewpoints and sociocultural factors that are constantly changing. To overcome the limitations of ethnicity/race categories, genetically-defined ancestry can be used in research contexts (38) as an alternative measure of population-specific variation. In contrast to the discrete nature of ethnicity/race categories, genetic ancestry can be expressed as several continuous variables based on genotype variation that reflect ancestry composition (39).

1.4.1 Existing approaches to account for population variation in EWAS

Although the importance of accounting for human population-based variation in EWAS is widely acknowledged, methods for DNAm-specific analysis are underdeveloped. Barfield 2014 developed a principal components analysis (PCA)-based approach to summarize genetic variation in PC variables that can be later incorporated into downstream analyses (31) (Table 3.1). This approach, which is highly similar and likely inspired from genotype-based methods of assessing genetic ancestry that are routinely used GWAS (40), was effective in a population of North American Caucasian and African American peripheral blood samples. In contrast to genotype-based genetic ancestry, population-specific signal in DNAm data is not nearly as strong; therefore, Barfield 2014 found improvements by using PCA on CpG sites that are nearby (<50bp) single nucleotide polymorphisms (SNPs), which are where the DNA sequence at a particular genomic position often varies between individuals, and in this case, human populations. Rahmani 2017 developed an

approach they refer to as EPISTRUCTURE, which is also a PCA-based approach using a subset of CpGs. In contrast to Barfield 2014's "proximity-based" approach, EPISTRUCTURE uses CpG sites that were directly associated with a nearby SNP. These SNPs that are associated with CpG methylation, referred to as methylation quantitative trait loci (mQTLs), were identified in one of the two cohorts included in Rahmani 2017. Although EPISTRUCTURE had improvements in capturing population-specific genotype-based variation in their DNAm data, a limitation of this type of approach is that many mQTLs are highly tissue-specific (30,41,42), and moreover SNPs can be highly population-specific (43). Since EPISTRUCTURE was developed in blood and in Europeans, it remains unclear if EPISTRUCTURE can provide reasonable performance in other tissues and populations. In contrast to Barfield 2014 and EPISTRUCTURE, Zhou 2017 developed an entirely genotype-based approach by training a random forest predictor of ethnicity, using the 59-65 SNP probes on the 450k and 850k DNAm microarrays (44). They show that their SNP predictor performs well for at least 3 broad ethnicities in USA, that they defined as White, Black or African American, and Asian. However, these SNP probes were designed for population-specific analyses in mind, and because of the small number of probes, it is unclear how robust and sensitive to additional populations is Zhou 2017's SNP predictor.

1.5 Epigenetics and placenta cell types

In the early blastocyst, the majority of gamete-derived epigenetic marks have been erased. The trophectoderm will develop into one of the layers of the chorion as well as the villous trophoblast, which retains this low methylated state. Hypomethylation of chorionic villi as a whole reflects that trophoblast is the predominant cell type within this tissue (46). In the mouse, trophoblast hypomethylation is associated with down-regulation ofthe DNA methyltransferases, Dnmt3a, Dnmt3b, and Dnmt1a (47). However, the role of DNMTs in human trophoblast is less clear (48). In contrast, the inner cell mass undergoes *de novo* global DNAm. These cells are the origin of the fetal tissues, but also contribute to primitive endoderm, from which placental endothelial cells are derived, and to the extraembryonic mesoderm, from which villus stroma is derived. In addition to differences in DNAm, the early trophectoderm shows higher H3K27 methylation as well as lower levels of histone H2A and/or H4 phosphorylation (49). Further differentiation processes in both trophoblast and other lineages are associated with additional DNAm changes and histone modifications at specific regulatory loci (50).

The tree-like structures of chorionic villi that compose the placenta consist of an outer layer of stem-like cytotrophoblasts (CTBs), which fuse together to form syncytiotrophoblast (STB). This syncytium facilitates maternal-fetal exchange and produce hormones to support pregnancy, such as progesterone, leptin,

Table 1.1: Description of methods to infer self-reported ethnicity or genetic ancestry using HM450K data.

				Sample Characteristics		
method	Statistical Approach	Input HM450K sites	Output	Tissue	Populations*	Cohort Location
Barfield et al. 2014 (31)	PCA	7703 DNAme sites with a 1000 genomes project SNP at the CpG site	Genetic ancestry as PC scores	Blood	Caucasian-Americans, African-Americans	USA
EPI-STRUCTURE (32)	PCA	4913 DNAme sites associated with local genetic variation (mQTLs)	Genetic ancestry as PC scores	Blood	Europeans, Puerto Ricans, Mexicans	Southern Germany; USA
Zhou et al. 2017 (33)	Predictive-modeling	59/65 SNP sites	Ethnicity	Multiple	White, Black or African American, Asian	Many
PlaNET (45)	Predictive-modeling	15 SNPs; 1845 DNAme sites	Ethnicity and Genetic Ancestry	Placenta	Caucasians, Asians, Africans	Canada, USA

* Ethnicity/ancestry as defined in associated study

human chorionic gonadotropin, and human placental lactogen. Transcriptional and epigenetic events control CTB differentiation and are triggered by pregnancy-associated levels in oxygen content (51); (52)]. Under low oxygen conditions, CTB fusion is impaired, and certain STB-specific CpG sites become hypermethylated (53). For example, syncytin-1, which is critical for STB formation, is down-regulated in response to hypoxia through DNMT3A-dependent hypermethylation (54). Histone deacetylase complexes (HDAC) and histone acetyltransferases (HAT) are also involved through their effects on the transcription factor GCMa, which regulates syncytin-1 expression (55).

The terminally differentiated syncytium contains nuclei derived from the fusion of CTBs at various stages in gestation. As a result, nuclei in the syncytium are heterogeneous in their age and structure. Nuclei from recently fused CTBs tend to contain a transcriptionally active euchromatic structure whereas older heterochromatic nuclei form condensed structures called syncytial knots (56). The majority of STB nuclei are transcriptionally active and lacking repressive epigenetic marks, such as DNAm, H3K9me3, and H3K27me3 (57). However, there is also an enrichment of H4K20me3, a repressive chromatin mark, that is suggested to potentially originate from the older nuclei.

CTBs can also differentiate into extravillous trophoblasts (EVTs), which exhibit strikingly different transcriptional, epigenetic, and morphological characteristics compared to STBs. During this process, expression of cell adhesion molecules and polarity are lost and expression of mesenchymal markers, migratory, and invasive properties are gained (58). Key to EVT differentiation is hypoxia-inducible factor (HIF) and Notch1 signalling, which both can be triggered by the low oxygen environment of the early gestation placenta (59–61). Transcription factors essential to the initiation of this transition process, *SNAIL* and *SLUG*, are also differentially methylated between CTBs and EVTs (62).

Histone acetylation is also essential for trophoblast differentiation, and is thought to be controlled by placental-utero oxygen content through the activity of HIF-1 and HDACs (52). The HDAC inhibitor, trichostatin-A, results in increased histone acetylation and over-expression of tumor suppressor gene *SERPINB5*, leading to reduced EVT motility and invasion (63). In contrast, acetylation of histones H2A and H2B by the CREB-binding protein (CBP) acetyltransferases inhibits the epithelial-mesenchyme transition and promotes epithelial characteristics in mouse studies (64). A higher-order chromatin structure mediated by histone modifications in combination with a locus repeat structure was demonstrated to play a role in STB-specific expression within the human growth hormone gene cluster (65), illustrating that the DNA sequence itself can affect chromatin structure.

Our knowledge of cell-specific epigenetics is confined primarily to trophoblast populations. How-

ever, additional placental cell populations are functionally important and may also be implicated in disease. Chorionic villus cultures yield the predominant growth of fibroblast cells, which have a DNAm profile that is much more similar to the fetal membranes than to whole placental villi (66). Placental-specific macrophages, Hofbauer cells (HBs), lie in the mesenchymal core of the chorionic villi and in comparison to maternal and fetal macrophages, show hypermethylation of many immune response-related and classical macrophage-activation associated genes, possibly as a result of high expression of DNMT1, DNMT3A, and DNMT3B (67).

Our current understanding of pathological diseases that affect the placenta is contributed predominantly from bulk-tissue profiling studies, which suffer from the limitation that cellular heterogeneity can result in the dilution of cell-specific signals and the creation of spurious ones. More targeted studies investigating cell-specific signatures will contribute to our understanding of changes associated with normal placental functioning as well as of abnormal ones.

1.5.1 Cell-specific placental DNA methylation

Because heterogeneous tissues, such as the placenta, are made up of several cell types (Figure 1.1), each with a distinct DNAm signature, whole-tissue measurements are ultimately an average of the DNAm signatures of the constituent cell types, weighted by their respective frequency in the bulk tissue sample. Therefore, changes in DNAm measured in complex tissues can often be attributed to variation in cell composition rather than DNAm changes that occur in the constituent cell populations (68). This makes interpretation of placental DNAm studies difficult until placental DNAm is characterized at a cell-specific resolution.

During the first few cell divisions after fertilization, there is a wave of genome-wide erasure of DNAm, followed by *de novo* DNAm in the inner cell mass (69). Deriving from the inner cell mass are fetal tissues and the mesenchymal core component of the placental chorionic villi. Within the mesenchymal core, stromal cells (SC) and HBs can be seen in the placental stroma as early as 18 days post conception (70), which are thought to derive from mesenchymal stem cells. HBs are distinct from decidual macrophages and fetal/maternal monocytes (67); they display high phenotypic diversity, promoting angiogenesis early in gestation and later participating in the immune response to pathological processes and infection (71,72). Placental vasculature is critically important for proper functioning of the placenta, and depends on the development of vessels beneath the trophoblast layer. These vessels are formed from endothelial cells (EC) that derive from the chorionic mesoderm (73). Encompassing the mesenchymal core is a thick trophoblast TB epithelial cell layer, which displays a hypomethylated profile (66). TBs comprise a set of functionally distinct subtypes, each with their own unique function (53,74): CTBs are stem-like cells that harbor regenerative

abilities and give rise to the two major trophobalst subtypes, EVTs and STBs. EVT are motile cells that travel to maternal tissue and remodel maternal vasculature, while STB are a multi-nucleated epithelial layer lining the chorionic villi that perform critical roles in hormone production and nutrient transfer.

As a consequence of its distinct developmental origin, dramatic differences in DNAm between placenta and somatic tissues have been observed (66). Globally, the placenta is hypomethylated compared to other tissues, which was originally attributed to reduced methylation of repetitive element DNAm (75,76), but was later resolved to be primarily due to placental-specific PMDs (77). It is unclear if these PMDs have a distinct function or are footprints of earlier developmental events between embryonic and extraembryonic tissues. Parent-of origin specific DNAm, which is associated with genomic imprinting, is also more commonly found in the placenta than other tissues (78). Almost all known imprinted genes are imprinted in the placenta, and many are exclusively imprinted in the placenta (79–82). Interestingly, a number of placental-specific imprinted genes are polymorphically imprinted (79). It is possible that cellular and genetic heterogeneity can contribute to polymorphic imprinting, as well as variability in DNAm generally. Supporting this, a significant role for genetic control of placental DNAm variation was recently characterized (41). These studies have contributed to our understanding of the unique epigenetics of the placenta, but it remains unclear if these features are maintained in all constituent placental cell types or are confined to specific ones.

1.5.2 Cell deconvolution

Placental DNAm is often studied in the context of disease and environmental exposures that may affect health. A common study design is the EWAS (83), where differentially methylated CpGs (DMCs) are identified in a high-throughput manner, usually with microarray or sequencing based approaches. However, placental DNAm studies are almost all carried out using whole CV and are therefore subject to challenges of interpretability due to potential cell composition variability (84). Unlike other tissues, such as adult blood and umbilical cord blood, addressing cell composition variability in placenta is difficult due to a lack of reference placental DNAm profiles, which enables bioinformatic estimation of cell composition from cellular deconvolution techniques (85). These methods operate by modelling the whole tissue measurements as a weighted sum of cell type -specific DNAm signatures, where the weights correspond to the relative proportion of each constituent cell type in the whole tissue sample, and can be determined using least-squares or non-constrained regression approaches (85–87). Without reference DNAm profiles for each cell population, researchers sometimes account for cell composition using reference-free deconvolution methods (88). However, the effectiveness of reference-free deconvolution in capturing cell composition variation has not yet been assessed in placenta.

1.6 Maternal signal in placental DNA methylation data

Illumina high-density microarray profiling for DNAm has advanced placental research by identifying epigenetic processes associated with placental function in relation to maternal environment, genetics, and perinatal health (89). Critical to these studies is the development of bioinformatic tools to process and analyze DNAm data. Recently, sample contamination, which can result in obscured measurements, has been shown to be prevalent in EWAS data (90). Sex mismatches and, mis-annotation of samples is a problem that has been previously demonstrated in other types of genomic studies (91). Heiss and Just 2018 used two bioinformatic approaches to determine whether samples measured on the Illumina Infinium HumanMethylation450 BeadChip (450k) and Infinium MethylationEPIC DNAm BeadChip (850k) microarrays are contaminated or misannotated.

The first approach estimates the amount of mixing of distinct genotypes in a DNAm array sample, by measuring the amount of outlying signal at 59-65 single-nucleotide polymorphism (SNP) probes, which increases as mixing between genetically distinct individuals occur. The second approach identifies samples that have a pattern of signal intensity on the X and Y chromosome probes that is consistent with mixing of samples of different sex. This second approach is only sensitive to those cases where the sexes of the mixed samples are distinct, Despite most epigenome-wide association studies (EWAS) having some level of data quality control and sample preprocessing. Heiss and Just 2018 demonstrate that misannotation and sample contamination is relatively common in DNAm studies, but also importantly note that contamination is identifiable by their approaches and therefore addressable in downstream analysis.

In placental research, where conceptus-derived chorionic villi are normally sampled, there is an added challenge of avoiding contamination with maternal tissue, as the placenta is directly attached to the maternal uterine wall and bathed in maternal blood (91). Because maternal cells from decidual tissue are distinct in both function and epigenetic profile than placental (conceptus-derived) cells, contamination from maternal tissues can obscure measurements of DNAm and make interpretation difficult (92). While small numbers of maternal macrophages are a normal component of placental cell composition throughout gestation (93), placental inflammation can sometimes be associated with infiltration of maternal immune cells into the intervillous space and even the villi themselves (e.g. acute chorioamnionitis (94)). Due to these technical and biological factors that can contribute to maternal DNA presenting in placental samples, it can be difficult to determine whether DNAm variation reflects phenotype-associated epigenetic reprogramming or if simply the underlying cell composition is changing.

Although careful sample processing can generally avoid most maternal contamination, knowing

the degree of maternal DNA contribution for each sample would aid downstream analyses and may be relevant to the biological processes of interest. Despite this, the degree that maternal DNA contamination influences placental studies has been minimally explored. Wan 2019 describe a set of quality control measures based on clustering DNAm patterns between control placental and maternal tissue samples, and partially methylated domains (PMDs), that were used to infer that early termination pregnancies are especially at risk of contamination during sample processing (95). PMDs are long regions of intermediate/low DNAm surrounded by regions of higher DNAm that exist in a highly cell-specific fashion (96). In particular, they emphasized that samples obtained from early termination pregnancies are especially at risk of contamination during sample processing. Maternal contamination can also present a risk to cord blood studies. Morin 2017 used tissue-specific approaches to assess maternal contamination in DNAm data derived from cord blood, such as using highly cord blood -specific sites to predict contaminated and non-contaminated samples (97). Interestingly, they explored using quality measurements (i.e. "no-calls") from genotyping data for samples that had also been measured on the Illumina PsychChip (genotyping array) to confirm contamination calls, but in contrast to Heiss 2018, found that the sensitivity of the genotyping data for genotype mixing was poor.

1.7 Research objectives and hypothesis

Previous work in placental DNAm research involves mostly EWAS-type study designs assessing placental phenotypes such as preeclampsia, fetal growth restriction, chorioamnionitis, neural tube defects, and environment. However, these studies are limited to existing bioinformatic tools, which have not been extensively explored for their appropriateness and effectiveness in placental data. Moreover, the underlying DNAm of the various cell components of the placental chorionic villi has not yet been characterized.

In this dissertation I aimed to characterize and identify the major sources of variation in placental DNAm, and develop bioinformatic approaches to infer, model, and characterize these variables from placental DNAm data directly. I hypothesize that cell composition, genetic ancestry and ethnicity, and maternal DNA contamination can be identified and predicted from placental chorionic villi DNAm data. To assess this, I performed:

1. A study of developing and comparing approaches to predict ethnicity and genetic ancestry from placental DNAm data using multiple placental GEO datasets.
2. A study of characterizing the DNAm of placental cell types across gestation, and developing a reference to estimate cell composition from placental chorionic villi DNAm.

3. A study of using bioinformatics to identify maternal signal from placental DNAm data from previously published placental DNAm studies.

These studies contribute to further insights into major contributors of variation in placental DNAm, and provide useful bioinformatic approaches for conducting better placental DNAm research.

2 Accurate ethnicity prediction from placental DNA methylation data

2.1 Introduction

Population stratification or confounding effects from variation in genetic ancestry can confound placental EWAS. However, methods to estimate population effects either by inferring genetic ancestry (31,32) from highly genotype-associated CpG methylation (32) or by predicting ethnicity, may have population- and tissue- specific effectiveness.

In this study, we developed a placenta-specific approach to predicting ethnicity, which we refer to as planet (Placental DNAm Elastic Net Ethnicity Tool). Planet uses DNAm and genotyping data measured on the 450k array and was developed using multiple cohorts of placentas from North America. To ensure compatibility with future studies, planet was developed on overlapping sites from 450k and the newer 850k array. We show that planet out-performs existing methods in predicting ethnicity in placental DNAm data and can produce accurate measures of genetic ancestry. Our method can be used to classify individuals into discrete groups based on self-reported race/ethnicity or to describe individuals on an ancestral continuum that more accurately reflect the nature of human populations. In this study, self-reported ethnicity/race categories sometimes differed between cohorts. Ethnicity defined in this study is a collection of self-reported ethnicity and/or race information that was recoded such that populations were grouped into 3 populations with high cultural similarity and common ancestral origin (recoding: "African" = African / African American / Black; "Caucasian" = Caucasian / Non-hispanic White; "Asian" = East Asian / Asian).

In studies where ethnicity information is unavailable, planet can be applied to predict ethnicity after obtaining DNAm data, and used to investigate population-specific differences or to minimize confounding by population stratification in downstream DNAm statistical analyses.

2.2 Results

2.2.1 Datasets

Our goal was to develop a placental DNAm-based ethnicity classifier, which could learn ethnicity-specific DNAm patterns from one set of samples in order to assign ethnicity labels to a new set of samples. We searched for placental 450k data on the Gene Expression Omnibus (98) that contained more than one ethnicity group and made sample-specific ethnicity information available (Table 2.1). Five distinct cohorts met these criteria (labelled C1-C5), with three major North American ethnicities represented by sufficiently

large numbers across more than one dataset: African (n = 58), Asian (n = 53), and Caucasian (n = 389). We opted to include samples from both healthy and abnormal pregnancies (preeclampsia, gestational diabetes mellitus, fetal growth restriction or overgrowth) (Table 2.1) (79,99–104). Though there were significant cohort-specific effects on DNAm that may reflect batch/technical variation (Figure A.1), we included these multiple datasets and phenotypes to enable the development of a robust classifier that would generalize well in future studies (105).

2.2.2 Development of a placental DNA methylation ethnicity classifier

To determine the best machine learning classification algorithm that could learn ethnicity-specific patterns from DNAm microarray data, we compared four algorithms previously shown to be well-suited for prediction using high-dimensional genomics data (105–107): generalized logistic regression with an elastic net penalty (GLMNET) (108,109), nearest shrunken centroids (NSC) (106), k-nearest neighbours (KNN) (110), and support vector machines (SVM) (111). For each algorithm, hyperparameter(s) were selected (e.g. k number of neighbours for KNN) that resulted in the highest performance estimated by repeated five-fold cross validation (three repeats). All algorithms performed favorably (logLoss = 0.170 - 0.276; Figure A.2), except KNN (logLoss = 1.82). However, all algorithms showed a bias for high predicatability of Caucasians (average accuracy = 0.980), and low predictability of Asians (average accuracy = 0.448) (Figure A.2). Considering overall- and ethnicity-specific performance, the GLMNET algorithm was used for the remainder of the study (accuracy = 0.866, 0.625, 0.998 for Africans, Asians, and Caucasians, respectively), and we refer to this classifier as planet (Placental DNAm Elastic Net Ethnicity Tool).

For each sample, planet returns a probability that the sample is African, Asian or Caucasian and the final classification is defined by the ethnicity class with the highest of these probabilities. We reason that these probabilities have the potential to identify samples with mixed ancestry or ethnicity. Therefore, we implemented a threshold function on planet's probability outputs that classifies samples as 'Ambiguous' if the highest of the three class-specific probabilities is below 0.75 (Material and Methods, Figure A.3). This resulted in 7 self-reported African, 12 Asian, and 13 Caucasian samples as being classified as ambiguous, which led to a slight decrease in performance (Figure 2.1a). However, we note that because genetic ancestry exists on a continuum and due to the limitations of self-reported ethnicity, there are likely to be individuals of mixed ancestry/ethnicity in our sample set, and therefore hypothesize that a model that includes an ambiguous class is more realistic and accurate than one without. Cross validation, where training/validation subsets were created based on cohort-identity, yielded an overall accuracy of 0.900, a Kappa of 0.738, and a positive predictive value of 0.944 (Figure 2.1a), which was consistent when examining performance by

Table 2.1: Description of 450k DNAm datasets used to develop and test planet.

Cohort	n	GEO accession	Description	Location	Self-reported ethnicity			Additional genetic data
					AFR (n=57)	ASI (n=53)	CAU (n=389)	
C1 (103)	72	GSE70453	36 controls, 36 gestational diabetes mellitus	Boston, MA, USA	13	13	46	NA
C2 (101)	24	GSE73375	13 controls, 11 preeclampsia (PE)	Chapel Hill, NC, USA	13	1	10	NA
C3 (102)	289	GSE75248	289 samples from infants with variable newborn neurobehaviour	RI, USA; MA, USA	23	9	257	NA
C4 (100)	44	GSE100197	17 controls, 27 PE	Toronto, CAN	7	12	25	50 AIMs
C5 (99)	70	GSE100197, GSE108567, GSE74738, Unpublished	35 controls, 35 fetal growth restriction, PE, and/or preterm birth	Vancouver, CAN	1	18	51	50 AIMs; Omni2.5

AFR African, ASI Asian, CAU Caucasian

dataset (Figure A.4).

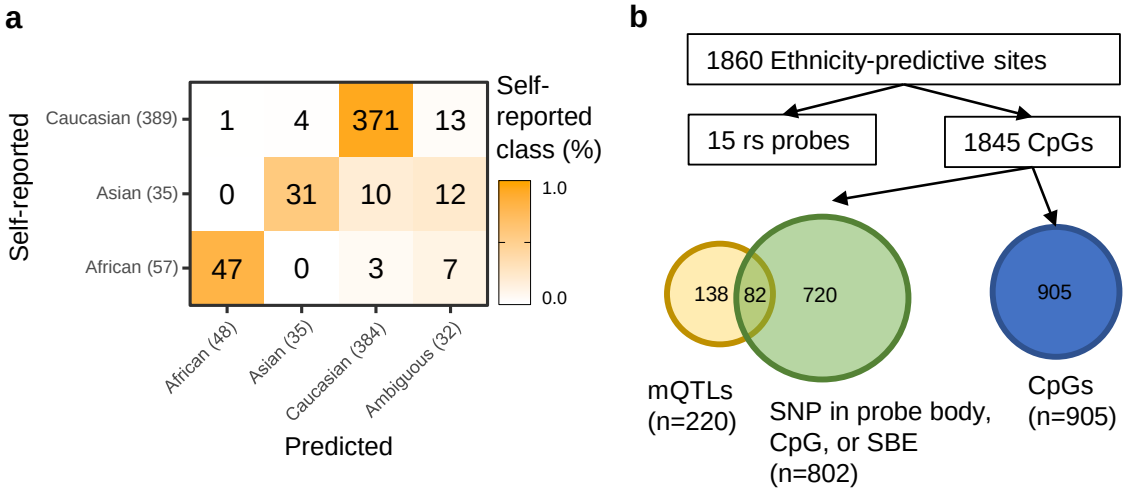

Figure 2.1: Evaluating planet's performance and characterizing ethnicity-predictive 450k sites. We developed planet (Placental elastic net ethnicity classifier), using placental 450k data and evaluated its classification performance using leave-one-dataset-out cross validation. **a** Each sample's ethnicity classification is shown with respect to their self-reported ethnicity. Samples were called 'ambiguous' if their predicted probability fell below a 'confidence' threshold of 75%. **b** planet utilizes a subset of ethnicity-predictive sites from the 450k. To investigate whether genetic signal is present in the measurement for these sites, we cross-referenced ethnicity-predictive sites to an existing placental mQTL database (41) and determined whether any sites had SNPs present in either the probe body, CpG site of interrogation, or single base extension sites, based on dbSNP137

2.2.3 Ethnicity-predictive sites on the 450k array are largely linked to genetic variation

To better understand the basis of planet's ethnicity prediction, we examined the 1860 sites automatically-selected by the GLMNET model. These sites were enriched for SNP probes, containing 15 of the 59 SNPs explicitly measured on both 450k and 850k DNAm arrays ($p < 1e\text{-}16$). Of the remaining 1845 DNAm sites, we found significant enrichment for sites linked to genetic variation: 802 sites (43.1%) have a documented SNP in either the probe body, CpG site of interrogation, or the single base extension site ($p < 1e\text{-}16$) (112), and 220 sites (11.8%) corresponded to previously identified placental-specific methylation quantitative trait loci (mQTLs) (41) ($p < 1e\text{-}16$, Figure 2.1b). With respect to chromosomal location, we found significant enrichment for ethnicity-predictive sites on chromosomes 2 ($p < 0.01$), 15 ($p < 0.05$), and 17 ($p < 0.05$) (Figure A.5 file 2: Figure A.5). With respect to CpG density, we found significant enrichment for ethnicity-predictive sites in OpenSea ($p < 0.001$) and South Shore ($p < 0.05$) regions (Figure A.5), where relatively neutral (unselected) genetic variation is more likely to be located (113). Pathway analysis for GO and KEGG terms for genes associated with the 1860 sites, found only one significant ($p < 0.05$) GO term (homophilic cell adhesion via plasma membrane adhesion molecules).

2.2.4 DNAm -inferred ethnicity and genetic ancestry

To test the ability of planet to identify individuals of mixed ancestry, we examined whether samples classified as 'ambiguous' were also intermediate with respect to genetically-defined ancestry. Genetic ancestry was inferred from 50 ancestry informative genotyping markers (AIMs) in samples from cohorts C4 and C5 (n = 109), using 1000 Genomes Project samples as reference populations (43,114). These 50 markers were previously selected based on their ability to differentiate between African, European, East Asian, and South Asian populations (114). Plotting the first two multi-dimensional scaling coordinates calculated on the 50 AIMs in (Figure 2.2), shows a handful of samples intermediate to three more distinct ancestry clusters. The samples with less extreme genetic ancestry coordinates based on AIMs tended to have lower planet-calculated probabilities associated with the ethnicity classification matching the individual's self-reported ethnicity (Figure 2.2) confirming that planet provides some information on the genetic ancestry composition.

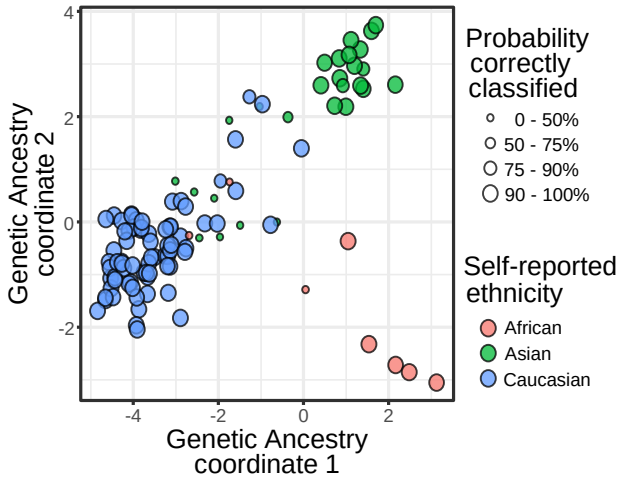

Figure 2.2: Probabilities associated with planet ethnicity predictions and genetic ancestry inferred from AIMs. Probabilities associated with planet ethnicity predictions and genetic ancestry inferred from AIMs. Ethnicity classifications from planet and associated confidence/probability scores were compared to genetic ancestry inferred from 50 AIMs (n = 109, cohorts C4, C5), represented by the first three coordinates from multidimensional scaling using 1000 genomes project samples as reference populations

Although genetic ancestry can be adequately inferred from a small set of AIMs, it is best obtained from a large number of unlinked markers (115). Therefore, we also inferred genetic ancestry in a smaller number of samples from C5 (n = 37) with high density genotyping array data (Omni 2.5, >2.5 million SNPs), again using 1000 Genomes Project samples as reference populations (43,116,117), and compared this to planet's predicted membership probabilities for each ethnicity (Figure 2.3a-c). 10 of these 37 samples were not initially used for previous analyses due to a lack of available self-reported ethnicity information (Figure 2.3a). We found that genetic ancestry coefficients reflected the probabilities associated with ethnicity

classification to a high degree (Figure 2.3bc, R2 = 0.95-0.96, p < 0.001).

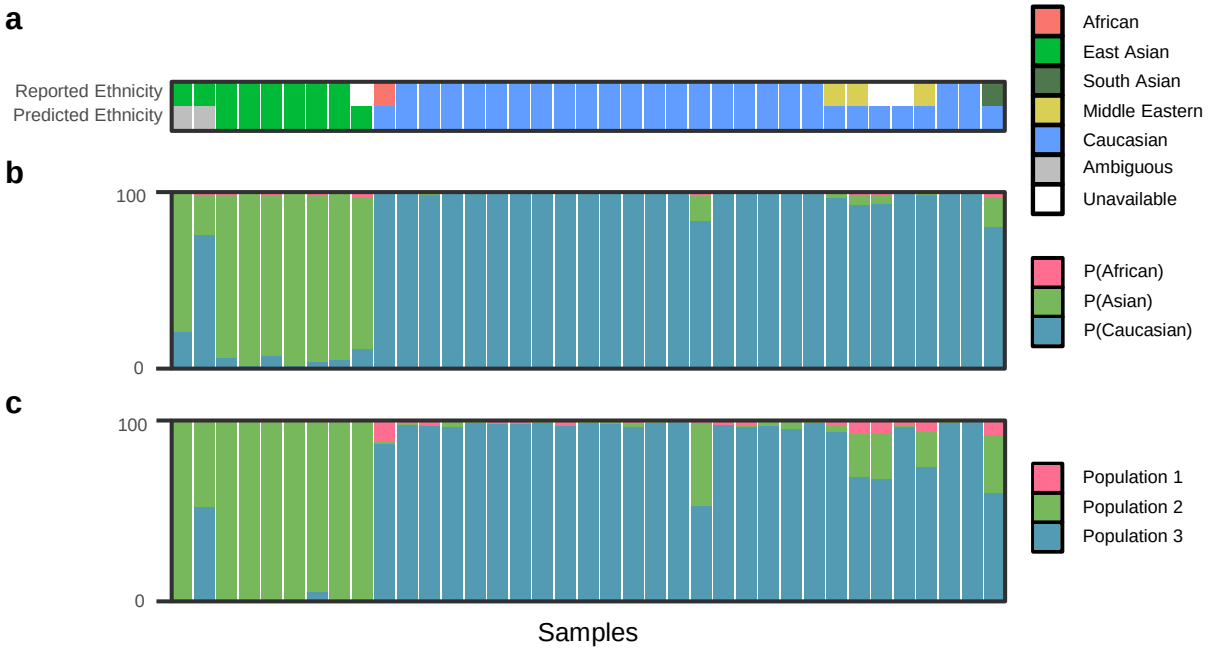

Figure 2.3: Probabilities associated with planet ethnicity predictions and genetic ancestry inferred from high density genotyping data. Probabilities associated with planet ethnicity predictions and genetic ancestry inferred from high-density genotyping data. planet was tested in a subset of cohort C5 (n = 37). **a** planet's ethnicity classifications were compared with self-reported ethnicity. **b** Ethnicity probabilities generated by planet were compared to **c** genetic ancestry coefficients determined from high-density genotyping data (Omni 2.5, > 2 million SNPs), using the function snmf() from the R package LEA, and found to be highly correlated (R2 = 0.95–0.96, p < 0.001) determined by linear regression.

2.2.5 Characterizing existing methods to infer population structure in placental DNA methylation data

To evaluate our hypothesis that a placental-specific approach to population inference would outperform existing methods developed in other tissues, we compared the performance of planet to three previously published 450k methods: Barfield's SNP-based filtering approach (31), EPISTRUCTURE (32), and Zhou's SNP-based classifier (33). To address the differences in the type of outcomes produced by each method (e.g. PCs or ethnicity classifications), we used PCA to generate metrics that could be compared between methods. PCA was performed on the set of 450k sites corresponding to each method (Table 1) to determine the amount of variance explained in self-reported ethnicity (Figure 2.4a; n = 499, cohorts C1-C5), genetic ancestry (Figure 2.4b,c; n = 109, cohorts C4 and C5 only), and cohort-specific patient variables (e.g. microarray batch, sex, gestational age; Figure A.6), by each of the top ten PCs corresponding to each of the four population inference methods. For computation of PCs on planet's sites, we used a cohort-specific cross validation

framework to account for bias that could be introduced by using the same samples for development and testing. Specifically, planet's PCs were computed separately for each cohort using ethnicity-predictive sites selected in all other cohorts (methods). We found that for all cohorts, the first two PCs computed on planet's sites and the 59 SNPs was highly correlated with self-reported ethnicity (Figure 2.4a, R2 = 0.649 ± 0.087, 0.697 ± 0.110, respectively), genetic ancestry coordinate 1 (Figure 2.4b, R2 = 0.680 ± 0.086, 0.721 ± 0.019), and genetic ancestry coordinate 2 (Figure 2.4c, R2 = 0.296 ± 0.418, R2 = 0.356 ± 0.497; Figure 2.4a). In contrast, the first PC computed on Barfield's and EPISTRUCTURE's sites showed almost no correlation with self-reported ethnicity (Figure 2.4a, R2 = 0.0452 ± 0.060, 0.066 ± 0.082), genetic ancestry coordinate 1 (Figure 2.4b, R2 = 0.044 ± 0.060, 0.040 ± 0.055, respectively), or genetic ancestry coordinate 2 (Figure 2.4c, R2 = 0.0178 ± 0.0236, 0.0228 ± 0.0321). Instead, for Barfield and EPISTRUCTURE, the PCs that correlated with ethnicity/ancestry were confined to PCs 3-6 (Figure 2.4a), while often the top PCs (e.g., 1-4) for these two methods were associated with variables other than ethnicity/ancestry (A.6). For example, in cohort C4, EPISTRUCTURE PC1 was most correlated with row position on the 450k array (R2 = 0.482), PC2 with gestational age (R2 = 0.315), PC3 with genetic ancestry coordinate 1 (R2 = 0.450) and PC5 with ethnicity (R2 = 0.579; A.6).

Limiting to methods that predict ethnicity classes, we compared the performance of planet to Zhou et al. 2018's SNP-based classifier (Figure A.7). Both classifiers demonstrated similar accuracy in classifying self-reported Africans (87.1% for planet; 90.3% for Zhou) and Caucasians (96.7% vs 97.9%), but planet was more accurate in classifying self-reported Asians (74.4% vs 41.0%).

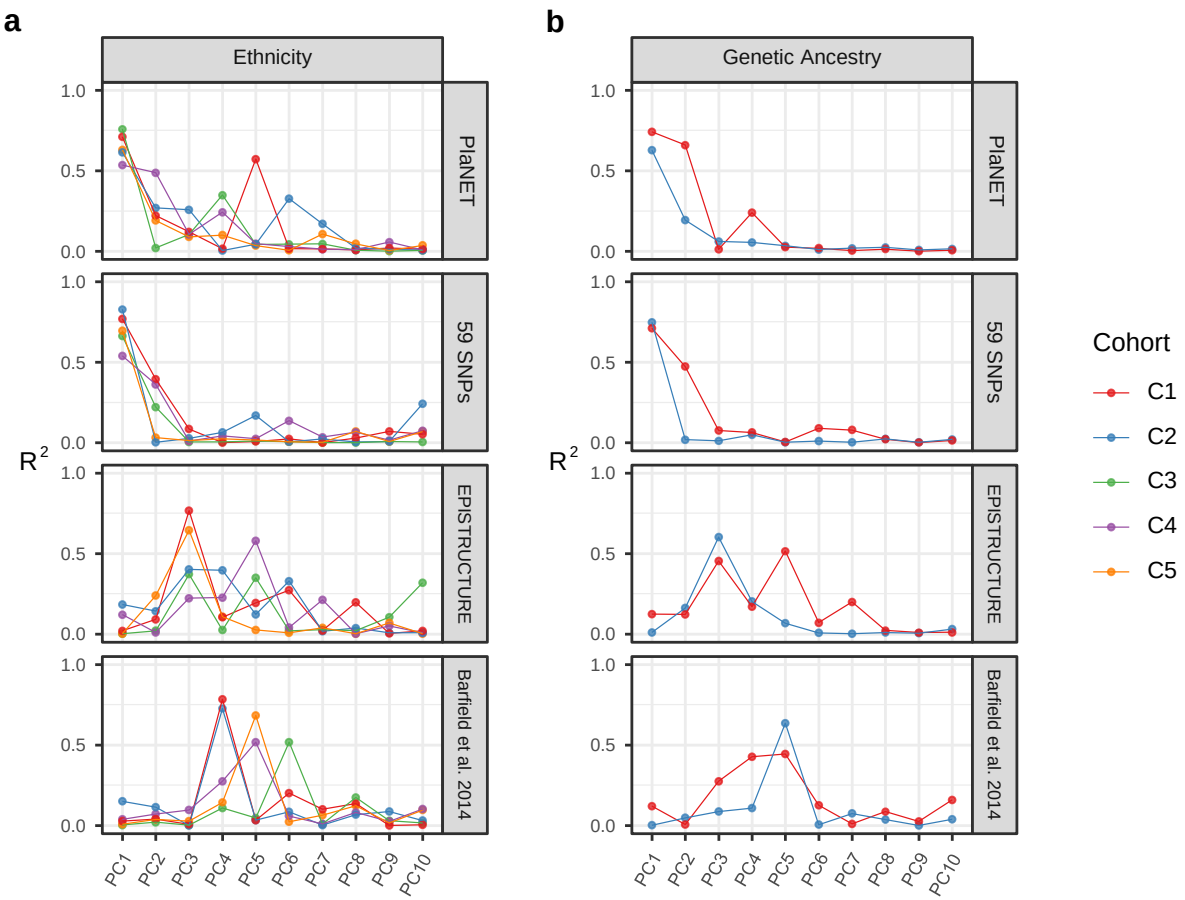

Figure 2.4: Comparing planet to existing methods to account for population stratification using 450k data. **a** For each cohort, principal components analysis was conducted on planet using a model trained on all other cohorts. planet's principal components (PCs) were then compared to the PCs computed on sites from EPISTRUCTURE (32), Barfield's method (31), and the 59 SNPs. a Amount of variance explained from a series of linear models where principal component "i" is a function of self reported ethnicity encoded as a dummy variable. **b** This was then repeated using AIMs coordinates 1 and 2 instead of ethnicity as the independent variable (n = 109)

2.2.6 Application of planet in an EWAS setting

Lastly, to demonstrate the utility of applying planet to placental DNAm data, we applied planet to obtain ethnicity classifications across two previously published EWAS studies using three datasets (Table 2.2, Figure A.9). We note that this includes samples from cohorts C4 and C5 that were used to develop planet.

One study used two distinct cohorts from Vancouver, Canada (GSE100197, = 102) and Toronto, Canada (GSE98224, n = 48) to investigate DNAm alterations associated with preeclampsia status (99). We reasoned that correction for ethnicity should decrease false positives in the EWAS and therefore increase concordance between hits identified in the two data sets. In the original EWAS, with no adjustment for ethnicity, our group reported that 599 out of the 1703 (35.1%) significant associations found in the discovery cohort

were also significant in the validation cohort, and the correlation of the difference in mean DNAm between controls and preeclampsia-affected samples (i.e. delta betas) at FDR significant sites between discovery and validation was 0.62 (99). When we repeated the analysis while adjusting for ethnicity determined by planet, the number of preeclampsia-associated sites that overlapped between cohorts increased to 651/1614 (40.3%) [Table A.3], and the correlation between delta betas increased to 0.66. We also found that repeating gene set enrichment analysis, which originally found nothing significant (99), yielded several significantly enriched GO terms such as developmental process, inflammatory response and cell adhesion.

Next, because adjustment for population stratification can not only be done via correction in linear modelling, but can also be done by stratifying an analysis by population identity, we performed a secondary EWAS confined to samples predicted as Caucasians (n = 71/102 for discovery, n = 28/48 for validation). This resulted in a decrease in overlap in preeclampsia-associated sites between cohorts: 359/1488 (17%) [Table A.4], although the correlation between delta betas remained high (r = 0.67), indicating the observed decrease in overlap between significantly differentially methylated sites was likely due to a decrease in power from smaller sample size (particularly in the validation group) rather than a decrease in concordance between cohorts.

The application of planet can be useful for checking for discrepancies in self-reported ethnicity information. We tested whether planet could identify the ethnicity of samples from an all-Caucasian population. GSE71678 (n = 343), a cohort not used in the development of planet, consisted of DNAm data from placental samples collected from a New Hampshire, USA birth cohort that investigated the effects of arsenic exposure on placental DNAm (118). 342 samples were classified as Caucasian by planet, and 1 sample had a high probability of belonging to the Caucasian group (Probability = 0.73) but was below our confidence threshold and was therefore classified as 'ambiguous,' confirming ethnic homogeneity was high in this cohort and adjustment for population stratification was not needed in this study.

2.3 Discussion

In this study, we developed planet, a method to predict Asian, African, and Caucasian ethnicity using placental 450k array data. To enable compatibility with future studies, planet was developed on sites (452,453 CpGs and 59 SNPs) overlapping between 450k and 850k DNAm arrays. Although all samples in this study were reported as a single ethnicity/race, we expected that there would be significant population substructure that might limit our ability to develop predictive models of ethnicity and to assess their performance. Despite this limitation, ethnicity could be predicted with high accuracy as assessed by cross validation. planet's DNAm-based ethnicity classification relies on 450k sites with large amounts of genetic signal, which supported

Table 2.2: Distribution of planet ethnicity predictions across previously published placental EWAS datasets.

GEO Accession	Primary groups	African	Asian	Caucasian	Ambiguous
GSE98224	EOPET	5	4	10	0
	Preterm Controls	1	3	5	0
	LOPET	1	1	8	1
	Term Controls	0	4	5	0
GSE100197	EOPET	1	5	15	1
	Preterm Controls	1	4	19	0
	LOPET	0	6	12	0
	Term Controls	0	2	17	0
	IUGR	0	3	8	0
GSE71678	NA*	0	0	342	1

EOPET - Early Onset Preeclampsia, LOPET - Late Onset Preeclampsia.
* Phenotype of interest is a continuous variable (arsenic concentration).

our initial efforts to filter our data to enrich for genetic-informative sites prior to classifier development (methods) [41, 50, 51]. When examining planet's 1,860 sites used to predict ethnicity, more than half could be linked to a nearby genetic polymorphism. Of these, 802 CpG sites have documented SNPs in their probe body, single base extension or CpG site of interrogation, which previously have been identified to differ between European and East Asian populations (112). Several studies have suggested the genetic influence on DNAm at these sites is primarily technical in nature (112,119,120), suggesting the patterns in DNAm at these sites are likely tissue-agnostic, warranting further investigation in their utility in predicting ethnicity and/or genetic ancestry in tissues other than the placenta. A significant proportion of other ethnicity-predictive CpG sites (n = 220) were previously found associated with placental mQTLs in a population with similar demographics to the ones studied here (41). This finding, together with EPISTRUCTURE—a method that also relies on mQTLs (32)—suggests that leveraging the tissue- and population- specificity of mQTLs can produce highly effective DNAm -based population structure inference methods.

Of the existing methods to assess population stratification from DNAm data, we note that Barfield's method and EPISTRUCTURE infer continuous measures of genetic ancestry, while Zhou's SNP-based classifier returns discrete ethnicity classifications, however ours produce both (31–33) (Table 1.1). EPISTRUCTURE and Barfield's method are unsupervised PCA-based approaches, which rely on the empirical observation that specific DNAm sites can be highly correlated with PCs computed on genome-wide genotype data in adult blood samples (31,32). However, we found that DNAm at these sites did not produce PCs that are highly associated with genotype data in placental samples. Instead, top PCs were more often associated with non-ancestry related variables in the placental samples included in this study, such as gestational age,

preeclampsia, and technical variables. Ethnicity and genetic ancestry -associated PCs were confined to the third to sixth component of variation, suggesting that application of these methods may require filtering of PCs to those that are ethnicity / ancestry-specific, which is impossible when self-reported ethnicity and genetic ancestry information is unavailable (i.e. when these methods are needed most). Future improvements to these types of methods can aim at improving the amount of ethnicity and genetic ancestry -associated signal in the sites used to ensure the top two-three PCs are always associated with ethnicity and ancestry. This aim could also be supported in identifying ethnicity and ancestry -associated sites that are also robust to changes in non-genetic drivers of DNAm such as cell type, gestational age, and severe pathology.

Supervised population inference approaches such as ethnicity classifiers can return an explicit assignment of samples into distinct groups. In comparison to self-reported ethnicity, an assessment based on DNAm/genetic data is more objectively defined, which allows for more robust investigation of ethnicity-specific effects. An important goal of any population structure inference method would be to account for samples of mixed ancestry, a capability not well supported by Zhou's ethnicity classifier (33). In contrast, planet produced membership probabilities corresponding to each ethnicity group that were highly correlated with genetic ancestry estimated from genotyping data. This was consistent whether we used principal components analysis on AIMs data, or model-based estimation of ancestry on high density genotyping array data (116,121–123). In this study, we defined samples of potential mixed ancestry as those with a maximum membership probability of less than 0.75, but we note that this threshold can be manually adjusted by the user, and that the probabilities themselves can be used to adjust for population structure in study populations including significant numbers of samples with mixed ancestry.

Results of DNAm studies on genetic ancestry and ethnicity, such as this one, depend on the number and proportion of different populations sampled from, as well as the tissue studied. Due to limitations in sample availability, only African, Asian, and Caucasian ethnicities were included in our study. However, we note that these ethnicities are among the most common in North American populations—but future developments should consider inclusion of additional ethnicities. Furthermore, due to limited number of samples with high density genetic data, we were unable to address the extent of finer population structure that likely exists within the major ancestral groups studied. Differences in ethnic composition in samples from our study and samples used to develop Barfield's method and EPISTRUCTURE may also explain why Barfield's method or EPISTRUCTURE performed poorly in our study (31,32). A lack of generalizability of these methods to our placental samples was likely further compounded by the use of different tissues to develop each method—Barfield and EPISTRUCTURE were both developed and tested in blood tissue only. This is especially important to consider when applying these techniques to tissues with unique DNAm

profiles, such as placenta (66). It is possible that application of these approaches to other tissues that are more similar to blood (e.g. other somatically-derived tissues) may result in better performance compared to when applied to placenta as seen in this study. However, any DNAm-based test needs to be validated before application to new tissues, which has not yet been done for these methods.

A major goal of EWAS is to uncover signal truly associated with the phenotype/environment of interest that might generalize to other relevant populations. This is challenging given the wide host of technical variables that can affect DNAm measurements and the common finding that many phenotypes are associated with relatively small effect sizes (104,124). To this end, adjustment for major confounders such as genetic ancestry or ethnicity can significantly improve EWAS. We demonstrated, in a reanalysis of our previously published PE placentas, that adjustment for ethnicity, determined by planet, improved the replicability of significant associations between independent cohorts. Conversely, overadjustment can occur when populations are relatively homogeneous, resulting in bias and/or loss of precision. We showed that planet can indicate minimal population stratification when applied to a homogenous Caucasian population. Thus, planet will be useful in assessing population stratification in future placental EWAS, as well as conducting ethnicity-stratified analyses, which may lead to important insights into the disparities between populations of pregnancy-related outcomes (125–127).

2.4 Conclusion

We demonstrated that ethnicity and genetic ancestry can be accurately predicted using placental HM40K DNAm microarray data with respect to three major ethnicity/ancestral populations. Although samples that were used to develop planet were reported to come from single ethnic populations, our classifier was able to capture mixed ancestry, and outperformed existing prediction methods. Planet will be valuable in assessing and accounting for population stratification, which can confound associations between DNAm with disease or environment, in future studies using 450k or 850k arrays. The machine-learning approach used to develop planet can easily be applied for other tissues and populations for use in future DNAm studies.

2.5 Methods

2.5.1 Collection of previously published placental 450k DNA methylation data

Placental DNAm data from liveborn deliveries of healthy and mixed pregnancy complications (n = 585), were combined from seven GEO 450k datasets corresponding to five North American cohorts (summarized in Table 2.1) (79,98,99,101–103). Five unpublished samples from the C5 cohort were included and are available at GSE128827. Gestational ages of these pregnancies at delivery ranged from 26 to 42 weeks and

50.30% of samples were male. Samples were excluded (n=67) if their self-reported ethnicity was missing or did not fall into one of three major race/ethnicity groups: Asian/East Asian (n=53), Caucasian/White (non-hispanic) (n=389), or African/African American/Black (n=57). Based on census data (128), we note that self-reported Caucasian/White (non-hispanic) samples are typically of European ancestry, self-reported Asians are typically of East Asian ancestry and self-reported African represent diverse ancestries from Africa with a significant potential of admixture from other ancestries (129). When possible, data was downloaded as raw IDAT files (GSE75248, GSE100197, GSE100197, GSE108567, GSE74738), otherwise methylated and unmethylated intensities were utilized (GSE70453, GSE73375).

2.5.2 DNA methylation data processing

All samples were analyzed using the Illumina Infinium HumanMethylation450 BeadChip array. Array data analysis was performed using R version 3.5.0. To allow compatibility of planet with the newer 850k microarray, the raw 450k data (485,512 CpGs, 65 SNPs) was filtered to the 452,453 CpGs and 59 SNPs common between both platforms prior to classifier development (33). Because genetic variability can capture ancestry information, we omitted the common filtering step that would remove sites with probes that overlap SNPs (n = 52,116 at a minor allele frequency > 0.05). CpGs were removed if greater than 1% of samples had poor quality signal (bead count < 3, or a detection p-value > 0.01; n = 14,858). The remaining poor quality measurements were replaced with imputed values using K-Nearest Neighbours from the R package impute (130). Cross-hybridizing (n = 41,937) (119,120) and placental-specific non-variable sites (n = 86,502) (131) were also removed, leaving 319,233 sites for classifier development.

Biological sex was determined by hierarchical clustering on DNAm measured from sites on the sex chromosomes and then compared to reported sex. Samples with discordant reported and inferred sex were removed (n=3). Samples were also removed if they had a low mean inter-array correlation (< 0.95, n = 5). Intra-array normalization methods, normal-exponential out-of-band (NOOB) (132) and beta mixture quantile normalization (BMIQ) (133) were used from R packages minfi (version 1.26.2) (134) and wateRmelon (version 1.24.0) (135) to normalize data.

2.5.3 Genotyping data collection and genetic ancestry assessment

In a subset of C5 (n = 27) and 10 additional samples, high density SNP array genotypes were collected. DNA samples from one site from the fetal side of each placenta were collected as previously described (136) and quality was checked using a NanoDrop ND-1000 (Thermo Scientific) as well as by electrophoresis on a 1% agarose gel. Genotyping at ~2.3 million SNPs was done on the Illumina Infinium Omni2.5-8 (Omni2.5)

array at the Centre for Applied Genomics, Hospital for Sick Kids, Toronto, Canada. For inferring genetic ancestry, the data for these 37 samples was combined with a previously processed 1000 Genomes Project Omni2.5 dataset (n = 1,756) to use as reference populations (43,117). Genotypes in this combined dataset were filtered for quality (missing call rate > 0.05, n removed = 31,604), minor allele frequency (MAF > 0.05, n removed = 114,628), and linkage disequilibrium pruning was performed to select representative SNPs (R2 < 0.25, n removed = 919,824) for a final dataset of 218,732 SNPs and n = 1793 samples. Genetic ancestry coefficients were estimated using the R package LEA, which utilizes sparse non-negative matrix factorization to produce similar results to model-based algorithms ADMIXTURE and STRUCTURE (116,123). Cross-entropy criterion was used to assess the number of ancestral populations (Figure A.8) (137).

A smaller panel of 50 ancestry-informative genotyping markers (AIMs) was collected in a subset of samples from cohorts C4 (n = 41) and C5 (n = 68). AIMs were selected based on their ability to differentiate between African, European, East Asian, and South Asian populations (138–140). Results from cohort C5 have been published elsewhere (114), and genotyping data was collected for cohort C4 in the same manner. Briefly, these markers were measured in placental villus DNA using the Sequenom iPlex Gold platform (Génome Québec Innovation Centre, Montréal, Canada). Genetic ancestry inferred from 50 AIMs markers was computed using multi-dimensional scaling after combining with the same 50 AIMs from the 1000 Genomes Project samples, as previously described (114).

2.5.4 Developing the ethnicity classifier and assessing its performance

To develop and assess the performance of planet we used a 'leave-one-dataset-out cross-validation' (LODOCV) approach. This approach uses four out of five datasets to develop a predictive model (training), which is then used to generate ethnicity classifications on the samples in the remaining dataset (testing). This differs from the traditional cross validation approach of randomly splitting the full dataset into training and testing. LODOCV produces more accurate estimates of classifier performance for future studies, and has been previously used for evaluating age-predictive models (141). Each iteration of LODOCV generates dataset-specific estimates of performance (accuracy, Kappa). After all iterations, overall performance was assessed by aggregating classifications across all datasets.

For fitting predictive models within LODOCV-generated training sets, we used the R package caret (142) Several algorithms were compared: logistic regression with an elastic net penalty (GLMNET) (143,144), nearest shrunken centroids (NSC) (145,146), K-nearest neighbours (KNN) (147), and support vector machines (SVM) (111). To determine optimum tuning parameters for each algorithm (e.g., 'k' number of neighbours for KNN, alpha and lambda for GLMNET), we built several models while varying the tuning

parameter(s) and compared the performance of these models within each training set using repeated (n = 3) five-fold cross validation. Hyperparameter values were left as default settings in caret (148), or a grid of values for GLMNET (alpha = 0.025 – 0.500, lambda = 0.0025 – 0.2500). We compared the performance of these models using accuracy, positive predictive value, cohen's Kappa (149), and logLoss (a measure of classification accuracy that heavily penalizes over-confident misclassifications). After assessing the classifier performance using LODOCV, a final GLMNET model was fit to the entire dataset (cohorts C1-C5) using the same model fitting procedure described above and is available for use in future datasets (https://github.com/wvictor14/planet).

2.5.5 Enrichment analysis

The DNAm sites and SNPs selected to predict ethnicity in this final model (n = 1860) were used for enrichment analysis. For DNAm sites, we looked for enrichment for SNPs in the probe body, CpG site, and single base extension sites based on Illumina's 450k annotation version 1.2 (150). We looked for enrichment for placental mQTLs (41), chromosomes and CpG islands (HG19; Figure A.5). Fisher's exact test was used for all enrichment tests using a p-value threshold of < 0.05, and was carried out in R using the function fisher.test(). GO and KEGG pathway analysis was done using the R package missMethyl version 3.8 (151).

2.5.6 Threshold analysis

We explored the use of a 'threshold function' to identify samples that are difficult to classify into discrete ethnicity groupings because of mixed ancestry. Because planet's ethnicity classifications are associated with varying degrees of confidence (i.e., probabilities), we reasoned that a sample's most probable ethnicity classification (i.e., max(P(Asian), P(African), P(Caucasian)) would be lower with a higher degree of mixed ancestry. Therefore, we implemented a threshold function on planet's probability outputs that classifies samples as 'Ambiguous' if the highest of the three class-specific probabilities is below a certain threshold. We explored several thresholds and decided on 0.75, which minimized the resulting decrease in predictive performance (Figure A.3).

2.5.7 Comparison of methods for inferring genetic ancestry / ethnicity from 450k data

Because existing population inference methods and planet use different statistical approaches to infer genetic ancestry/ethnicity (PCA-based vs predictive modeling), we compared each method based on the amount of population-associated signal in DNAm from each method-specific subset of sites. This was done by applying PCA to standardized beta values for HM450k sites associated with each method (Table 1.1) (31–33) within each cohort. To avoid bias, the PCs associated with planet were calculated for each cohort using a classifier

trained on all other cohorts (generated from LODOCV). To test for the amount ethnicity and genetic ancestry –associated signal in the sites corresponding to each method, we applied several simple linear regression models to estimate the amount of variance explained in PCi (i = 1, 2, 3, ..., 10) by self-reported ethnicity and genetic ancestry when available. To determine other factors that might affect signal in these sites, we also tested for the association between PCi and each covariate available for each cohort. All simple regression tests were done in R using the function lm().

To compare planet to Zhou et al. 2017's SNP-based classifier (33), we used the package R package sesame (version 1.1.0) (152) to obtain SNP-based ethnicity classifications for samples with idats available (cohorts C3, C4, and C5).

2.5.8 Application of planet to previous EWAS

To demonstrate application of planet, we downloaded placental 450k DNAm datasets GSE98224, GSE100197, and GSE71678. We note that GSE100197 and GSE98224 overlap cohorts C4 and C5, respectively. To apply planet to obtain ethnicity information, raw data was downloaded from GEO in the form of IDATs and loaded into R using minfi (version 1.26.2). Both NOOB and BMIQ normalization were applied before applying planet. The R package limma (version 3.36.2) was used to test for differentially methylated sites. For GSE98224 and GSE100197, the processed DNAm data was used, and statistical thresholds were chosen the same as the published analysis (99). For enrichment analysis, differentially methylated CpGs were inputted into the gometh function from the R package missMethyl (version 1.16.0) using all filtered sites as background, and default settings.

3 Cell-specific characterization of the placental methylome

3.1 Introduction

The placenta is composed of several distinct cell populations, each carrying out a specific biological role, that together create a healthy placenta. Measurements of DNAm using whole placental chorionic villi, are therefore, an average measurement of the variety of cells that make up a placental sample. Grigoriu 2011 found that major placental cell types, cytotrophoblasts and fibroblasts have highly distinct DNAm profiles (46). Fogarty 2015 compared epigenetic markers such as 5'hydroxymethylcytosine (5'hmc), H3K9me3, H3K27me3, and others in CTB and STB (57). They found that STB had lower levels of H3K27me3 and H3K9me3, and higher H4K20me3 and 5'hmc compared to CTB. Because placental cell types are epigenetically distinct, interpretation of DNAm measurements in whole chorionic villi are difficult to interpret without addressing cell composition variability. Without reference DNAm profiles of constituent cell types, bioinformatic approaches to account for cell composition variability, such as reference-based cell deconvolution, cannot be used (85).

To address this, in this study we generated DNAm reference profiles for 4 major human placental cell populations using the 850k DNAm microarray. Our study is the first to characterize the DNAm of major placental cell populations with a high-resolution approach, across first trimester and term placentas. We show that cell-specific DNAm occurs at thousands of CpG sites, of which a subset can be used to infer cell composition using cellular deconvolution. Our study underscores the importance of cell-specific approaches in placental studies, especially when measuring epigenetic features such as DNAm.

3.2 Results

3.2.1 Major human placental cell types have highly specific methylation patterns

To characterize the dynamics of CpG methylation during human placental development, we performed DNAm microarray profiling (n CpGs=737,050 after removal unreliable probes) in samples of matched CV and 4 fluorescence-activated cell sorted (FACS) cell- types (Figure B.1A), from 9 first trimester (6.4–13 weeks gestational age) and 19 term (36.4–40.4 weeks) pregnancies (Table 3.1). Immunofluorescence staining of flow cytometry sorted cells (Figure B.1B-E) determined high purity for TB (KRT7+, 97%), HB (CD68+, 95%), and EC (CD31+, 88%) and lower purity for SC (VIM+, 73%). Several bioinformatic approaches, such as array-based sex inference (90), and genotype clustering, were used to identify contamination with maternal DNA (Figure B.2A-F). We restricted analysis to samples with an estimated maternal cell contamination of less than 35%, with the majority of first trimester samples having less than 20%, and term samples less than

Table 3.1: Number of cell-specific and matched chorionic villi samples from first trimester and term placentas, measured on the Illumina 850k DNAm microarray. Surface markers for flow cytometry and immunofluorescence staining are shown in brackets.

	First trimester	Term
Chorionic villi	7	19
Trophoblast (EGFR+/KRT7+)	5	19
Hofbauer (CD14+/CD68+)	3	18
Endothelial (CD34+/CD31+)	8	19
Stromal (VIM+)	9	19
Mean Gestation age (mean and range in weeks)	10.8 (6.4–13)	39.0 (36.4–40.4)
Sex (n Males)	4	9

10% (Figure B.2G). This resulted in the exclusion of: 6 HB, 1 EC, and 4 TB from first trimester, and 1 HB from term samples. Final sample numbers in all downstream analyses are shown in Table 3.1.

To determine major factors that drive DNAm variation, we first applied principal components analysis (PCA) to all 126 CV and cell samples. Three distinct clusters were observed when samples were projected onto PCs 1 and 2 (total percent variation explained = 64%; Figure 3.1a). Samples in these clusters were i) TB and CV, ii) SC and EC, and iii) HB. Cell type was strongly associated with the first 3 PCs (p<0.001), while gestational age (i.e. "Trimester") was the second strongest identifiable factor driving DNAm variation, being associated with PCs 4 and 5 (p<0.001, Figure B.3). Technical variables such as "Batch," "Row," and "Chip ID" explained less variation in comparison to biological variables. Sex was associated with PCs 6 and 8–11 (p<0.01). The close clustering of TB with CV (original unsorted tissue) is consistent with this being the predominant cell type in whole villi.

We next wanted to define the extent and patterns of cell-specific DNAm. At a Bonferroni-adjusted p<0.01 and an absolute difference in mean methylation (Δ)>25%, we found 75,000–135,553 and 9136–117,528 (term and first trimester, respectively) cell-specific differentially methylated CpGs (DMCs; Figure 3.1b). The differences in the number of DMCs between first trimester (n = 3–9) and term (n = 18–19) are likely due to less power from the smaller sample size for first trimester samples compared to term. When comparing across term samples, we detected more DMCs for TB and HB (n = 135,553 and 130,733) compared to SC and EC (80,153 and 75,525; respectively). This was also true for first trimester samples: there were more DMCs for TB and HB (117,528 and 78,309) than SC and EC (9136 and 18,867). We further classified these DMCs by whether their methylation was in the "less than" (compared to all other cell types) or "more than" direction. Most TB DMCs were in the less methylated direction (61% - first trimester, 88% term), whereas HB DMCs were often more methylated than other cell types (74% - first, 72% term). A list of 38,656–86,355 differentially methylated regions (DMRs) were identified (FDR<0.01) using the R package

dmrcate for each cell type and gestational age.

To characterize the functional relevance of placental cell-specific DMCs, we tested these CpGs for enrichment in various genomic elements (chi-squared test, FDR<0.05; term DMCs in Figure 3.1c, first trimester DMCs in Figure B.4). Cell-specific DMCs were depleted in gene-related elements such as promoters, exons, 5 UTRs, and 3 UTRs. Instead, we saw significant enrichment in non-coding regions, such as open seas, CpG island shores, intergenic regions, introns, and enhancers. The level and direction of enrichment was highly consistent across first trimester and term cell DMCs. Less methylated DMCs were enriched for placental PMD regions (77) for TB but depleted for all other cell types. Functional enrichment analysis tested if GO or KEGG pathways were associated with cell-specific DMCs. We adjusted for the variable number of CpGs per gene to reduce bias in gene set analysis. EC and HB DMCs were enriched (FDR<0.05) for terms related to intercellular interactions such as "cellular response to external stimulus," whereas stromal DMCs yielded more intracellular processes related to maintaining tissue structure, such as "actin cytoskeleton" and "collagen binding." Trophoblast DMCs were enriched for two KEGG pathways, "ECM-receptor interaction" and "Regulation of actin cytoskeleton" (Table B.1 and B.3).

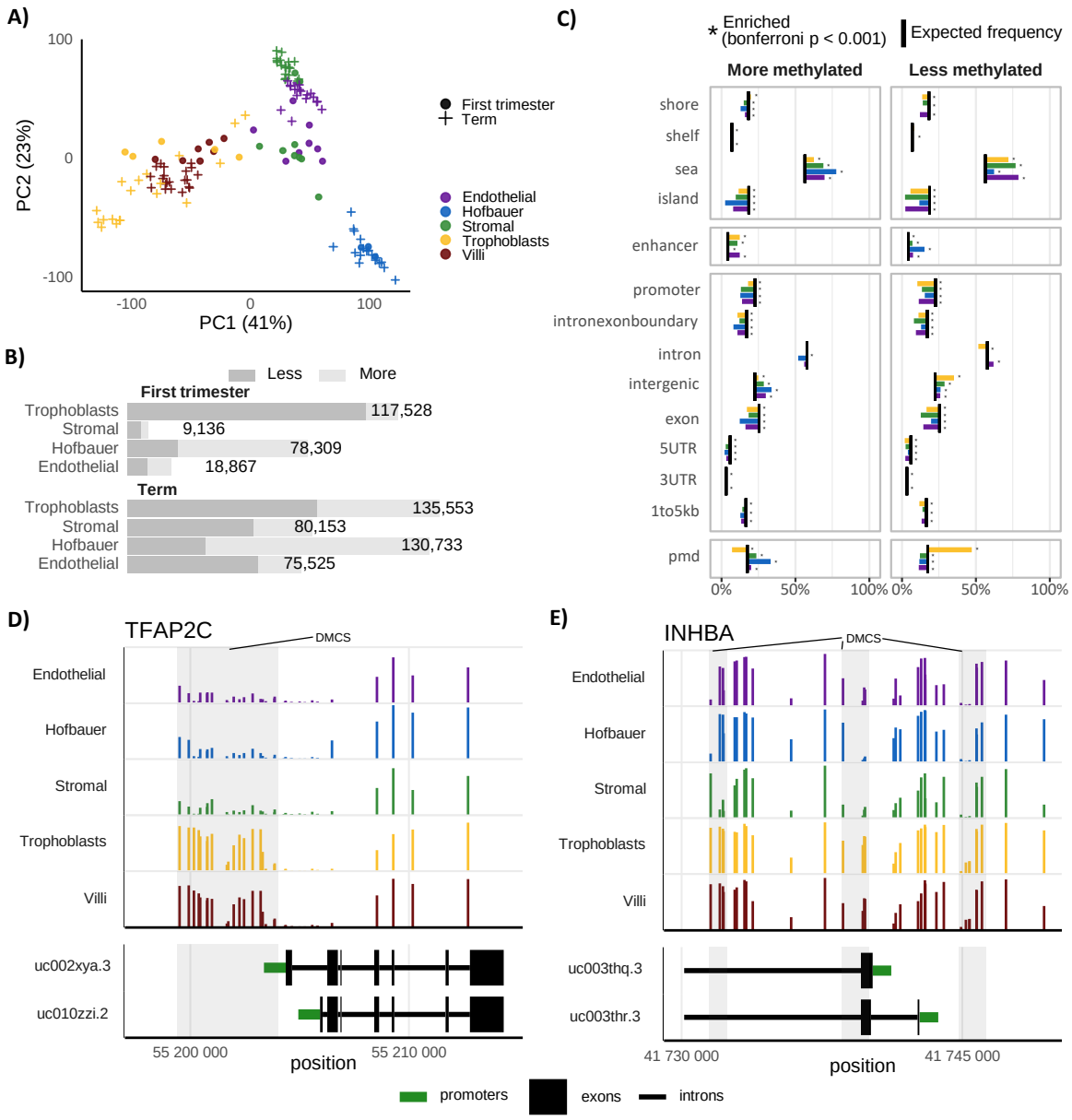

Figure 3.1: Genome-wide characterization of placental cell DNAm **a** Principal components analysis (PCA) was applied to all samples and CpGs. Samples are projected onto axes PC1 and PC2 which account for 41% and 23% total variance, respectively. **b** Results from the differential methylation analysis using the R package *limma* are shown here. DMCs, defined as those tests passing a Bonferroni-adjust p-value <0.01, and a difference in group means >0.25, were divided into less methylated and more methylated compared to all other cell types. **c** Enrichment analysis of term cell-specific DMCs was carried out on genomic elements using a chi-squared test and a Bonferroni-adjusted p-value <0.01. The expected (background) frequency, which is the percentage of total tested CpGs in each genomic element, is shown as a black line. **d** Average term placental cell-specific DNAm across *TFAP2C* transcripts on chromosome 6, and **e** *INHBA* transcripts on chromosome 7. Differentially methylated regions (defined as regions with a high density of differentially methylated CpGs), are highlighted with a grey background. Y axis ranges from 0 to 100% DNAm

Table 3.2: Number of preeclampsia-associated CpGs from Wilson et al. 2018 that are cell-specific DMCs for term samples. Enrichment for preeclampsia-associated CpGs was statistically significant for each term cell-specific set of CpGs at a Bonferroni-adjusted p<0.01

	n cell-specific DMCs	n DMCs that are preeclampsia-associated	Proportion out of 599 preeclampsia CpGs that are also cell-specific DMCs	Odds ratio
Trophoblast	135553	147 (0.11%)	27.20%	1.66
Stromal	80153	105 (0.13%)	19.40%	1.98
Endothelial	75525	109 (0.14%)	20.20%	2.22
Hofbauer Cells	130733	131 (0.10%)	24.30%	1.49

3.2.2 Cell-specific DNAm occurs at highly functionally-relevant genes

A number of regions with a high density of DMCs were located in or nearby functionally- and pathology-relevant genes. *TFAP2C*, which encodes a pan-trophoblast marker, were highly methylated in TB compared to other cell types in the promoter and upstream region; whole CV showed a similar profile to TB (Figure 3.1d). This region contains several predicted enhancers (153), which may require DNAm for recruiting transcription factors. Alternatively, other regions more distal to *TFAP2C* may be responsible for regulation of this gene's transcription. Other trophoblast-specific markers, such as *GCM1*, *MMP2*, *SLC1A5*, and *GATA3*, also had regions of highly cell-specific DNAm localized near their transcription start sites (Figure B.5). We also observed high DMC density regions in genes for which placental DNAm and/or expression differences have been associated with preeclampsia (99), including *INHBA* (Figure 3.1e), *JUNB, TEAD3*, *NDRG1*, and *CGA* (Figure B.6). Out of 540 preeclampsia-associated CpGs previously identified by Wilson et al. 2018 that were also captured in our processed data, a statistically significant (Bonferroni adjusted $p<0.01$) fraction ranging from 19.4–27.2% were also identified as exhibiting cell-specific DNAm for term samples (Table 3.2) (99).

We hypothesized that genome-wide differences in DNAm could in part relate to differences in the expression and DNAm at genes that regulate the deposition, maintenance, and removal of DNAm, such as *DNMT1*, *DNMT3A*, *DNMT3B*, *DNMT3L*, and *TET1*. In these genes, we found that a high proportion of CpGs in the promoter region (61, 36, 31, 83, 18%, respectively) were differentially methylated by cell type. However, considering the variable number of CpGs associated with each gene's promoter, these percentages were not significantly greater than genes of similar CpG coverage (Figure 3.2ab). Differential methylation within DNAm-regulating genes was highly localized (Figure 3.2c). The promoter of DNAm-maintenance gene *DNMT1*, which is known to be specifically imprinted in the placenta (154), shows the expected intermediately methylated (i.e. ~50%) pattern for all cell types except HB, which is completely unmethylated (Figure 3.2c).

This suggests that *DNMT1* is imprinted in TB, SC, and EC, but not in HB.

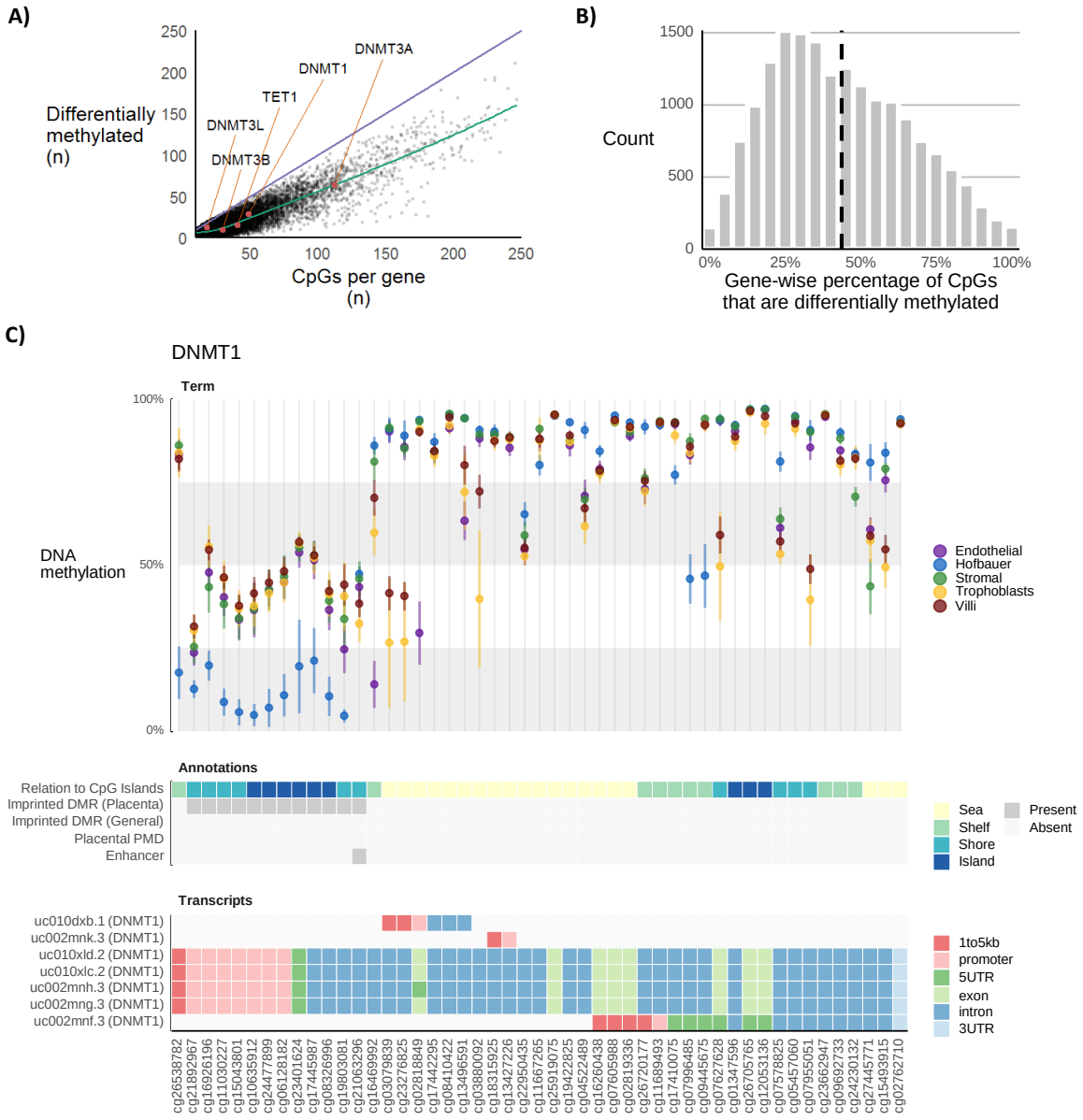

Figure 3.2: Differential methylation at DNAm -regulating genes. **a** On a per-gene basis, the number of promoter CpGs that are differentially methylated by at least one cell type, out of the total number of promoter CpGs per gene. The y=x line is shown (blue), where genes with 100% of promoter CpGs are differentially methylated. The green line is a smoothed average. **b** Distribution of the percentage of promoter CpGs per gene that are differentially methylated. The dotted line represents an array-wide average. **c** DNAm at CpGs associated with *DNMT1* for term placental samples (top). CpGs in CpG islands, regions of genomic imprinting, PMDs, and enhancers are indicated (middle). Associated UCSC transcripts and their genomic elements (promoter, 5 UTR, exons, introns, 3'UTR) are displayed (bottom)

3.2.3 DNA methylation characterization of Syncytiotrophoblast and Hofbauer cells

We used the pan-trophoblast marker EGFR to isolate TB using FACS. Because mature EVTs exist primarily in maternal tissue, and STBs are structurally incompatible with FACS isolation protocols, our TB sample likely consists primarily of CTB. In order to better understand the relationship between STB and the isolated TB cells, we compared a subset of TB with matched STB from the same placenta that was obtained from enzymatic separation using Collagenase IA (referred to as eSTB; n=5) from term CV samples. This digestion protocol which extracts the outer layer of the CV, produces a sample enriched for STB, but is likely to also contain a proportion of non-STB cell types. To compare eSTB samples globally to other cell types, we projected eSTB onto PCs 1 and 2 to see where they cluster in relation to other samples. On PCs 1 and 2, eSTB clustered closely with TB and CV samples, indicating high similarity between these three populations (Figure 3.3a). Throughout gestation, the STB proportion increases, and is greater in nuclei number compared to CTB at term (155). To determine if TB or eSTB samples were more similar to CV, unsupervised hierarchical clustering was applied on the top 1000 most variable probes, and resulted in CV clustering with eSTB (Figure 3.3b), which is consistent with the expectation that CV consists primarily of STB. Supporting this, we found more DMCs (Bonferroni p<0.01, absolute difference in mean DNAm $>$ 25%) between TB and eSTB (n DMCs=4666), than between CV and eSTB (n DMCs=72). Differential methylation at specific CpGs localized to genes known to be expressed in STB, such as *CGA*, *CYP19A1*, *PAPPA2*, *PARP1*, *SLC13A4*, and *SLC22A11* (Figure 3.3c) (156–159). The direction of DNAm at these CpGs was mostly consistent with expected patterns of genes that are more active in eSTB compared to TB and other placental cell types (i.e. more methylation at introns, less methylation at promoters)

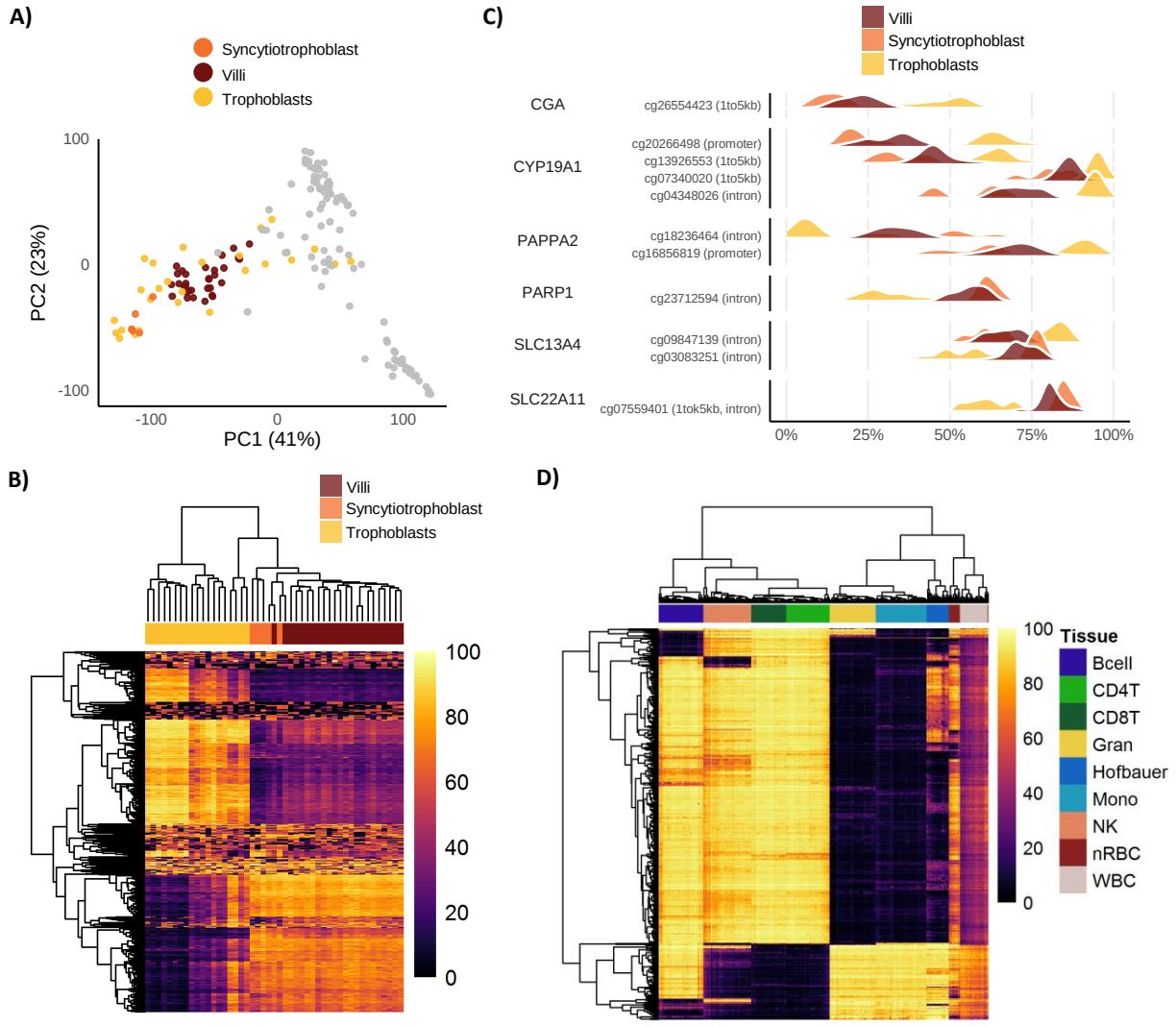

Figure 3.3: Characterization of enzymatically-separated syncytiotrophoblast and Hofbauer cell DNAm to closely related cell types. **a** Syncytiotrophoblast samples (n=5) were projected onto principal components PC1 and PC2. Original samples used for constructing these PCs (Fig. 1a) are shown (chorionic villi: dark red, trophoblast: yellow, all others: grey). Syncytiotrophoblast (orange) cluster with the chorionic villi and trophoblast samples. **b** Clustering on the top 1000 variable CpGs between chorionic villi, syncytiotrophoblast, and trophoblast samples. Hierarchical clustering with Euclidean distance was used for both CpG-wise (rows) and sample-wise (columns) clustering. DNAm is shown as a range between 0 and 100%. **c** Density plots are shown for select differentially methylated CpGs, which were identified using limma, with a Bonferroni adjust p<0.01, and a mean difference in DNAm >25%. CpGs are shown along the y-axis with their locational relationship (shown in brackets) to their associated gene (left). DNAm is shown on the x-axis. **d** Clustering on the top 1000 variable CpGs between Hofbauer cells and cord blood cell types. Hierarchical clustering with Euclidean distance was used for both CpG-wise and sample-wise clustering. WBC: whole cord blood, nRBC: nucleated red blood cells, NK: natural killer cells, CD4T: CD4+ T cells, CD8T: CD8 T cells, Gran: granulocytes, Bcell: B cells, DNAm: DNA methylation.

The distinct DNAm profiles observed in placental HB suggests a distinct developmental trajectory. Indeed, the functional role and phenotypic diversity of HBs is complex and thought to vary across gestation,

however, they show similar morphological and cell marker characteristics as adult and fetal monocytes (71). Therefore, to compare placental HBs to other immune cells, we compared their DNAm profiles to a curated 450k DNAm database of flow-sorted cord blood cell types (n=263) (160). We included only term HBs in this comparison since the available cord blood data was collected from term samples. To determine which cord blood cell types HB are most similar to, we applied unsupervised hierarchical clustering on the top 1000 most variable CpGs across each dataset. We observed that HB form their own distinct cluster (Figure 3.3d), indicating they likely have unique functional properties compared to other immune cells at similar developmental stages. This finding supports previous reports of distinct DNAm between HBs, fetal/maternal monocytes, and decidual macrophages (67). HBs cluster most closely with monocytes and granulocytes, consistent with them having a common developmental origin.

3.2.4 Canonical placental epigenetic features are not always present in all constituent cells

To determine if previously identified placental specific features of DNAm are cell specific, we compared cell-type specific DNAm at partially methylated domains (PMDs), genomic imprinting, and repetitive elements (77,79,161). PMDs are large (>100kb) regions of lower average methylation (<70%) compared to surrounding regions. Placental PMDs are thought to contribute to the observation that placental DNAm on average is much lower than other human tissues (66). To characterize their cell-specificity, we calculated the percentage of CpGs that are found in previously defined placental PMDs (77) with DNAm falling into 20% intervals (0–20%, 20–40%, 40–60%, 60–80%, 80–100%). We observed that DNAm levels in PMDs is highly cell-specific (Figure 3.4a). TB, like CV, have more CpGs with low levels of DNAm in PMDs (0–40%) compared to other cell types. HB show a strong bias towards higher DNAm levels, with over 43% of CpGs in PMDs exhibiting >80% DNAm. We observed some changes within cell types between trimesters. All cell types have lower levels (0–40%) of methylation in term compared to first trimester. All cell types except TB have less intermediately (40–60% intervals) methylated CpGs at term compared to first trimester. HB, in contrast, have more intermediately (40–60% intervals) methylated CpGs in third trimester. In summary, the methylation levels at CpGs in PMD regions were at the expected levels (relatively low methylated compared to surrounding regions) for CV and TB; sometimes hypermethylated for EC and SC; and were almost always highly methylated for HB, at levels typically found in somatic cells.

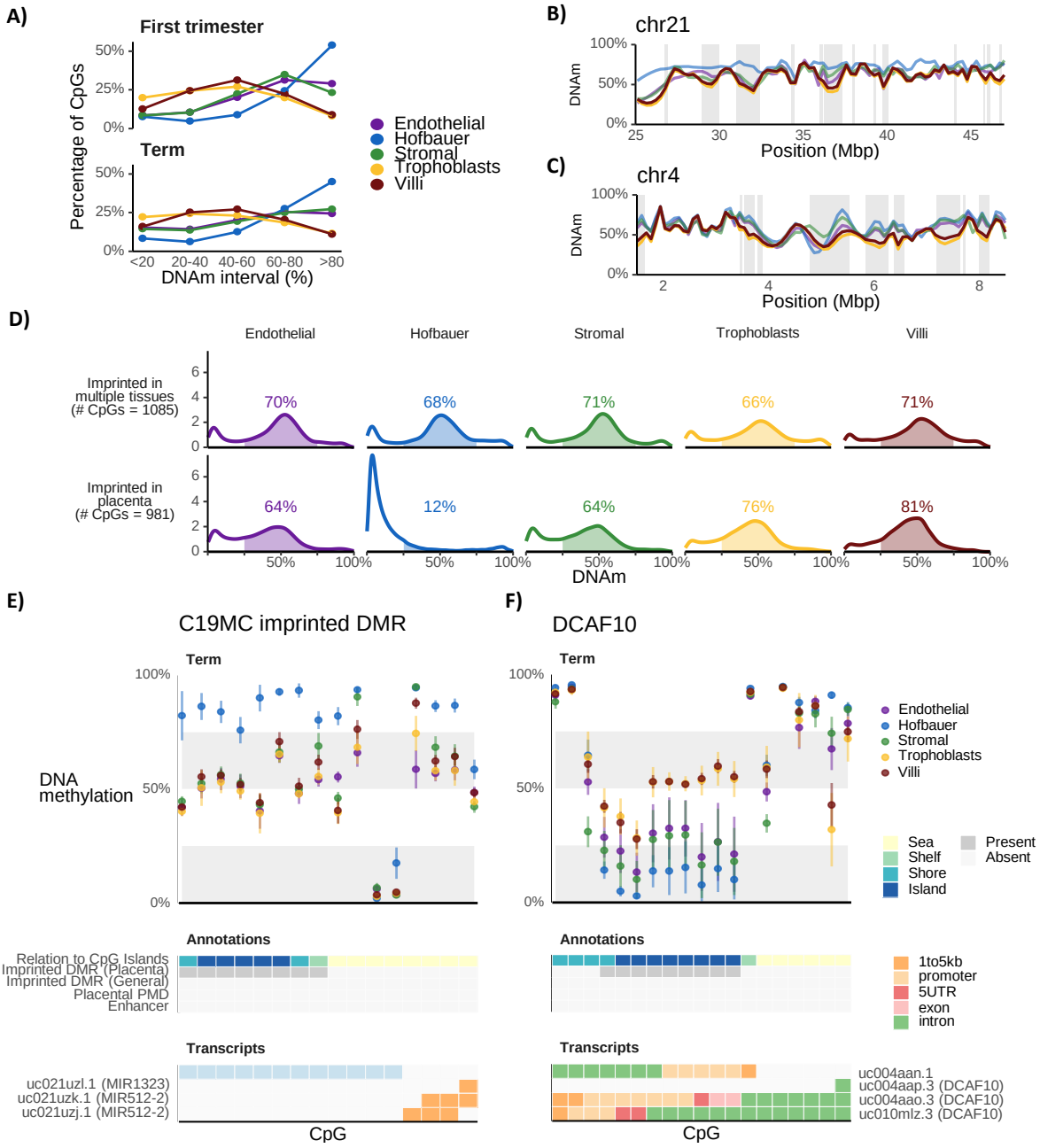

Figure 3.4: DNAm at partially methylated domains (PMDs), and imprinted differentially methylated regions (DMRs). **a** The percentage (y-axis) of CpGs in placental PMDs, falling into specific methylation intervals (0–20%, 20–40%, 40–60%, etc.) is shown for each cell type and trimester. **b** DNAm across specific regions on chromosome 21 (B) and 4 (**c**). PMDs are highlighted with a grey background. **d** Density plots (y-axis) of imprinted DMRs in term samples, divided into those that are imprinted in multiple tissues, (i.e. non-placental-specific; 1085 CpGs total; top) and placental-specific (981 CpGs total; bottom). The percentage of CpGs falling within 25%–75% is shown above each plot. **e** Cell-specific DNAm at the C19MC placental-specific imprinted DMR. This placental-specific imprint overlies a CpG island upstream of the miRNA cluster. **f** DNAm at placental-specific imprinted region for DCAF10

In examining specific regions containing PMDs, a strong bimodal pattern of methylation was ob-

served, where regions of lower methylation (overlapping known placental PMD regions), which were surrounded by regions of higher methylation (Figure 3.4bc). TB DNAm levels followed closely the levels measured in CV, supporting that placental PMDs are likely reflecting mainly TB-specific DNAm patterns. In contrast, DNAm in HB often deviated from the other cell types, typically showing higher levels of methylation within PMDs. SC and EC often "followed" CV DNAm levels, but were not nearly as consistent as TB cells in this respect.

We also looked at imprinted differentially methylated regions (DMRs) that are covered by the 850k array. While many imprinted DMRs are maintained in somatic tissues, others are highly specific to the placenta (79–82). To evaluate whether placental-specific imprinting is maintained in constituent placental cell populations, we first combined the results from four studies (79,80,82,162) to form a list (B.3) of placental-specific (n CpGs = 981; n genes = 111) and non-placental specific (i.e. imprinted in other tissues) DMRs (n CpGs = 1085; n genes = 307). To determine if CpGs were intermediately-methylated, as would be expected for an imprinted DMR, we counted the proportion of CpGs with an average DNAm across both alleles that were in a range between 25 and 75% methylation. For CpGs in non-placental specific imprinted DMRs, the mean percentage of CpGs in the intermediate range across each cell type and in CV in term samples was 69% (Figure 3.4d). For placental-specific imprinted CpGs, the percentage of CpGs falling into this DNAm range was much more variable. As expected, in the term placental samples, TB and CV had a high percentage (76, 81%, respectively) of CpGs in this DNAm range. SC and EC had a lower, but still a majority, percentage of CpGs in this range (64, 64%, respectively). In contrast, HB cells had almost no CpGs (12%) in this range; almost all CpGs were unmethylated (<25%). These proportions were similar in first trimester samples, except with EC and SC showing less intermediate methylation and more CpGs with less methylation at placental-specific imprints (Figure B.7A). These results suggest that placental-specific imprinting is maintained primarily in TB, and to a lesser degree EC and SC, and is virtually absent in HB. When considering the parental origin of imprinted DNAm (79–82,162), paternally- methylated regions had more CpGs falling within 25–75% as compared to maternal ones (Figure B.7BC). We only estimated this in non-placental specific imprinted DMRs, since almost all validated placental-specific imprinted DMRs are maternally methylated.

DNAm at specific imprinted DMRs was examined. As described above, TB and CV had intermediate (>25, <75%) DNAm at nearly all CpGs located in placental-specific imprinted regions (Figure 3.4d). Most of these CpGs, in contrast, are hypomethylated for HB cells, consistent with this cell type having a different developmental origin than other placental components (embryonic versus extraembryonic). However, at the imprinted DMR associated with the placental-specific expressed microRNA cluster C19MC, this pat-

tern is reversed: HB have hypermethylation at this region (Figure 3.4e) as is reported for somatic adult/fetal tissues (79). For SC and EC, these cell types generally show lower levels of DNAm than TB/CV at the placental-imprinted DMRs, sometimes matching that in HBs and other times showing levels somewhere between HB and TB/CV. Such patterns are observed for genes such as *DCAF10* (Figure 3.4f), fibroblast growth factors *FGF8*, *FGF12* (Figure B.8AB), and at epigenetic regulator *JMJD1C* (Figure B.9A). However, for a few DMRs, levels of DNAm in SC/EC matched that of TB/CV, such as ones associated with the DNAm maintenance gene *DNMT1* (Figure 3.2c) and *FGF14* (Figure B.9B). Higher DNAm than TB/CV was only observed for 1 gene (*RASGRF1*, Figure B.10).

DNAm at repetitive elements, such as Alu and LINE1 elements, can be placental-specific and have been hypothesized to often be important regulatory components of placental processes (163). To determine if DNAm at repetitive elements is consistent across placental cell populations, we analyzed the subset of 850k CpGs that map to Alu (n=15,289) and LINE1 (22,006) elements. Compared to CV, TB had lower LINE1 DNAm (mean difference in DNAm=-1.5%, p=0.04), and HB had much higher DNAm (+9.7%, p<0.001; Figure B.11A). Similar relationships are seen for Alu elements. TB had lower (-1.2%, p=0.02), HB had higher (+7.0%, p<0.001), and EC had higher (+2.1%, p<0.001) DNAm in Alu CpGs, when compared to CV. To explore large-scale DNAm differences, we averaged DNAm across all 850k probes and compared each cell type to CV. We found these relationships to be similar to those with the subset of repetitive elements probes. HB had higher DNAm compared to CV (+5%, p<0.001), and all other cell types had lower DNAm (Table B.4). The relationships we found for repetitive elements and global DNAm between cell types and CV were also largely consistent in our first trimester samples (Table B.4, Figure B.11B). To determine genome-wide repetitive element DNAm, we used the random forest -based 'REMP' algorithm (134) to predict 438,664 Alu CpGs and 39,136 LINE1 CpGs that are not covered by the 850k array. Relationships between cell types and CV for predicted and non-predicted repetitive elements were mostly the same, except TB DNAm in predicted Alu and LINE1 CpGs was not significantly different compared to CV (Table B.4, Figure B.11C).

3.2.5 Cell-specific DNAm dynamics across gestation

To determine how DNAm changes in placental cell populations over gestation, we compared first and third trimester cell samples at 737,050 CpGs. We found 108,814 (TB); 94,619 (SC); 63,433 (EC) and 1550 (HB) significant cell-specific gestational-age dependent DMCs (Bonferroni p<0.01, Δ >0.05). Strikingly, almost all of the TB DMCs show an increase in DNAm from first trimester to term (98.2%; Figure 3.5a). Most gestational-age DMCs for HB and SC also show an increase in DNAm from first trimester to term (75.6 and 56.6%, respectively). In contrast, EC DMCs show less DNAm in the term compared to first trimester

(77.1%).

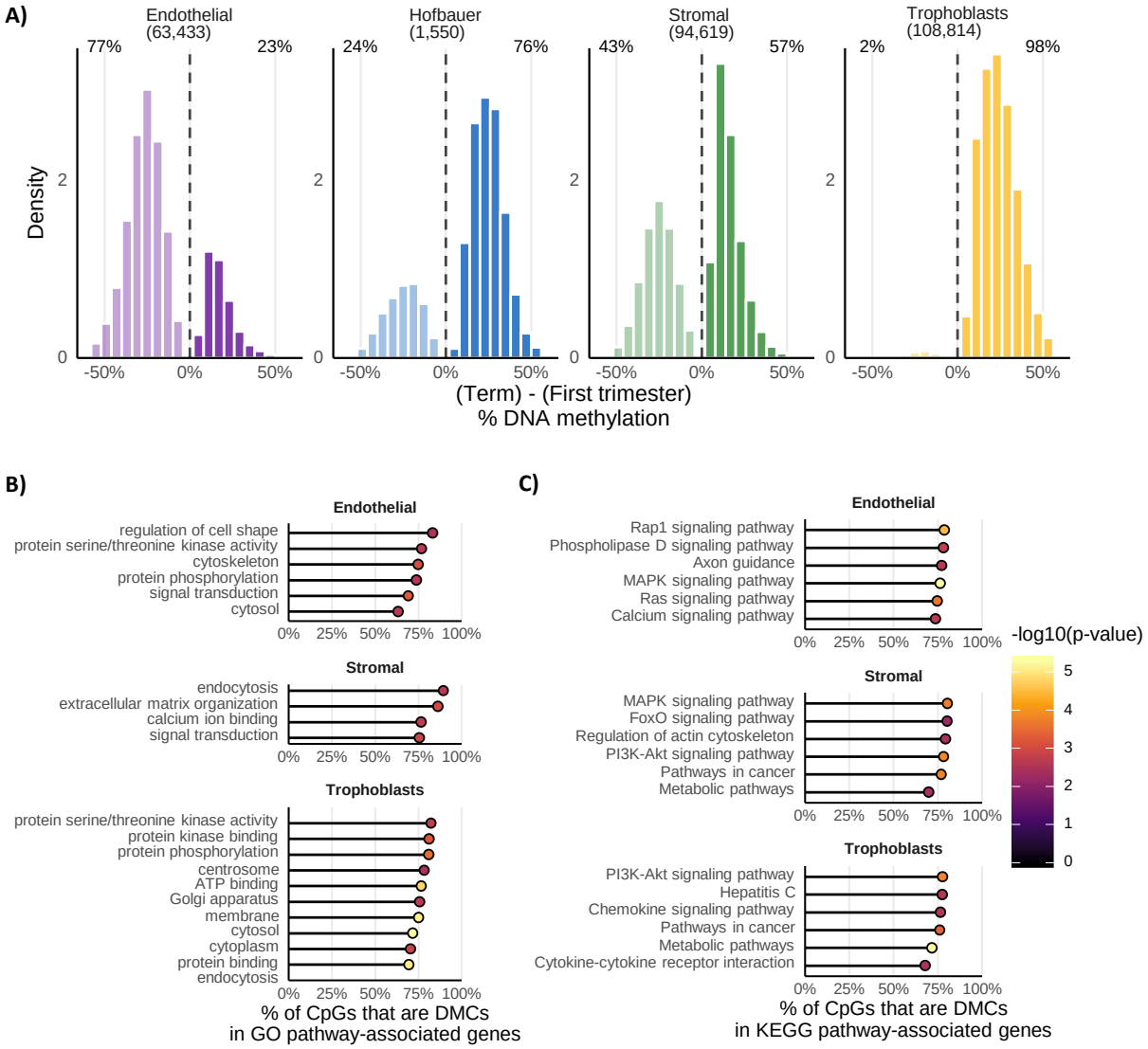

Figure 3.5: Gestational-age dependent DNAm within each placental cell population. **a** The distribution of the changes in DNAm between first and third trimester, within each cell type. Only statistically significant (Bonferroni p<0.01) and biologically relevant (mean change in DNAm >5%) differences are shown. Number of gestational age associated DMCs are labelled above each plot. **b** Functional enrichment analysis for GO terms tested with the R package *missMethyl*. **c** Functional enrichment analysis for KEGG pathways tested with the R package *missMethyl*. D) Enrichment for genomic features: CpG island-related elements, enhancers, PMDs, and gene features

Several interesting KEGG pathways and GO terms were significant (FDR<0.05) in our functional enrichment analysis (Figure 3.5bc). Immune pathways ("Cytokine-Cytokine receptor interactions") and metabolism-related terms ("metabolic pathways," "ATP binding," "kinase activity") for trophoblast

gestational-age dependent DMCs suggest a highly active state throughout gestation affecting multiple placental functions. As expected, stromal terms were highly associated with cellular/tissue structure -related terms, such as "extracellular matrix organization," and "Regulation of actin cytoskeleton." No significant pathway or GO terms were found significant for HB gestational-age DMCs. Most gestational-age dependent DMCs were enriched with open sea regions, regardless of direction of methylation (Figure 3.5d). HB DMCs that increase in methylation with gestational age were the only cell type DMCs that were heavily enriched for enhancers (Bonferroni $p<0.001$). Trophoblast DMCs that increase in methylation with gestational age were enriched for CpG island shores, open seas, and intergenic regions (Bonferroni $p<0.001$). All cell type-specific gestational-age dependent DMCs were depleted (Bonferroni $p<0.001$) for promoter regions, suggesting that genome-wide promoter DNAm is mostly stable from first trimester to term.

3.2.6 Assessing cell composition in chorionic villi

Using placental cell DNAm profiles as a reference, we assessed cellular composition in CV samples using cellular deconvolution. To select cell-type discriminating CpGs, the *pickCompProbes* function from the R package *minfi* (134) was used, which takes the top 100 most hypo- and hyper-methylated CpGs ranked by F-test statistic for each cell type. Gestational-age specific references were created for first trimester and term. For first trimester samples, reference probes were selected from all first trimester cell samples, but also term nucleated red blood cells (nRBCs) and eSTB samples were used since these cell types are also present in early gestation (164). For nRBC samples, 11 DNAm profiles from umbilical cord blood from public databases were included (160). Reference CpGs determined from first trimester (Figure 3.6a) and term (Figure 3.6b) placental samples were highly cell-specific.

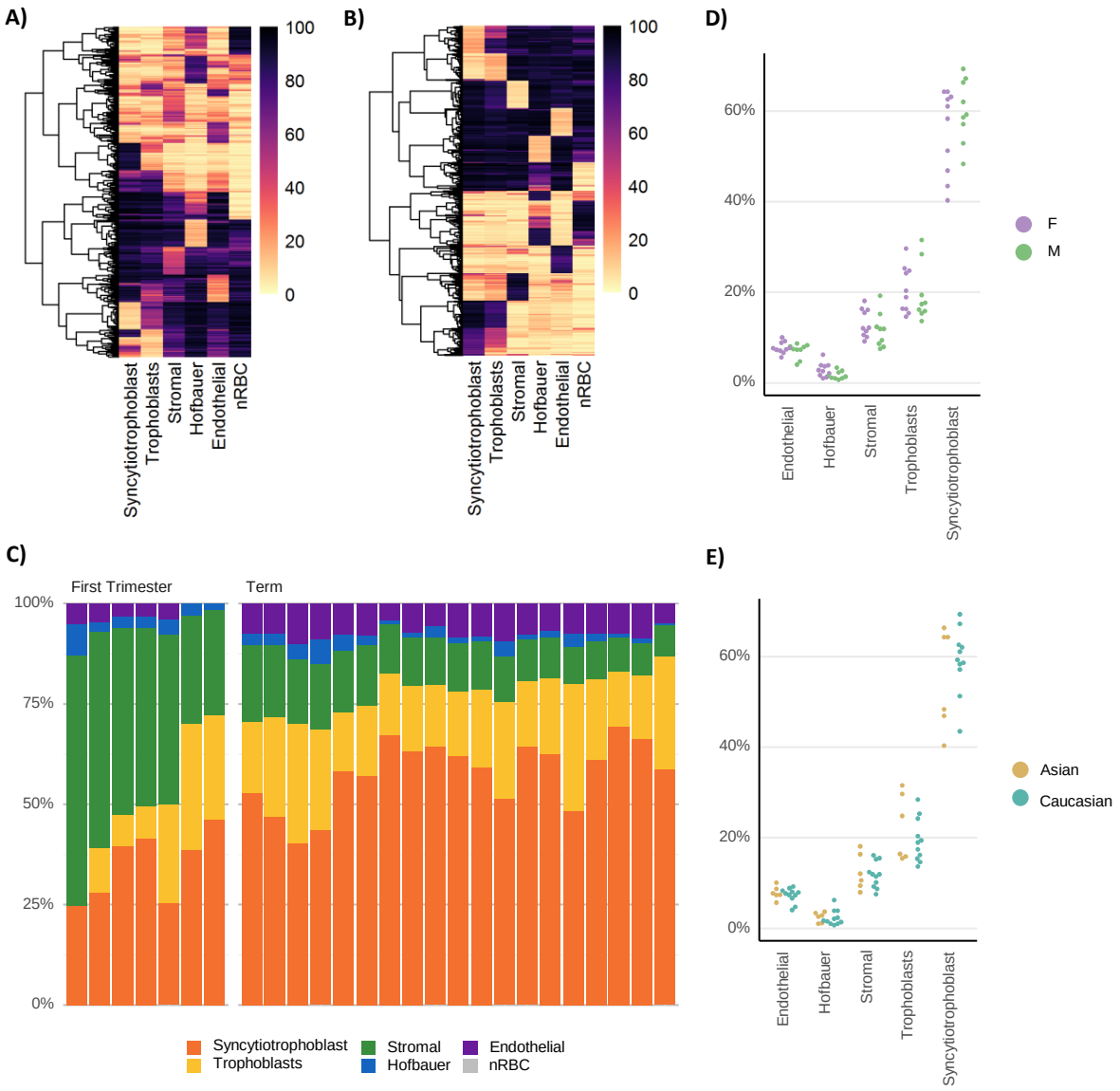

Figure 3.6: Assessing cell composition in first trimester and term CV samples. **a** Mean DNAm across each cell type (columns) for 600 first trimester deconvolution reference CpGs selected by minfi::pickCompProbes. CpGs (rows) are hierarchically clustered using euclidean distance. **b** Term reference CpGs. **c** Cell composition of 7 first trimester and 19 third trimester CV samples, estimated with cellular deconvolution using RPC. **d** Cell composition is similar between male and female term samples with respect to estimated percentage of each cell type (y-axis). F: female, M: male. **e** Cell composition is similar between Asian and European/Caucasian third trimester samples

To determine the best-performing cellular deconvolution method, 1500 in silico bulk mixtures were generated based on our cell data with known cellular composition proportions. These deconvolution methods were compared: constrained projection (CP) (86), robust partial correlations (RPC) (85), and support-vector regression / CIBERSORT (CBS) (87). All three methods were tested using the implementation from the R

package *EpiDISH* (165), and the constrained projection approach was used from implementations in both *EpiDISH* and *minfi* (134) R packages. Performance was high and consistent across algorithms and cell types (R^2=0.88–0.99, RMSE=0.02–0.08, MAE=0.01–0.4; Figure B.12A; Table B.5). However, RPC slightly outperformed other approaches (R^2=0.96, MAE=0.024, RMSE=0.045). Biases towards under−/over- estimation for certain cell types were small but were consistent across algorithms (Figure B.12B): SC tended to be overestimated (mean difference between estimated and actual= +0.33% to 0.98%), HB were underestimated (−0.03% to −0.38%). TB were underestimated (−0.07% to 0.94%), and nRBCs do not show as much bias (+0.03 to −0.21%).

To assess the validity of placental cell deconvolution estimates, we applied deconvolution to previously published placental samples that are enriched for specific cell populations. Deconvolution was applied to cultured trophoblast samples (n=90) from Yuen et al. 2013 (53), that were cultured to 24h (predominantly CTB phenotype) or to 48h, after which many CTB cells have fused into STB; each set of samples was also subjected to varying oxygen levels (1%, 8%, 20%). Cultured STB had higher estimated STB relative to sample-matched cultured CTB (Figure B.13A). The small changes in STB:CTB between culturing times are consistent with the small DNAm differences that were reported in Yuen 2013 (53), and suggest that although fusion of cytotrophoblast was achieved, further culturing would be required to produce a mature STB phenotype akin to term placenta. We then applied deconvolution to first, second trimester, and term enzymatically separated mesenchyme (n=3) and matched samples of outer TB layer of chorionic villi (n=3) from Hanna et al. 2016 (79), the latter of which were isolated in the same manner as eSTB in the present study. Despite batch effects and array differences (450k vs 850k), the term TB sample was estimated to be mostly syncytiotrophoblast (97%; Figure B.13B). Deconvolution estimates for trophoblast isolated from first and second trimester placentas were also mostly TB with some presence of the mesenchymal components, in particular some SC. Matched mesenchyme samples, as expected, were enriched for SC, and EC. Overall, these findings are consistent with our understanding that enzymatic separation enriches for certain populations but cannot produce homogenous cell populations. Lastly, we applied cell deconvolution to chorionic villi samples (n=5) that were enriched for large visible stem villi. These samples had cell compositions that were heavily enriched for SC (mean=51%, sd=4%), compared to matched "normally" -processed chorionic villi (mean=11%, sd=2%; Figure B.13C).

RPC cellular deconvolution was applied to our 7 first trimester and 19 term CV samples. There was significant gestational-age specific variation in the estimated percentage of eSTB, TB, and SC (Table 3; Figure 3.6c). eSTB were the most abundant cell type in all (19/19) term samples (mean=58%), whereas SC was the most abundant in most (5/7) first trimester samples (mean=43%). There were significant changes

Table 3.3: Mean of cell composition estimates (%) for first trimester and term CV samples using RPC cellular deconvolution. Standard deviation is shown in parentheses

	First (n = 7)	Term (n = 19)
Syncytiotrophoblast	35 (9)	58 (8)
Trophoblast	16 (12)	20 (6)
Stromal	43 (13)	12 (3)
Hofbauer cells	3 (2)	2.34 (1)
Endothelial	3 (2)	7 (1)
Nucleated red blood cells	0 (0)	0 (0)

from first trimester to term samples: there was a significant mean increase of 23% in eSTB (Bonferroni-adjusted $p<0.001$), a decrease in SC (−31%; adjusted $p<0.001$) and a small increase in EC (+5%; adjust $p<=0.005$). A detectable contribution of nRBCs was not estimated in any sample using RPC deconvolution. No significant (adjusted $p>0.01$) differences in cell composition were observed between male (n=9) and female (n=10) samples (Figure 3.6d; Table B.6), or between European/Caucasian (n=11) and East Asian (n=6) samples (Figure 3.6e) for term CV. Within-trimester gestational age (estimated and reported) was not significantly associated with cell composition (Table B.6), although numbers were small.

3.3 Discussion

We performed a comprehensive analysis of DNAm for human placental cell types using the Illumina 850k methylation array. Previous placental cell DNAm studies have focused on a lesser number of cell types (166), used lower resolution approaches (46), or focused on a narrow gestational age range (e.g. only first trimester, or only term). Using the 850k array, which targets CpG sites in gene-rich regions and non-coding regulatory elements, this study describes the DNAm profiles of major human placental cell types from first trimester and term placentas, and identifies cell-specific and gestational age –dependent DNAm.

After the wave of *de novo* DNAm in the inner cell mass and trophectoderm, global differences in DNAm exist between these two blastocyst cell layers and their derivatives. These differences result in genome-wide patterns with the placenta showing a unique hypomethylated DNAm profile compared to other somatic tissues (167). Earlier studies suggested that the hypomethylated placenta was partly due to lower DNAm at repetitive elements such as LINE1 (75,76,163). We show that LINE1 and ALU DNAm is higher in HB compared to other placental cell types, but otherwise displays low cell-specificity. Later studies indicated that placental hypomethylation could be largely attributed to long regions of consistently low methylation (PMDs), and that this type of patterning was unique to the placenta (77). We found that PMDs are more pronounced in TB, and are absent from HB. The impact of PMDs is unclear and may in part reflect that in

the blastocyst, the trophectoderm does not undergo *de novo* DNAm. Whether PMDs serve a functional role in the placenta is also unclear, but our understanding of their relevance would benefit from characterizing the timing of their development. We note that although our study is genome-wide, the number of CpG loci analyzed (n=737,050) is only a fraction of the epigenome (~3%), and is biased towards genomic regions with annotated functionality (e.g. near genes and regulatory elements). Therefore, findings of this study should be interpreted with these limitations in mind. To comprehensively understand repetitive element and PMD DNAm, higher resolution approaches such as whole genome bisulfite sequencing will be necessary.

Not all CpG sites undergo dynamic changes in DNAm status throughout development. Genomic imprinting, defined as parent-of-origin specific gene expression, is typically associated with regulatory regions (promoters/enhancers) that exhibit parent-of-origin dependent DNAm. Imprinting is an evolutionary phenomenon that exists only in eutherian mammals (78), which suggests a potential important relationship between placental function and imprinting. Consistent with this, there is an enrichment for imprinted genes that are specific to the placenta (79–82). Although our study lacks parental information, previously identified placental-specific imprinted DMRs tend to show the expected intermediate DNAm levels in TB, and to a lesser degree, in EC and SC. For example, the placental-imprinted gene, *DNMT1*, is only unmethylated at its promoter in HB, while other placental cell types are hemi-methylated. DNAm-mediated down-regulation of *DNMT1* expression has been shown in whole placental tissue (48), and our data suggest that that DNAm-mediated regulation of DNMT1 varies by placental cell population. All placental-specific imprints examined in this study showed the expected intermediate methylation for TB and CV samples, and hypomethylation for HB. The methylation patterns of HBs are consistent with an origin from fetal monocytes. The variability in the patterns of DNAm at these imprinted CpGs for mesenchymal components, EC and SC, could be from variability in the timing of erasure of these imprints. Future investigations in resolving parental-origin-specific DNAm and expression are needed. In contrast, our data suggest that common (non-placental specific) imprinted CpGs are maintained in all placental cell populations.

To address the challenges of cell composition variability in placental DNAm studies, we have generated DNAm profiles for 4 major human placental cell populations as well as enzymatically isolated STB, and assessed their utility as references for cellular deconvolution. Like other tissues (86,160,168), placental cell composition can be estimated with any of the commonly used deconvolution approaches. However, it was not possible to independently validate the DNAm-based cell composition estimates presented in this study with other quantitative measures of cell composition (e.g. with histology) and it is not possible to get measures on the identical sample assayed for DNAm. Instead, we validated bioinformatically estimated cell composition in cultured trophoblasts and in previously published samples that are enriched for certain

populations (e.g. by enzymatically stripping away the outer layers of chorionic villi). Estimated cell composition in first trimester and term samples was also consistent with our understanding of how placental cell composition changes across gestation. The ratio of CTB:STB is relatively equal at 13 weeks gestation (155). But as trophoblastic surface area increases as pregnancy progresses (169), at term, 90% of nuclei exist in STB and the remainder are in CTB (155). This corresponds to a large observed increase in the eSTB component (+58% from first trimester), becoming the predominant cell population at term. However, we note that within the STB, nuclei are also heterogeneous in their chromatin state, where there are 4 times more transcriptionally inactive nuclei compared to active ones (57). This property, combined with how similar CTB and STB methylation profiles are, may limit the ability to accurately estimate eSTB proportion. There were also no nRBCs estimated as present in either first trimester or term placentas, suggesting that their contribution to placental cell composition may be very small, at least in uncomplicated pregnancies. Together, these observations suggest that this approach is able to capture large relative changes, but may be imprecise when assessing smaller changes. Future studies with independent measures of cell composition, such as from histology, will be essential for assessing the accuracy of this approach, as has been done for cell deconvolution in other tissues such as adult/cord blood and brain (86,160,168).

There was also significant interindividual variation in cell composition that could not be fully explained by within-trimester gestational age variation, suggesting that other factors contribute to cell composition variability. In this study, we found that chronological (i.e. reported) and biological (i.e. estimated from DNAm) gestational age, sex, and ancestry were not significantly associated with cell composition. But the sample size supporting these findings was small and future studies with more appropriate power are needed to answer how much these factors play in contributing to placental cell composition variability. We also caution that the accuracy of cell composition estimates on first trimester samples relies on the degree of gestational-age dependent variation in term eSTB and nRBC reference CpGs, which could not be assessed in this study.

Another challenge to this study, and others which use a single or few marker genes/proteins to isolate/define cell populations, is addressing heterogeneity within relatively homogenous cell populations. As mentioned previously, TB contains several subtypes, such as CTB, STB and EVTs. In this study, our TB is likely mostly CTB but contains some proportion of immature precursors to the other TB subtypes, given that pan-trophoblast markers EGFR+ were used for cell isolation. HB (CD14+/CD68+) and SC (VIM+) can also able to be divided into meaningful subtypes (71). It will be essential to placental epigenetics research to develop DNAm references for other placental cell subtypes, such as extravillous trophoblast. This will be especially important in studies on placental pathologies (and likely also in many other phenotypes), where

certain TB subtypes are more affected than others, such as preeclampsia (170) and placenta accreta (171). Associated changes in cell composition with preeclampsia may explain the finding that many CpGs with altered DNAm in preeclampsia (99) are also highly cell-specific. Cellular heterogeneity will always be a challenge when using techniques that take measurements in samples that consist of a mixed population of cells. However, cell deconvolution applied in placental DNAm studies will significantly improve interpretation of the resulting measurements and findings.

3.4 Methods

3.4.1 Patient recruitment

Placental tissues were obtained with approval from the University of British Columbia / Children's and Women's Health Centre of British Columbia Research Ethics Board (H04–70488, H16–02280, H13–00640). Women for a scheduled C-section with a healthy term (>37 weeks) singleton pregnancy were recruited with written informed consent at BC Women's Hospital, Vancouver Canada. In addition, first trimester samples from elective terminations were obtained in a deidentified manner. A total of 9 first trimester (6.4–13 weeks) and 19 term (36.4–40.4 weeks) placental samples were obtained; all were screened for large chromosome abnormalities using CNV calling on the 850k array, and found to be normal. No gross pathologies were noted.

3.4.2 Tissue processing and cell isolation

Fresh term placental samples from 3 to 4 sites were taken from the fetal-facing side of the placental disc to avoid maternal contamination and pooled for processing. Chorionic villi samples were washed several times in 1X PBS to eliminate all traces of visible blood and physically homogenized using razor blades. For term placental samples the tissue was then incubated twice in a denuding/digestion HEPES buffer containing HBSS, Dispase, trypsin and DNase I for 30 min at 37C°, to allow the separation of most of the syncytiotrophoblast layer of the chorionic villi. The remaining tissue was then washed in HBSS media with 2% FBS (HF media) and subsequently digested using Collagenase/Hyaluronidase Digestion DMEM Buffer with DNase I, at 37oC for 1h with vortexing every 30 min. The supernatant was collected. This is followed by a wash of the remaining cell pellet with HF media, gentle centrifugation at 4°C for 10 min and further digestion of the cell pellet, with gentle mixing with a pipette, using of 0.25% trypsin solution for 2 min at room temp. The pellet is then washed again with HF media and digested once again with a Dispase/DNase I solution by gentle mixing with a pipette. This is followed by a final wash in HF media and filtering of the sample using first 100um and then 40um sieves to eliminate any remaining chunks of tissue. The cells are

then counted and frozen in freezing media at −80°C until used for FACS. The freezing process eliminates a great deal of the remaining non-nucleated red blood cells.

For first trimester samples, the entire placental sample was processed after identification and removal of most of the decidual tissue. The sample was mechanically homogenized using razor blades and then digested with Collagenase/Hyaluronidase DMEM Buffer at 37°C for 1h. The tissue was then washed with HF media and further digested with a 0.25% trypsin solution by gentle mixing with a pipette for 2 min. The pellet is then washed again with HF media and digested once again with a Dispase/DNase I solution by gentle mixing with a pipette for 2 min. The sample was finally washed with HF media and filtered through a 40um sieve. The cell pellet was resuspended in HF and cells were counted and subsequently frozen in freezing media at -80oC until used for FACS.

To isolate human placental cell types with fluorescence-activated cell-sorting (FACS), cells were first thawed and then washed using HF media. Suspended cells in HF media were then filtered through a 40-um sieve (VWR, CA21008–949) and then counted using a hemocytometer. Trypan Blue (0.4%, Amresco, K940-100ML) was used to identify live / dead cells. A final cell solution was made at a concentration of 10 million cells per ml, which was then stained with the following antibodies purchased from eBioscience: 7-AAD (1:25, 00–6993-50), CD235a FITC (1:50, 11–9886-42), CD45 APC-eFluor780 (1:100, 47–0459-42), CD14 PE (1:50, 12–0149-42), CD34 APC (1:25, 17–0349-42), and EGFR PeCy7 (Biolegend, 1:50, 352,909). Approximately 200,000 cells for term placental samples and 125,000 cells for first trimester were obtained for each cell type using the BCCHR FACS Core equipment. DNA was extracted from cell-sorted samples and matched whole villi using Qiagen DNeasy Blood & Tissue kit (Qiagen, 69,504 / 69,506).

Enzymatically isolated syncytiotrophoblasts (eSTB) were obtained from term villi samples using an enzymatic digestion protocol. Briefly, approximately 0.5 cc of chorionic villi were washed thoroughly several times with 1X PBS to eliminate all visible traces of blood without disrupting the tissue and then incubated for 10 min in 1ml of Collagenase IA 1mg/ml (Sigma). The tube was then vortexed for 30s, if cloudy, 3 ml of Hanks Balanced salt solution (HBSS) was added to the digest letting it settle for 2 min. The supernatant containing mostly syncytiotrophoblast (STB) and some cytotrophoblast (CTB) was collected in a separate tube, the pellet was centrifugated and washed in 1X PBS before DNA extraction. This HBSS step is repeated once and all supernatant is pooled in the same tube. If the initial collagenase IA digest is not cloudy after the initial 10 min digestion, the whole villi were digested for an additional 2 min before adding the HBSS.

3.4.3 Measuring DNA methylation in placental samples

DNA quality was checked using a NanoDrop ND-1000 (Thermo Scientific) as well as by electrophoresis on a 1% agarose gel. Bisulfite conversion was carried out using the EZ DNA Methylation Kit (Zymo, D5001 and D5002), before amplification and hybridization to the Infinium Methylation EPIC BeadChip (Illumina, WG-317) following the manufacturer's protocol. An Illumina iScan reader was used to scan the chips and produce raw data files (IDATs).

3.4.4 DNA methylation data processing

To assess various quality metrics, IDAT files were loaded directly into R (v3.6.1) using the *minfi* package (v1.32.0) and *ewastools* (90) (v1.6). Poor quality and unreliable probes (detection p-value > 0.01, bead count <3, cross-hybridizing (33), probes with SNPs within 5 bp of the CpG site in the probe direction (33)), and probes located on sex chromosomes were removed (n=109,410). Analysis was restricted to a final set of 737,050 autosome probes. All samples had high (7500–15,000) average median intensity readings in the methylated and unmethylated channels, and passed manufacturer-determined default thresholds for 17 control probes. The possibility of sample mislabelling was verified comparing reported sex and inferred sex based on X chromosome copy number (*ewastools*) (90). Identical genotypes between matched cell-sorted and whole chorionic villi samples were verified using the 850k array's 59 SNP probes (*ewastools*) (90). This genotype-check also identified a number of first trimester cell-sorted samples with evidence of maternal contamination, which were removed from further analyses (n=12). Upon inspection of global DNAm patterns with PCA, we identified and removed 2 outlier samples that we suspect were contaminated with cells from other genotype-matched samples. After quality control and probe filtering, noob (132) and BMIQ (133) normalization was applied to the DNAm data.

3.4.5 Differentially methylated CpGs (DMCs) analysis

All analyses were conducted in R version >3.6.1. To identify differentially methylated CpGs (DMCs), the R package *limma* (172) (v3.42.0) was used to apply CpG-wise linear models with empirical Bayes posterior variance estimators (173). Unless otherwise stated, the "one-versus-all" approach was applied, where for each CpG, the mean DNAm of one cell type was compared to the mean of all other samples (excluding villi). DMCs were defined as those tests that were statistically significant at a bonferroni-adjusted p-value of <0.01, and also a showed a difference in mean DNAm >25%. For functional enrichment analysis of identified DMCs, the R package *missmethyl* (v1.20.0) was used to account for the variable number of CpGs that can be associated with each gene (151). For testing enrichment of DMCs for various genomic features

(e.g. CpG islands, promoters, enhancers) and preeclampsia-associated CpGs, chi-squared tests were applied using the base-R function *chi.sq.test*. Annotations for UCSC transcripts (e.g. promoters, introns, exons, etc.), enhancers, and CpG islands were taken from the R package *annotatr*, which downloads annotation data from UCSC directly. Significant enrichment/depletion was defined as those with a Bonferroni-adjusted $p<0.01$. DMRs were identified using the R package *dmrcate* (v2.0.7), using an FDR cutoff of <0.01, with default settings.

3.4.6 Partially methylated domains

To assess cell-specific placental DNAm in partially methylated domains (PMDs), coordinates for previously identified placental PMDs were taken from Schroeder et al. 2011 (77). Original hg18 coordinates were mapped to hg19 using the UCSC LiftOver tool implemented in the R package *liftover* (v1.10.0). Due to differences in genomic content between the two genome versions, remapping broke up many PMD regions into smaller ones. Fifteen of these smaller "pieces" mapped to different chromosomes, so were removed from further analysis. To account for bias in array-specific coverage towards CpGs lying in promoters, CpG islands, and CpG island shores, these CpGs were removed, as previously described (77).

3.4.7 Imprinted regions

Location data for imprinted regions was created by combining results from five previous human imprinting studies (79,80,82,162). A variety of approaches and technologies were used in these studies, such as whole genome bisulfite sequencing, methyl-sequencing, and Infinium 450k methylation arrays. We took outer coordinates for overlapping regions. A final list of imprinted regions can be found in Table B.3.

Repetitive element mappings were determined by downloading the "rmsk" track from UCSC genome browser (hg19) and then overlapping these regions with the 850k array CpGs. Mean DNAm was calculated by averaging over each set (Alu, LINE1, all CpGs) of CpGs for each sample. Predicted genome-wide DNAm for Alu and LINE1 CpGs was done using the Bioconductor R package *REMP* (174) (v1.8.2), using default settings.

3.4.8 Public cord blood DNAm data

A curated database of cord blood cell types DNAm data ran on the 450k methylation array was used. This data was downloaded from the R package *FlowSorted.CordBloodCombined.450k* and noob normalized (160). For associated analyses, the common probes from this dataset and our 850k data were used. Heatmaps/clustering was applied using the R package *pheatmap*. Where possible, colour-blind friendly

palettes were used with the R package *viridis*.

3.4.9 Cellular deconvolution

Reference probes for cellular deconvolution were determined using the *pickCompProbes* function from the R package *minfi* (v1.30.0) separately for first trimester and third trimester samples. The Houseman et al. 2012 constrained projection (CP) approach was applied using implementations in the *minfi* and *EpiDISH* (v2.0.2) R packages. Other algorithms tested were robust partial correlations (RPC) and CIBERSORT (CBS), both implemented in the *EpiDISH* package. Default parameters were used for all functions, except "constraint" was set to "equality" for using the CP approach from *EpiDISH*. In silico mixtures were generated by the following procedure: 250 proportion samples were drawn from a uniform distribution between 0 and 1. These are the first 250 proportions for one cell type. Two hundred and fifty additional proportions were drawn from a uniform distribution between 0 and the first proportion for the next cell type. This was repeated for a total of 6 times for 6 cell types. These 5 sets of 250 sampled proportions make up 250 in silico mixtures. Because this procedure only ensures that the first set of percentages are uniformly distributed from [0,1] and the remainder are biased towards increasingly smaller values, we repeated this entire procedure for each cell type, each time starting with a different cell type, for a total of 1500 in silico mixtures. Performance metrics to compare algorithms were computed using the r package *yardstick* (v0.0.4). Linear modelling with Bonferroni- multiple testing adjustment was done to test differences in cell composition by sex and by ethnicity. Inferred ethnicity was computed via the R package *planet* (v0.2.0) (45), and corroborated with the first 2 principal components of high density (~2.3 million SNPs) genotyping data

4 Maternal contamination in placental DNA methylation studies

4.1 Introduction

In this study, we provide a comprehensive assessment of maternal cell contamination in placental DNAm studies. Building on some of the existing bioinformatic approaches described above, we designed a multi-faceted placental-specific approach to assess maternal cell contamination in DNAm data.

4.2 Methods

DNAm data was obtained from the data repository GEO for 11 placental studies, 3 of which were from our own lab (Table 4.1). These included studies of acute chorioamnionitis (dataset 1), preeclampsia (datasets 2, 3, 10, 11), neural tube defects (dataset 4), assisted reproductive technologies (dataset 5), neurobehavior of newborns (dataset 7), arsenic exposure (dataset 8), and extremely early gestation preterm birth (dataset 9). Decidua and maternal blood samples from additional datasets GSE113600 (dataset 12) and GSE74738 (dataset 13) were included for methylome-wide principal components analysis (PCA). All studies reported collecting 1-4 sites of chorionic villi from the fetal facing side of the placenta, which is done to avoid contaminating maternal tissue. DNAm data was normalized with BMIQ and noob; and filtered for poor quality, cross-hybridizing, and probes with nearby SNPs according to the 2021 June 15th update of the original Zhou 2017 DNAm annotation (33).

4.2.1 Identifying mixing of genetically distinct DNA samples

To quantify the amount of mixing of genetically distinct DNA in each sample, a mixture model was fit to the 59 SNPs on the 450k/EPIC array using the R package ewastools (90). Mixing of genotypes is expected to result in SNP measurements with a high probability for the "outlier" distribution (referred to as the "P(outlier)" value), which corresponds to measurements that deviate from the 3 possible genotypes. P(outlier) values range from 0 to 1, and increases with mixing of distinct genotypes (e.g. fetal and maternal DNA).

4.2.2 Cell composition outliers

We hypothesized that placental samples significantly contaminated with maternal decidua or blood could be identified from DNAm-based estimates of placental cell composition. Specifically cell composition inferred from our previously developed placental cell deconvolution approach (175), and PCA clustering with maternal blood and decidua samples, were tested for their association with genotype-based contamination. This latter

approach is a variation of the PCA approach described previously from Wan 2019 (95), whereby, instead of using all CpG sites, we used a specific subset of highly decidua and blood-specific CpG sites that we identified through linear modelling. Decidua and blood-specific CpG sites were determined using the R package limma (172), by comparing the DNAm profile from decudua and maternal blood independently to that for placenta; the top 500 hypo- and hyper- methylated CpGs (total was 2000 cpgs for two tissue-specific comparisons) were used for PCA.

4.2.3 Epigenetic age

Epigenetic age is a measure of biological aging based on aging-associated DNAm changes. Epigenetic "age clocks" are usually built using supervised machine learning trained to predict chornological age using CpG methylation (176). Because epigenetic age should be highly increased in decidua and maternal blood (i.e. adult tissues) compared to conceptus-derived placental tissue, we hypothesized that estimated epigenetic age may significantly differ between maternally contaminated and non-contaminated placental samples. To estimate epigenetic age, placental and adult epigenetic age clocks were applied to chorionic villi samples (105,177).

4.2.4 Sex chromosome analysis

Placental samples from male fetuses that are contaminated with maternal (female) DNA, should appear more "female-like" when examining sex chromosome DNAm. Using the R package ewastools (90), DNAm intensity measurements on X and Y chromosome probes were normalized to autosomal signal, and then outliers on X and Y chromosome plots were identified in a dataset-specific manner (termed "XY outliers").

4.3 Results

4.3.1 Genotype mixing

The estimated genotype mixing between studies was highly variable (Figure 4.1A). While some datasets showed no evidence of mixed genotype contributions, other datasets showed a number of placental samples with a high P(outlier) value (Table 4.1). Due to the potential for maternal tissue contamination during placental sampling, we hypothesized that much of the observed genotype mixing can be explained by the presence of maternal cells, either due to sample processing technique (e.g. failure to avoid maternal infarcts or to wash sample well), or biological processes such as inflammation. We noted the number of samples with P(outlier) values greater than 0.15 for each dataset in Table 4.1. As high P(outlier) values may also result from poor data quality or non-maternal sample contamination (e.g. unintentional mixing of placental

Table 4.1: Placental datasets and number of samples with evidence of contamination.

Placental datasets			-	Contamination measures (n samples)			
			n villi / n total	Genotype mixing	PCA outliers	Epigenetic age	XY outliers
1	GSE115508	Konwar 2018 (7)	48	1	15	2	1
2	GSE100197	Wilson 2018 (10)	97			2	
3	GSE98224	Leavey 2018 (11)	48		2		
4	GSE69502	Price 2016 (12)	51	6	32	6	
5	GSE120250	Choufani 2019 (13)	88	1	38		1
6	GSE98938	Zhang 2021 (14)	2		1		4
7	GSE75248	Paquette 2016 (15)	335	5	3	2	3
8	GSE71678	Green 2016 (16)	343	6	18		1
9	GSE167885	ELGAN (17)	411	28	34		6
10	GSE125605	Wang 2019 (18)	42	2	3		
11	GSE75196	Yeung 2016 (19)	24				

samples), it is important to follow up flagged samples with additional measures to determine the cause of these results.

4.3.2 Association with low quality data

We first looked for a relationship between data quality measures and P(outlier values). Relatively few samples failed 1 or more Illumina data quality control checks (Figure 4.1B), likely because most studies eliminate poor quality samples after performing baseline data processing and quality control. However, in 3 out of 11 datasets, samples that failed control checks showed statistically significantly higher P(outlier) values (Figure C.1A), and in dataset 5, the number of samples that failed at least one control check was a large proportion of the dataset, 32/88 (36%) (Table 4.2). However each of 27 Illumina data quality control checks measures a different aspect of DNAm signal quality, and even if samples fail some Illumina controls, other measures of DNAm data quality may be normal (90). A list of the specific controls that failed for each dataset is described in Table C.1. Overall, however, poor quality data, as measured by the number of samples that failed these quality control checks, could not fully account for samples with high P(outlier) values, leaving the possibility that genotype mixing could be occurring in the remainder of the samples.

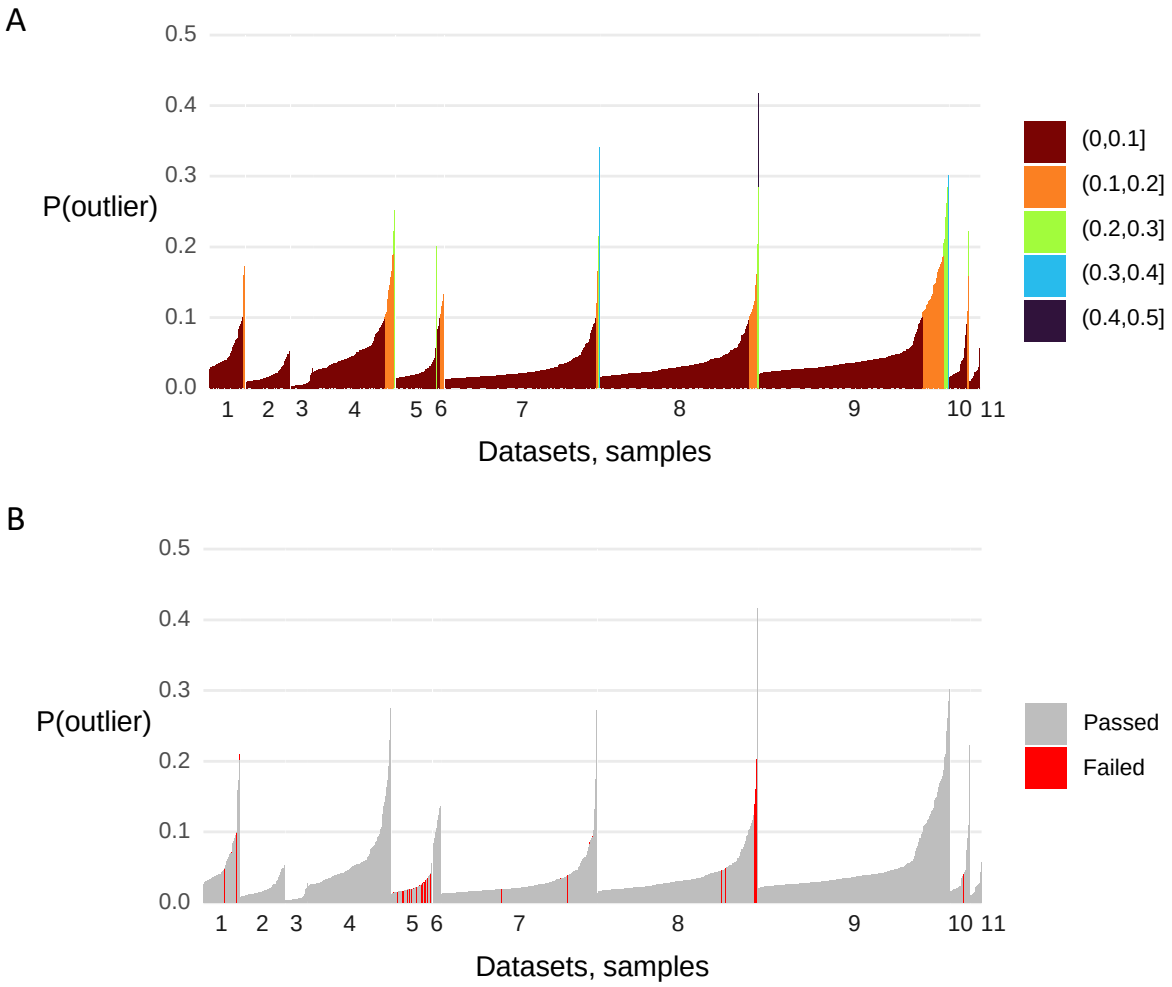

Figure 4.1: Estimated DNA contamination in placental DNAm data. Placental DNAm microarray (450k/850k Illumina) from 13 datasets was downloaded. Genotype contamination was estimated using a mixture model approach from ewastools R package. A) P(outlier), a measure of genotype mixing, estimated across all 13 placental datasets. B) Samples that failed 1 or more Illumina quality control checks tended to have higher P(outlier) values.

4.3.3 PCA assessment of sample contamination

To estimate the proportion of placental samples with P(outlier)>0.15 that were likely contaminated with maternal cells, we performed several other analyses based on DNAm variation. Wan 2019 used a PCA based approach that identified a set of placental samples that clustered near maternal blood and decidua, reflecting a globally altered methylation profile that was increasingly similar to the putative contaminating tissues (95). To increase the power of this approach, we modified the PCA approach by using only highly tissue-specific (maternal blood vs placenta, decidua vs placenta) CpG sites. PCA using the top 500 hypo- and hyper-methylated CpGs between placenta and either maternal blood or decidua (2x tissues = 2000 CpGs total)

Table 4.2: Number of samples that failed 1 or more of 27 Illumina data quality control checks.

Dataset	n	n failed	Some controls that failed
1	78	4	Bisulfite Conversion I Green, Bisulfite Conversion II
2	97	0	
3	48	0	
4	179	0	
5	88	32	Non-polymorphic Green
6	17	0	
7	335	8	Non-polymorphic Green
8	343	8	Non-polymorphic Red
9	411	0	
10	42	1	Bisulfite Conversion I Red, Bisulfite Conversion II
11	24	0	

resulted in the first principal component PC1 (94% variance explained) being able to separate placental and maternal samples (Figure 4.2A). This tissue-specific approach was more sensitive than using PCA on all CpGs (Figure C.2A). Placental samples were labeled based on how many standard deviations (SD) away from the placenta mean on PC1, and towards decidua and maternal blood samples (Figure 4.2A, Table 4.1). PC1 was positively associated ($p<0.001$, reference category: "x < mean") with high P(outlier) values (Figure 4.2B), indicating that more maternal tissue-like DNAm profiles in placental samples had correspondingly higher genotype mixing. Treating PC1 as a continuous variable, a linear model between P(outlier) versus PC1 was statistically significant ($p<0.001$, slope = 3.26, Figure 4.2C). Interestingly, we observed a concave shape to the nonlinear fit using the "LOESS" regression option in ggplot2). This "increasing and then decreasing" relationship of P(outlier) is consistent with a model of genotype mixing, where after a sample becomes more than 50% contamination, the contaminant becomes the primary component of the sample and P(outlier) will start to decrease. Two extreme outliers, both from dataset 11, drive the decreasing side of this relationship 4.2C) and displayed a cell composition profile consistent with high blood contamination (increased nucleated red blood cells and Hofbauer cells, Figure 4.2D)), as estimated from DNAm-based placental cell deconvolution (84,86,175).

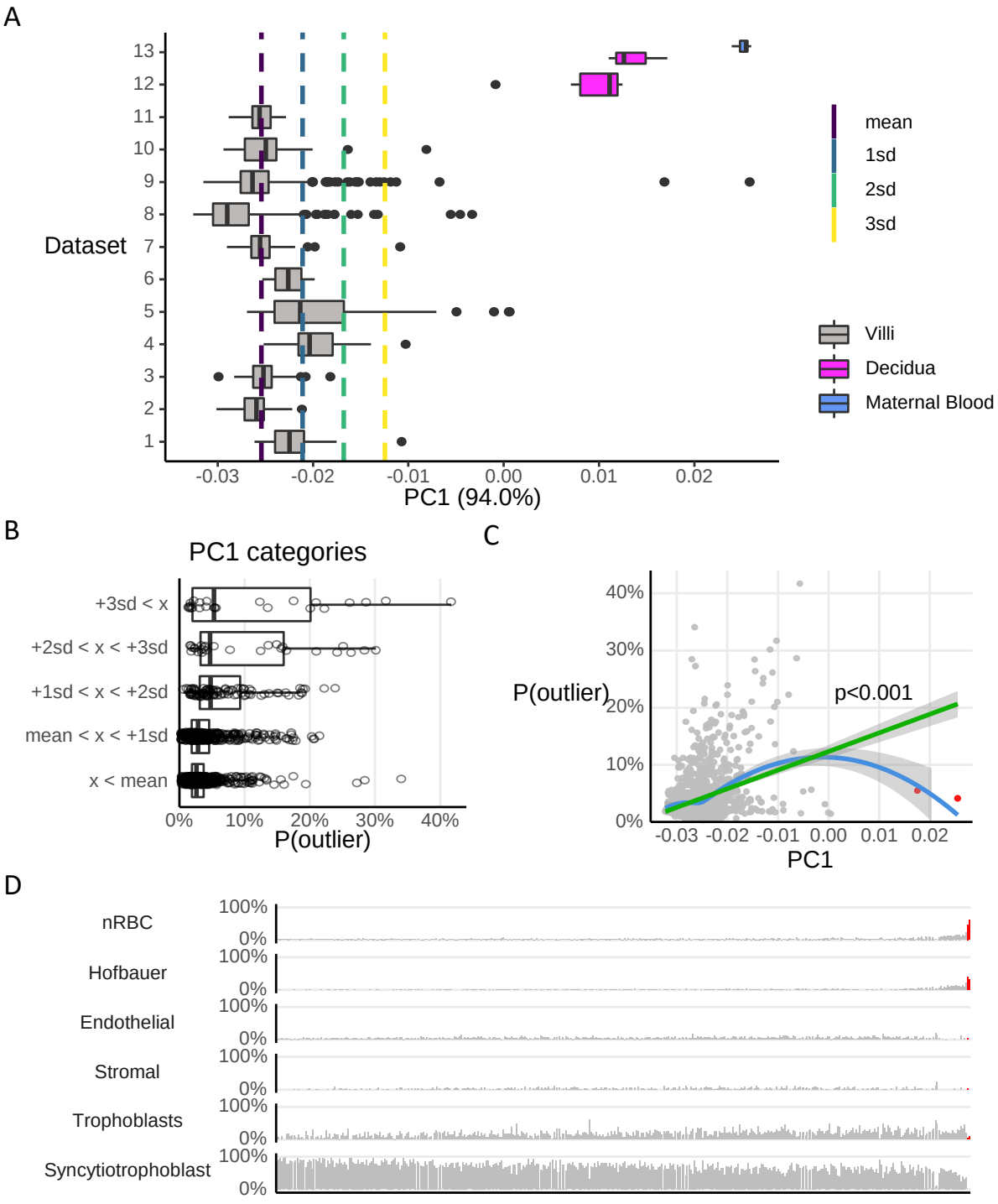

Figure 4.2: Maternal contamination alters DNAm profiles of placental samples. A) Principal component analysis (PCA) using 2000 tissue-specific CpG sites resulted in PC1 being able to discriminate between placental and maternal tissue samples. B) Categorization of placental samples based on PC1. Samples with higher PC1 values were associated with higher P(outlier) values. C) A positive statistically significant relationship between P(outlier) and PC1 was observed ($p<0.001$) using linear regression (green). A non-linear fit is shown in blue. Samples with high PC1 but low P(outlier) are shown in red. D) Cell composition of dataset 11, estimated from DNAm-based placental cell deconvolution. The two outlier samples (red) with high PC1 values are shown having a high immune cell profile. nRBC: nucleated red blood cells, sd: standard deviation, MBD: maternal blood decidua dataset.

4.3.4 Evaluating epigenetic age

We hypothesized that because of the large difference in epigenetic age between maternal (adult) and placental (fetal) tissue, maternally contaminated placental samples should display increased epigenetic age estimates. To test this we applied both adult and placental epigenetic age clocks on all placental samples (105,177). The adult epigenetic age (EA) clock (which was built on somatic tissues) was not effective in accurately separating GA in placental samples (Figure C.3A). However, there was a small, but statistically significant association between adult EA and P(outlier) in placental samples ($p < 0.01$, slope = 0.0125 P(outlier) / year; Figure 34.3A)), an effect that was largely driven by two datasets, 1 and 4 (Figure C.3B). Placental epigenetic gestational age (EGA), which was highly accurate on placental samples (mean difference = +1.07 weeks, sd = 0.88), yielded estimates of EGA that was not as high as expected for decidua (mean = 43.0 weeks) and maternal blood (mean = 44.2 weeks), indicating that EGA is likely to have low sensitivity for maternal contamination (Figure C.3C). The difference between EGA and reported was not significantly associated with P(outlier) (Figure 4.3B, slope = -0.002, n.s.).

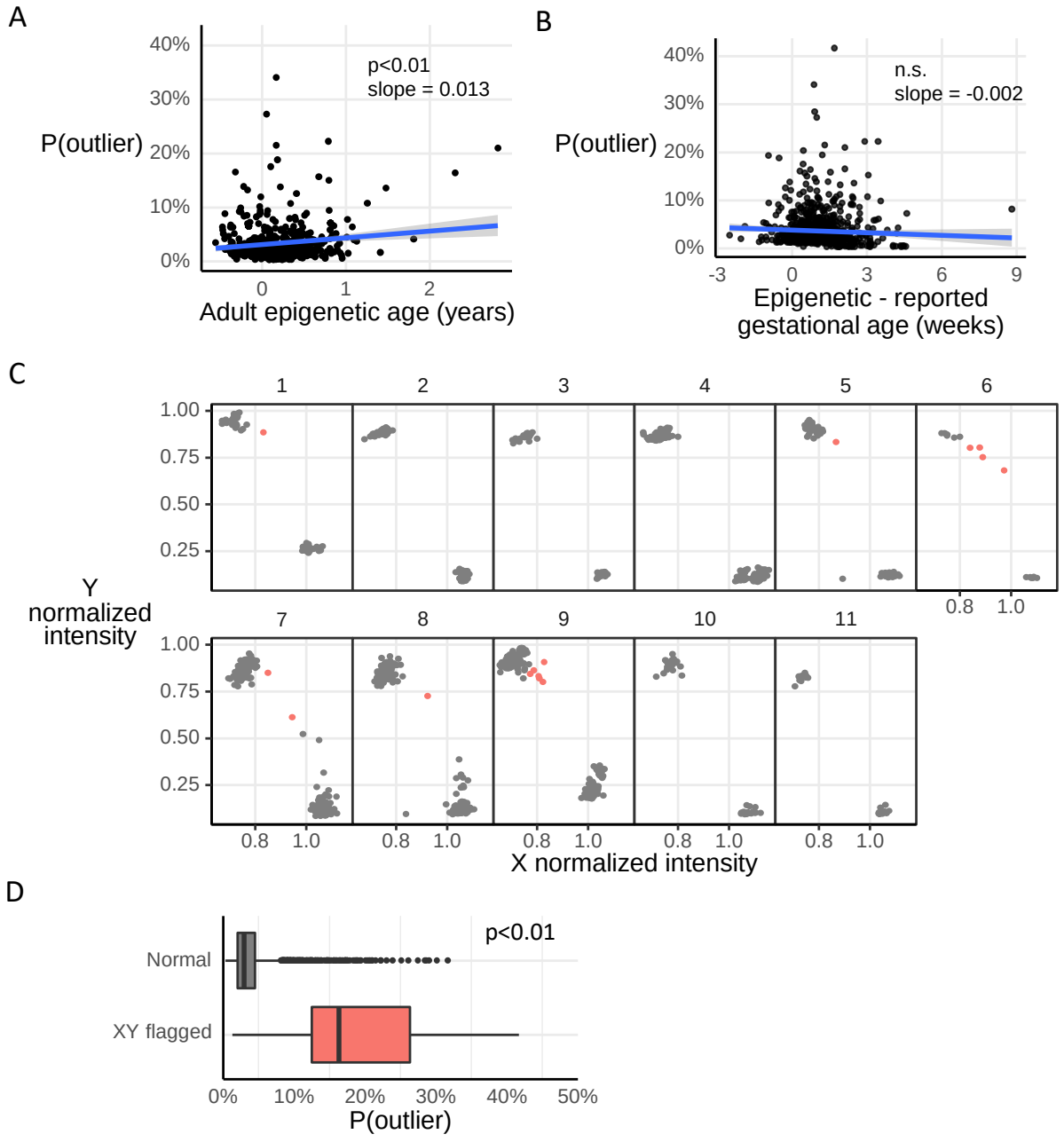

Figure 4.3: Epigenetic age and XY signal intensity analysis. A) Adult epigenetic age on placental chorionic villi samples was associated with P(outlier) ($p<0.01$, slope = 0.013). B) The difference between placental epigenetic gestational age and reported gestational age was not significantly associated with P(outlier). C) Total XY signal intensity normalized to total autosomal signal intensity. Male samples (top left) deviations towards female samples (bottom right) are consistent with maternal contamination. D) Males flagged in XY intensity analysis have significantly ($p<0.01$) higher P(outlier) values than other placental villi samples.

4.3.5 XY signal intensity analysis

A common approach to check for sample misannotation in DNAm analysis is to compare samples' recorded sex and overall signal intensity from the X and Y chromosome probes. Additionally, samples that deviate from the male and female sample clusters can be indicative of contamination with DNA from a different sex, although this can in rare cases also occur for biological reasons (i.e. sex chromosome mosaicism). We hypothesized that maternal contamination in the placenta from male pregnancies would be reflected using XY signal intensity analysis. Based on XY signal intensity normalized to autosomal signal intensity using ewastools (90), 6 datasets had male samples that appear more female-like (trending towards the female cluster) (Figure 4.3C). Although the number of samples flagged by this analysis was small (17/754 males total), flagged samples had significantly increased P(outlier) ($p < 0.01$, mean = 0.24 versus 0.04, Figure 4.3D).

4.3.6 No association of maternal contamination with pathology

Maternal immune cells infiltration into placental tissues can occur in response to placental inflammatory signals, as can occur in chorioamnionitis (178). Moreover, a very small proportion of the normal placental composition consists of maternal macrophages (93). Thus, maternal genotype mixing might be from infiltrating maternal cells as part of a biological process, not an unwanted technical artifact. To determine if putative contamination may be due to biological processes, we tested for the association between our contamination variables and available clinical and placental phenotypes. We were limited to sample variables made available on GEO, but in almost all cases we were able to test for the main phenotypes for each study (preeclampsia, acute chorioamnionitis, neural tube defects, assisted reproductive technologies). We found no statistically significant associations between placental phenotype and any of our measures of genetic and maternal contamination. P(outlier) was only significantly associated with late-onset preeclampsia samples in dataset 2 ($p < 0.05$, slope = -0.0084), but not in any of the other 3 preeclampsia datasets. PC1, generated from our tissue-specific approach described above, was associated with early-onset preeclampsia ($p<0.001$, slope = -0.0016), which does not necessarily indicate maternal cell presence given the large methylation alterations associated with severe preeclampsia (99). Overall, we were unable to robustly detect the presence of maternal cells in placental samples due to biological processes, which suggests that these approaches are not sensitive enough to detect the small number of maternal cells that infiltrate due to inflammation.

4.4 Discussion

We assessed whether the presence of maternal DNA could be detected in Illumina DNAm data derived from placental samples. Although DNAm microarray data was the subject of this study, our findings are relevant to all placental studies where technical of biological contamination from maternal tissue is a concern. While detecting contamination in genomic studies is a challenge, it can, to an extent, be addressed through bioinformatic data analysis after data has been generated. Our analysis builds upon the decade of bioinformatics development of DNAm-specific array data. We expect that accounting for contamination will improve and become more routine as bioinformatic approaches continue to develop, especially with the generation of new genomic technologies. Using this multi-pronged approach, we show that maternal contamination can likely influences a proportion of the 11 placental datasets included in this study. The dataset -specific variation observed in these placental datasets likely reflects variability in sample processing protocols between different research groups. As all data is from previously published studies, our findings represent a conservative estimate of the prevalence of maternal contamination in placental sampling in general, since highly contaminated samples may have been removed through standard QC and data analysis pipelines because of the resulting severely altered DNAm.

The variety of approaches used in this study have their own individual usefulness in identifying different aspects of maternal contamination in placental DNAm data. Although high values of P(outlier) (>0.15) were associated with altered clustering, XY signal intensity, and epigenetic (adult) age, we found that some datasets had high P(outlier) samples with no other evidence of maternal contamination. Because P(outlier) is essentially a measurement of noise and variability of signal intensity, we suspect that dataset-specific variability in background noise will be important to account for when employing this approach, and that dataset-specific considerations are likely important to fine-tune the thresholds of these types of analyses.

The combination of the genetic contamination measure with PCA-based clustering of chorionic villi samples with more decidua / blood-like DNAm profiles, is strongly indicative that maternal contamination is present. But in the absence of genotype contamination, deviations on the PC1 variable can be influenced by other biological factors, such as cell composition variability, disease, environment, and genetics. Selection of CpG sites that are highly tissue-specific can amplify the maternal signal and minimize these other sources of variation, as demonstrated by the increased discrimination we showed in our study compared to when using PCA with all array CpGs. However, this set of CpG sites can be improved since only a low number (n<13) of maternal decidua and maternal blood samples were used to for CpG selection, making the analysis underpowered.

Other limitations for the other approaches are also important to consider when identifying contamination. P(outlier) as noted previously, is also function of signal noise, and is not specific to maternal contamination but can be also useful in identifying sample-sample contamination which we did not explore in this study. Epigenetic age, we found to have little sensitivity to maternal contamination. We found that the placental age clock was not able to robustly discriminate maternal decidua and blood samples from villi, to the degree needed to use this as a measure of maternal contamination. Adult epigenetic age (105) was slightly associated with genetic contamination, indicating it has some use in identifying contamination. XY signal intensity analysis, which Heiss 2018 also explored to assess sample contamination in DNAm data, we found to have a strong association with genetic contamination. However, the restriction of XY analysis for identifying maternal contamination is only possible in male placental samples, and therefore we limit this approach for use as a secondary validation analysis after applying other contamination checks. Overall, our study highlights the importance of using a variety of bioinformatic approaches to identify and magnify different aspects maternal contamination from placental DNAm microarray data.

The bioinformatic approaches described in this study to assess different aspects of contamination, although are specific to the Illumina DNAm arrays and to placental tissue, these approaches can be adopted to other genomic data types and to other tissues. For example, in this study we used CpG sites that were differentially methylated between villi and decidua or maternal blood; however, in cord blood studies, where maternal contamination is also a challenge, a cord-blood specific set of CpG sites could be used instead. We found that P(outlier) variable based on the 59-65 SNP probes on the 450k/850k methylation arrays to be a fruitful measure of genetic contamination, but this approach could be generalizable to other sets of SNPs such as high-density genotyping arrays, of which many EWAS datasets have available too. Although, Morin 2017 found that genetic contamination measured from genotyping arrays was not correlated with other DNAm-based measures of contamination (97). Morin 2017 used the amount of genotype "no-calls," which are when a genotype cannot be confidently discerned from the raw signal intensity array data, as their measure of contamination; this approach is more stringent than the P(outlier) measurements used in this study. Further studies evaluating SNP data their usefulness in estimating contamination are needed.

5 Discussion

This chapter is original and unpublished

5.1 Summary and significance of findings

In this dissertation, I investigated some of the major sources of variation of placental DNAm: genetic ancestry and ethnicity, cell type, and the presence of maternal cells in placental tissues. Regarding these, I discovered the following insights. Ethnicity and genetic ancestry can be accurately predicted from placental DNAm microarray data, using both DNAm markers and SNPs. Many placental-specific DNAm features, such as PMDs, placental-specific imprinting, and repetitive element DNAm, can be found in trophoblast populations, and are often not present in the other constituent placental cell types. Lastly, I found that maternal contamination can occur in placental samples and can be bioinformatically identified from previous placental DNAm studies.

The ethnicity predictor trained on 450k placental DNAm data builds on previous studies and adds a placental-specific approach to the toolbox of methods for accounting for population stratification and genetic ancestry in EWAS (31–33). Although my placental ethnicity predictor, planet, was trained on self-reported ethnicity, I found that the modelled DNAm patterns are highly genotype-dependent, and that my ethnicity predictor is highly correlated with genetic ancestry. In the placental data for which I compared planet to other approaches of inferring genetic ancestry and ethnicity, I found that my tissue-specific approach performed better. This affirms this study's original hypothesis, which was that because DNAm is highly tissue-specific, and that genotype-DNAm relationships can also be highly tissue-specific (41), methods to infer biological variables such as ethnicity and genetic ancestry from DNAm are likely also to be tissue-specific.

Self-reported ethnicity has several shortcomings that my DNAm-based ethnicity predictor addresses. Self-reported ethnicity is often reported for the parents for placental studies, and is often missing from the paternal information. Self-reported ethnicity is also variably defined between countries, institutions, and individuals. In research of biological data such as DNAm, self-reported ethnicity is included as a often inadequate (36) surrogate of genetic ancestry, which should be accounted for in any population-based study such as the common EWAS design. Ethnicity or genetic ancestry, defined instead using DNAm and/or genotyping data is more reproducible and measurable, and therefore comparable across studies. It is unclear if DNAm-based genetic ancestry and ethnicity is more appropriate to use in EWAS compared to using genetic ancestry based on genotyping information, but in the absence of genotyping data, I demonstrate that DNAm-based genetic ancestry and ethnicity is a highly useful alternative to self-reported ethnicity.

Epigenetic modifications such as DNAm have highly cell-specific patterns to regulate cell-specific gene expression. To investigate the cell-specificity of placental DNAm, I characterized the methylomes of 5 major placental cell types in first trimester and term placentas. This work builds on previous research that investigated epigenetics in placental cell types, which was done using either lower resolution profiling technologies (46), and / or other epigenetic marks (57,179). I found highly placental cell-specific DNAm using the 850k Illumina DNAm microarray. Cell-specific DNAm often occured at promoters and regulatory elements (i.e. enhancers, CpG Islands) near known placental-specific genes, some of which had DNAm that previously had not been characterized before. Interestingly, I found that for many placental-specific DNAm features, such as PMDs, placental-specific imprinting, and repetitive elements, I found were present in trophoblasts, variably or intermediately present in endothelial and stromal cells, and mostly absent in Hofbauer cells. Conservation of placental features in trophoblasts is not unsurprising, given that trophoblast make up the bulk of placenta throughout pregnancy. However, trophoblast is itself a heterogenous population, with STB becoming the more predominant subtype in later gestation (155).

Lastly, I assessed cell composition in placental chorionic villi samples using the cell-specific methylomes as a reference for cell deconvolution. I found that cell composition of chorionic villi between first trimester and term placentas to be highly consistent with previous literature reports from using histology to measure cell composition. First trimester chorionic villi had little STB, and was instead mostly CTB, but at term, the bulk of chorionic villi becomes STB. This is consistent with our understanding of STB dynamics throughout gestation. STB undergoes little apoptosis but instead accrues inactive nuclei throughout gestation, while concurrently growing from constant fusion of the underlying CTB layer (155,179). Dieckmann 2021 also found similar cell composition patterns, estimated with our reference data, in chorionic villi samples from 3 cohorts, which included samples collected at term (n = 470, n = 139, n = 137) and from first trimester (n = 264) (180). Overall, Dieckmann 2021 found cell composition patterns that were largely consistent with ours, although there was significant variability between their cohorts, which may be in part attributed to the differing methods of placental sampling, as well as variation in factors such as environmental exposures and genetics (180).

It will be interesting to investigate whether cell composition changes associated with placental phenotype can be detected using placental DNAm cell deconvolution, since placental-mediated health conditions are often correlated with cell-specific and structural changes in the placenta. For example, fetal growth is associated with gross placental morphological changes, such as placental diameter and thickness, and placental weight relative to birth weight (181). However, it is unclear if cell composition also increases proportionally or if the relative proportion of specific cell populations becomes altered. Successes in applying

cell deconvolution to other tissues, such as blood, promises that the future will contain exciting insights for placental DNAm research. DNAm-based cell deconvolution has been demonstrated to be a predictor of disease, treatment response, and cancer survival (184). Moreover, regular improvements have been made to reference datasets and bioinformatic applications, allowing for selection of better quality reference CpGs, and accounting for other sources of variation, such as age (i.e. fetal versus adult) (160,185–188). Therefore, it is expected that similar improvements to placental reference data, and increased usage of DNAm-based cell deconvolution in varying types of placental data will occur, and in turn lead to a better understanding of the cellular variation associated with placental phenotypes.

In addition to variation in placental cell composition, maternal cells can be found present in placental samples. Although effort is often made to avoid maternal tissue contamination during placental tissue processing, it can be difficult especially in earlier gestation samples, where placentas are much smaller. In chapter 4, I assessed whether maternal cells can be detected in placental samples, using DNAm data measured from chorionic villi from 13 different placental datasets on GEO. I developed an approach to bioinformatically identify maternal cell contamination from placental DNAm data and found that maternal contamination is frequently found in placental samples from previous studies. The approaches I developed rely on existing ideas and tools from Heiss 2018 and Wan 2019(90,95), to create a highly placenta-specific approach that I show to be reliably able to identify maternal contamination from placental DNAm. Applying tissue-specific knowledge to develop bioinformatic tools to infer contamination from DNAm has been previously done for other tissues such as cord blood (189). This study demonstrates the importance of considering maternal cells in placental tissues in placental DNAm studies, which can arise either from sample processing, or due to biological processes such as inflammation.

5.2 Strengths and limitations

Throughout this dissertation, I supported accessibility and open data access practices through several concrete actions. I implemented user friendly R packages and functions to enable the easy usage of the bioinformatic approaches I developed. This R package, planet, now includes the placental ethnicity predictor and cell composition reference for deconvolution. Additionally, I developed the "Placental cell methylome browser," which is a web application that provides a user-friendly interface to access and explore the DNAm data. Lastly, throughout my dissertation, I made each of my studies' DNAm data available on GEO, for reproducibility and data sharing purposes. These accessibility practices, I am extremely proud of, and I hope will bring more value to the research I have conducted for my dissertation.

Public open-access data practices were essential to my research. Specifically, I relied heavily on

placental DNAm microarray data deposited on GEO for the bulk of my research data for chapter 2 and 4. In chapter 2, I used 5 placental DNAm datasets that were collected across various sites around North America. The variety in human populations that my datasets comprised of, was especially useful in chapter 2, since population stratification was a main focus. In chapter 4, I used 13 different GEO datasets comprising of various placental tissues to assess maternal contamination from placental DNAm data. Because maternal contamination is generally low and can be highly dataset-specific - a result of the variable sample processing protocols that different research groups use, having several different datasets increased my power to detect maternal contamination at level that was more appropriate for my analysis. Having access to a variety of placental datasets allowed me to develop the bioinformatics with the necessary sample sizes needed for sufficient power.

Without public data repositories like GEO, and an open-access data culture that promotes sharing of genomic data, my work would be greatly limited, and my findings less robust and generalizable. However, there are still significant improvements to data-sharing practices for genomic data. In particular, standardization of data files and types is needed. For example, DNAm microarray data should be deposited into repositories like GEO as .IDAT files, which are the most raw form of the data - the data that are downloaded from the microarray scanners. Instead, GEO allows researchers to deposit the data as plain text files, which means that some important quality control information is lost, and sometimes SNP probes can be missing. Another area in genomic data sharing practices needing significant improvement is the attention towards better quality sample information or study metadata. Often, genomic data is deposited without the necessary metadata (e.g. sex, ethnicity, gestational age, health information) to reproduce previous findings, or to conduct meaningful follow-up analyses. Part of my thesis addresses this limitation, by using the genomic DNAm data directly to infer such sample variables. Still though, DNAm-based inference of sample variables is limited, and having reported sample information has benefits such as cross-checking for misannotation and sample mixups.

The high resolution DNAm microarrays from Illumina were used to assess DNAm in all studies in this thesis. Although cost efficient, and ease-of-use, this results in several limitations. First, these microarrays are limited in their coverage to 450,000 and 850,000 CpGs (450k/850k arrays respectively), which amount to 1.5 and 3% of all CpGs in the genome. These arrays also do not provide uniform coverage; the 450k array focuses on gene elements, such as CpG island promoters (190), whereas the 850k array improves on the 450k design with the addition of coverage over distal elements such as enhancers (21). Although this specific coverage is aimed at including genomic regions with most likelihood of meaningful DNAm variation, improvements in total coverage are an obvious area to improve in future array generations.

The topic of my research was to identify and characterize major sources of variation in placental DNAm. One significant limitation of the studies in this dissertation, is a main focus was on term placenta, with exception of chapter 2 where first trimester samples were included, albeit at lower sample numbers than term samples (n=7-9 vs 19-20 for term). Placental DNAm varies significantly throughout gestation (177), likely from the dramatic development-related changes that occur within placental cell types and also in placental cell composition throughout pregnancy. From a bionformatics point of view, developing the placental analysis tools in term placenta was ideal because most placental research is conducted in term samples. However, many critical processes occur during gestation at earlier timepoints; for example at 10-12 weeks intervillous blood flow dramatically increases with corresponding increases in oxygen tension (191). Increased oxygen levels promotes placental extra-trophoblasts to start differentiating and to acquire invasive behaviour (192), which is critical to a healthy pregnancy. For exampe, in preeclampsia, extra-trophoblast invasion is reduced, resulting in lowered blood flow to the baby and hypertension in the mother (193). Investigation of the molecular processes, such as DNAm, that occur at early gestation timepoints in the placenta are therefore needed for a comprehensive understanding of placental biology and health.

Early gestation samples, however, are difficult to collect due to the safety risks posed by invasive sampling procedures. To this end, advances in non-invasive technologies and in-vitro models will be highly beneficial. Cell free DNA, which targets the placental DNA that has been shedded into the maternal blood stream, is a promising new non-invasive genomic technique. A key challenge of cell free DNA for measurement of placental DNA, however, is that sensitivity is a function of the proportion of placental DNA to maternal DNA, which increases with gestation itself (194). Currently, non-invasive prenatal testing can reliably detect large chromosomal abnormalities (195), but future developments will determine whether high resolution profiling such as sequencing and microarrays can be reliably used as well. Although *in utero* profiling provides the most accurate information, recent organoid developments promises a more convenient in vitro option that is not weighted with the challenges of non-invasive technologies (196).

5.3 Future directions

I developed several bioinformatic approaches to help analyze placental DNAm data, and implemented these in user-friendly applications for usage in future placental DNAm studies. These approaches allow factors that can significantly contribute to DNAm variability to be predicted, modeled, and accounted for. With these predicted variables, measuring placental DNAm provides a multi-facetted picture of each placental sample that can be useful for identifying changes associated with disease and environment. However, given the limitations of the studies and techniques outlined above, there are significant improvements that can be

still made.

It is untested how planet will perform in other continents of diverse populations. However, because planet was developed in diverse human populations from North America, it is possible that planet may generalize well to other populations. Related to this, more placental research in additional human populations are needed, and future improvements to planet and other population stratification methods can include additional populations. It is also still unclear if a machine learning -based approach is superior to semi-supervised PCA -based approaches for accounting for genetic ancestry in EWAS. Much of the existing literature focuses on PCA -based approaches, given their success in accounting for genetic ancestry in GWAS (197). However, machine learning approaches in DNAm data, to estimate or types of sample variables such as epigenetic age and cancer subtypes, have been highly robust and useful (141,177,198). In my study of predicting ethnicity from placental DNAm, I show that using both approaches can be complementary: machine learning to predict discrete categories (ethnicity), and semi-supervised PCA to generate continuous variables (genetic ancestry).

My study characterizing the DNAm in placental cell types from first trimester and term placentas, is the first to do so, and has provided a useful reference for placental studies already (15,180,199). Aside from improvements in genomic coverage, as mentioned previously, inclusion of additional cell populations and gestational timepoints are needed. In this study, trophoblasts represented by mostly CTB, and STB were collected. However, an important additional trophoblast subtype, EVTs, which are responsible for invasion and arterial remodelling in maternal tissue, were not included. EVTs play important additional roles to nutrient transport and placental development, and when impaired are associated with various placental-mediated disease phenotypes such as preeclampsia, intrauterine growth restriction, stillbirth, or recurrent abortion (200). STB were profiled for DNAm, but the nuclei in this cell layer condense of a heterogeneous mixture of gestationally varied nuclei, with a higher proportion of older nuclei condensing into syncytial knots at later gestation (201). Placental-associated maternal macrophages are also present in the placenta that adhere to and aid repair of the placental surface (202). Placental single cell RNA-seq (scRNAseq) studies have also demonstrated significant cell-type heterogeneity defined by gene expression clustering (203,204), which suggests that there there may be significant cell type heterogeneity in placental DNAm to be explored. To this end, single cell whole genome bisulfite sequencing (scWGBS) is an obvious next step for the unbiased characterization of DNAm in placental cell types (205). Novel cell types can be discovered using scWGBS, that can be missed when using other technologies like scRNAseq. For example, using scWGBS, Luo 2017 discoverd novel cell subtypes in brain cortex that were missed in past scRNAseq analysis (206). Additionally, because DNAm reflects developmental lineages and cell function, scWGBS is particularly well-suited for the

study of cell lineage, which can help resolve some areas of placental development, for example trophoblast differentiation (200). Therefore, placental scWGBS can clarify the relationships of various trophoblast subtypes, and additionally lead to insights into the epigenetic regulation of specific gene programs underlying placental development.

5.4 Conclusion

Our understanding of variation in placental DNAm is growing at an unprecendented rate, due to advances in genomic profiling technologies, bioinformatics, and placental-specific sampling and modelling technologies. This dissertation provides insights into placental DNAm and three contributing factors: ethnicity and genetic ancestry, cell-specific epigenetics and cell composition, and maternal cell contamination. These insights into placental DNAm improve our understanding of the molecular dynamics of placental biology, which may lead to future insights into how placental biology contributes to maternal and fetal health. Future studies, however, are still needed to further characterize the rich source of biological information, that is the placental methylome.

Bibliography

1. Jang HS, Shin WJ, Lee JE, Do JT. CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. Genes [Internet]. 2017 Jun;8(6):148. Available from: https://www.mdpi.com/2073-4425/8/6/148

2. Wang Y, Zhao S. Placental Blood Circulation [Internet]. Morgan & Claypool Life Sciences; 2010 [cited 2021 Oct 10]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK53254/

3. Malhotra A, Allison BJ, Castillo-Melendez M, Jenkin G, Polglase GR, Miller SL. Neonatal morbidities of fetal growth restriction: Pathophysiology and impact. Frontiers in Endocrinology [Internet]. 2019;10:55. Available from: https://www.frontiersin.org/article/10.3389/fendo.2019.00055

4. Napso T, Yong HEJ, Lopez-Tello J, Sferruzzi-Perri AN. The role of placental hormones in mediating maternal adaptations to support pregnancy and lactation. Frontiers in Physiology [Internet]. 2018;9:1091. Available from: https://www.frontiersin.org/article/10.3389/fphys.2018.01091

5. Napso T, Yong HEJ, Lopez-Tello J, Sferruzzi-Perri AN. The role of placental hormones in mediating maternal adaptations to support pregnancy and lactation. Frontiers in Physiology [Internet]. 2018;9:1091. Available from: https://www.frontiersin.org/article/10.3389/fphys.2018.01091

6. Guleria I, Sayegh MH. Maternal Acceptance of the Fetus: True Human Tolerance. The Journal of Immunology [Internet]. 2007 Mar 15;178(6):3345–51. Available from: https://www.jimmunol.org/content/178/6/3345

7. Wang X-Q, Li D-J. The mechanisms by which trophoblast-derived molecules induce maternal–fetal immune tolerance. Cellular & Molecular Immunology [Internet]. 2020 Nov [cited 2021 Nov 8];17(11):1204–7. Available from: https://www.nature.com/articles/s41423-020-0460-5

8. Hoo R, Nakimuli A, Vento-Tormo R. Innate immune mechanisms to protect against infection at the human decidual-placental interface. Frontiers in Immunology [Internet]. 2020;11:2070. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2020.02070

9. Delorme-Axford E, Sadovsky Y, Coyne CB. The Placenta as a Barrier to Viral Infections. Annual Review of Virology. 2014 Nov;1(1):133–46.

10. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell [Internet]. 1985 Jan 1;40(1):91–9. Available from: https://www.sciencedirect.com/science/article/pii/0092867485903125

11. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology [Internet]. 2013 Jan;38(1):23–38. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3521964/

12. Schulz WA, Steinhoff C, Florl AR. Methylation of Endogenous Human Retroelements in Health and Disease. In: Doerfler W, Böhm P, editors. Berlin, Heidelberg: Springer; 2006. p. 211–50. (Current Topics in Microbiology and Immunology). Available from: https://doi.org/10.1007/3-540-31181-5_11

13. Elhamamsy AR. Role of DNA methylation in imprinting disorders: an updated review. Journal of Assisted Reproduction and Genetics [Internet]. 2017 May;34(5):549–62. Available from: http://link.springer.com/10.1007/s10815-017-0895-5

14. Cotton AM, Avila L, Penaherrera MS, Affleck JG, Robinson WP, Brown CJ. Inactive X chromosome-specific reduction in placental DNA methylation. Human Molecular Genetics [Internet]. 2009 Oct 1;18(19):3544–52. Available from: https://academic.oup.com/hmg/article/18/19/3544/2385630

15. Del Gobbo GF, Yuan V, Robinson WP. Confined placental mosaicism involving multiple de novo copy number variants associated with fetal growth restriction: A case report. American Journal of Medical Genetics Part A [Internet]. 2021;185(6):1908–12. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.a.62183

16. Del Gobbo GF, Konwar C, Robinson WP. The significance of the placental genome and methylome in fetal and maternal health. Human Genetics [Internet]. 2019 Sep [cited 2019 Oct 31]; Available from: https://doi.org/10.1007/s00439-019-02058-w

17. Hackett JA, Sengupta R, Zylicz JJ, Murakami K, Lee C, Down TA, et al. Germline DNA Demethylation Dynamics and Imprint Erasure Through 5-Hydroxymethylcytosine. Science [Internet]. 2013 Jan 25;339(6118):448–52. Available from: https://science.sciencemag.org/content/339/6118/448

18. Hargan-Calvopina J, Taylor S, Cook H, Hu Z, Lee SA, Yen M-R, et al. Stage-specific demethylation in primordial germ cells safeguards against precocious differentiation. Developmental cell [Internet]. 2016 Oct 10;39(1):75–86. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5064860/

19. Fraser R, Lin C-J. Epigenetic reprogramming of the zygote in mice and men: On your marks, get set, go! Reproduction (Cambridge, England) [Internet]. 2016 Oct 20;152(6):R211–22. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5097126/

20. Xu R, Li C, Liu X, Gao S. Insights into epigenetic patterns in mammalian early embryos. Protein & Cell [Internet]. 2021 Jan 1;12(1):7–28. Available from: https://doi.org/10.1007/s13238-020-00757-z

21. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology [Internet]. 2016;17(1):208–8. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1066-1

22. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k DNA methylation data. Bioinformatics (Oxford, England) [Internet]. 2013 Jan;29(2):189–96. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23175756

23. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD, Siegmund KD. Low-level processing of illumina infinium DNA methylation BeadArrays. Nucleic acids research [Internet]. 2013 Apr;41(7):e90–0. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23476028

24. Chu S-K, Yang H-C. Interethnic DNA methylation difference and its implications in pharmacoepigenetics. Epigenomics [Internet]. 2017;9. Available from: https://doi.org/10.2217/epi-2017-0046

25. Smith JA, Zhao W, Wang X, Ratliff SM, Mukherjee B, Kardia SLR. Neighborhood characteristics influence DNA methylation of genes involved in stress response and inflammation: The multi-ethnic study of atherosclerosis. Epigenetics [Internet]. 2017;12. Available from: https://doi.org/10.1080/15592294.2017.1341026

26. Park SL, Patel YM, Loo LWM, Mullen DJ, Offringa IA, Maunakea A. Association of internal smoking dose with blood DNA methylation in three racial/ethnic populations. Clin Epigenet [Internet]. 2018;10. Available from: https://doi.org/10.1186/s13148-018-0543-7

27. Adkins RM, Krushkal J, Tylavsky FA, Thomas F. Racial differences in gene-specific DNA methylation levels are present at birth. Birth Defects Res A Clin Mol Teratol [Internet]. 2011;91. Available from: https://doi.org/10.1002/bdra.20770

28. Kwabi-Addo B, Wang S, Chung W, Jelinek J, Patierno SR, Wang B-D, et al. Identification of differentially methylated genes in normal prostate tissues from african american and caucasian men. Clinical cancer research : an official journal of the American Association for Cancer Research [Internet]. 2010 Jul;16(14):3539–47. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20606036

29. Liu J, Hutchison K, Perrone-Bizzozero N, Morgan M, Sui J, Calhoun V. Identification of genetic and epigenetic marks involved in population structure. PloS one [Internet]. 2010 Oct;5(10):e13209–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20949057

30. Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. Genome biology [Internet]. 2012 Feb;13(2):R8–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22322129

31. Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R. Accounting for population stratification in DNA methylation studies. Genet Epidemiol [Internet]. 2014;38. Available from: https://doi.org/10.1002/gepi.21789

32. Rahmani E, Shenhav L, Schweiger R, Yousefi P, Huen K, Eskenazi B, et al. Genome-wide methylation data mirror ancestry information. Epigenetics & Chromatin [Internet]. 2017;10(1):1–1. Available from: http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-016-0108-y

33. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of infinium DNA methylation BeadChip probes. Nucleic Acids Research [Internet]. 2017;45(4):gkw967–7. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw967

34. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: Population substructure and genome-wide association studies. Human molecular genetics [Internet]. 2008 Oct;17(R2):R143–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18852203

35. Sankar P, Cho MK, Mountain J. Race and ethnicity in genetic research. American journal of medical genetics Part A [Internet]. 2007 May;143A(9):961–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17345638

36. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: Its potential impact on understanding health disparities. Human Genomics [Internet]. 2015 Jan;9(1):1–1. Available from: http://www.humgenomics.com/content/9/1/1

37. Kaneshiro B, Geling O, Gellert K, Millar L. The challenges of collecting data on race and ethnicity in a diverse, multiethnic state. Hawaii Med J. 2011;70.

38. Banda Y, Kvale MN, Hoffmann TJ, Hesselson SE, Ranatunga D, Tang H. Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics [Internet]. 2015;200. Available from: https://doi.org/10.1534/genetics.115.178616

39. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet [Internet]. 1999;65. Available from: https://doi.org/10.1086/302449

40. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet [Internet]. 2006;38. Available from: https://doi.org/10.1038/ng1847

41. Delahaye F, Do C, Kong Y, Ashkar R, Salas M, Tycko B, et al. Genetic variants influence on the placenta regulatory landscape. Snyder M, editor. PLOS Genetics [Internet]. 2018 Nov;14(11):e1007785–5. Available from: http://dx.plos.org/10.1371/journal.pgen.1007785

42. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome biology [Internet]. 2011;12(1):R10–0. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21251332

43. Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, et al. A global reference for human genetic variation. Nature [Internet]. 2015 Oct;526(7571):68–74. Available from: http://www.nature.com/articles/nature15393

44. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of infinium DNA methylation BeadChip probes. Nucleic Acids Research [Internet]. 2017;45(4):gkw967–7. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw967

45. Yuan V, Price EM, Del Gobbo G, Mostafavi S, Cox B, Binder AM, et al. Accurate ethnicity prediction from placental DNA methylation data. Epigenetics & Chromatin [Internet]. 2019 Aug 9;12(1):51. Available from: https://doi.org/10.1186/s13072-019-0296-3

46. Grigoriu A, Ferreira JC, Choufani S, Baczyk D, Kingdom J, Weksberg R. Cell specific patterns of methylation in the human placenta. Epigenetics. 2011;

47. Oda M, Oxley D, Dean W, Reik W. Regulation of lineage specific DNA hypomethylation in mouse trophectoderm. PLoS ONE [Internet]. 2013;8(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3692478/pdf/pone.0068846.pdf

48. Novakovic B, Wong NC, Sibson M, Ng H-K, Morley R, Manuelpillai U, et al. DNA methylation-mediated down-regulation of DNA methyltransferase-1 (DNMT1) is coincident with, but not essential for, global hypomethylation in human placenta. The Journal of biological chemistry [Internet]. 2010 Mar;285(13):9583–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20071334

49. Nelissen ECM, van Montfoort APA, Dumoulin JCM, Evers JLH. Epigenetics and the placenta. Human Reproduction Update. 2011 Jun;17(3):397–417.

50. Hemberger M. Epigenetic landscape required for placental development. Cellular and Molecular Life Sciences [Internet]. 2007 Sep 1;64(18):2422–36. Available from: https://doi.org/10.1007/s00018-007-7113-z

51. Soncin F, Natale D, Parast MM. Signaling pathways in mouse and human trophoblast differentiation: A comparative review. Cellular and molecular life sciences : CMLS [Internet]. 2015;72(7):1291–302. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4366325/pdf/nihms645576.pdf

52. Maltepe E, Krampitz GW, Okazaki KM, Red-Horse K, Mak W, Simon MC, et al. Hypoxia-inducible factor-dependent histone deacetylase activity determines stem cell fate in the placenta. Development [Internet]. 2005;132(15). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4366325/pdf/nihms645576.pdf

53. Yuen RKC, Chen B, Blair JD, Robinson WP, Nelson DM. Hypoxia alters the epigenetic profile in cultured human placental trophoblasts. Epigenetics [Internet]. 2013 Feb;8(2):192–202. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23314690

54. Ruebner M, Strissel PL, Ekici AB, Stiegler E, Dammer U, Goecke TW, et al. Reduced syncytin-1 expression levels in placental syndromes correlates with epigenetic hypermethylation of the ERVW-1 promoter region. PloS one [Internet]. 2013;8(2):e56145–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23457515

55. Chuang H-C, Chang C-W, Chang G-D, Yao T-P, Chen H. Histone deacetylase 3 binds to and regulates the GCMa transcription factor. Nucleic Acids Research [Internet]. 2006 Mar 1;34(5):1459–69. Available from: https://doi.org/10.1093/nar/gkl048

56. Burton GJ, Woods AW, Jauniaux E, Kingdom JCP. Rheological and physiological consequences of conversion of the maternal spiral arteries for uteroplacental blood flow during human pregnancy. Placenta [Internet]. 2009;30(6):473–82. Available from: https://ac.els-cdn.com/S0143400409000666/1-s2.0-S0143400409000666-main.pdf?_tid=37fa7318-b80f-11e7-8208-00000aab0f01&acdnat=1508776296_f9cd4c6ca7d41cc181ac6c662ac9dc18

57. Fogarty NME, Burton GJ, Ferguson-smith AC. Different epigenetic states de fi ne syncytiotrophoblast and cytotrophoblast nuclei in the trophoblast of the human placenta. Placenta [Internet]. 2015;36(8):796–802. Available from: http://dx.doi.org/10.1016/j.placenta.2015.05.006

58. Davies JE, Pollheimer J, J Yong HE, Kokkinos MI, Kalionis B, Knöfler M, et al. Cell adhesion & migration epithelial-mesenchymal transition during extravillous trophoblast differentiation epithelial-mesenchymal transition during extravillous trophoblast differentiation. Cell Adhesion & Migration [Internet]. 2016;103:310–21. Available from: http://www.tandfonline.com/action/journalInformation?journalCode=kcam20

59. Haider S, Meinhardt G, Saleh L, Fiala C, Pollheimer J, Knöfler M. Notch1 controls development of the extravillous trophoblast lineage in the human placenta. Proceedings of the National Academy of Sciences [Internet]. 2016;113(48):E7710–9. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1612335113

60. Rodesch F, Simon P, Donner C, Jauniaux E. Oxygen measurements in endometrial and trophoblastic tissues during early pregnancy. Obstetrics and gynecology. 1992 Aug;80(2):283–5.

61. Wakeland AK, Soncin F, Moretto-Zita M, Chang C-W, Horii M, Pizzo D, et al. Hypoxia directs human extravillous trophoblast differentiation in a hypoxia-inducible factor–dependent manner. The American Journal of Pathology [Internet]. 2017 Apr 1;187(4):767–80. Available from: http://www.sciencedirect.com/science/article/pii/S0002944017300214

62. Chen Y, Wang K, Qian C-N, Leach R. DNA methylation is associated with transcription of snail and slug genes. Biochem Biophys Res Commun. 2013;230(3):1083–90.

63. Dokras A, Gardner LM, Kirschmann DA, Seftor EA, Hendrix MJ. The tumour suppressor gene maspin is differentially regulated in cytotrophoblasts during human placental development. Placenta [Internet]. 2002;23(0143-4004 LA - eng PT - Journal Article RN - 0 (Proteins) RN - 0 (RNA, Messenger) RN - 0 (Recombinant Proteins) RN - 0 (Serpins) RN - 0 (maspin) SB - IM):274–80. Available from: http://www.idealibrary.com

64. Abell AN, Jordan NV, Huang W, Prat A, Midland AA, Johnson NL, et al. MAP3K4/CBP-regulated H2B acetylation controls epithelial-mesenchymal transition in trophoblast stem cells. Cell Stem Cell [Internet]. 2011;8(5):525–37. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3091002/pdf/nihms285841.pdf

65. Tsai Y-C, Cooke NE, Liebhaber SA. Tissue specific CTCF occupancy and boundary function at the human growth hormone locus. Nucleic Acids Research [Internet]. 2014 Apr 1;42(8):4906–21. Available from: https://doi.org/10.1093/nar/gku139

66. Robinson WP, Price EM. The human placental methylome. Cold Spring Harb Perspect Med [Internet]. 2015;5. Available from: https://doi.org/10.1101/cshperspect.a023044

67. Kim SY, Romero R, Tarca AL, Bhatti G, Kim CJ, Lee J, et al. Methylome of fetal and maternal monocytes and macrophages at the feto-maternal interface. American journal of reproductive immunology (New York, NY : 1989) [Internet]. 2012 Jul;68(1):8–27. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3479407/

68. Holbrook JD, Huang R-C, Barton SJ, Saffery R, Lillycrop KA. Is cellular heterogeneity merely a confounder to be removed from epigenome-wide association studies? Epigenomics [Internet]. 2017 Aug;9(8):1143–50. Available from: http://www.futuremedicine.com/doi/10.2217/epi-2017-0032

69. Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, et al. A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature [Internet]. 2012 Mar 28;484(7394):339–44. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3331945/

70. Benirschke K. The human placenta. J. D. Boyd and W. J. Hamilton. Heffer, Cambridge, 365 pp. 1970. Teratology [Internet]. 1973;8(1):77–8. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/tera.1420080118

71. Reyes L, Wolfe B, Golos T. Hofbauer cells: Placental macrophages of fetal origin. In Springer, Cham; 2017. p. 45–60. Available from: http://link.springer.com/10.1007/978-3-319-54090-0_3

72. Reyes L, Golos TG. Hofbauer cells: Their role in healthy and complicated pregnancy. Frontiers in Immunology [Internet]. 2018;9. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2018.02628/full

73. Charolidi N, Host AJ, Ashton S, Tryfonos Z, Leslie K, Thilaganathan B, et al. First trimester placental endothelial cells from pregnancies with abnormal uterine artery Doppler are more sensitive to apoptotic stimuli. Laboratory Investigation [Internet]. 2019 Mar;99(3):411–20. Available from: https://www.nature.com/articles/s41374-018-0139-z

74. Novakovic B, Evain-Brion D, Murthi P, Fournier T, Saffery R. Variable DAXX gene methylation is a common feature of placental trophoblast differentiation, preeclampsia, and response to hypoxia. The FASEB Journal [Internet]. 2017 Feb 21;31(6):2380–92. Available from: https://www.fasebj.org/doi/full/10.1096/fj.201601189RR

75. Cotton AM, Avila L, Penaherrera MS, Affleck JG, Robinson WP, Brown CJ. Inactive X chromosome-specific reduction in placental DNA methylation. Human Molecular Genetics [Internet]. 2009 Oct 1;18(19):3544–52. Available from: https://academic.oup.com/hmg/article/18/19/3544/2385630

76. Perrin D, Ballestar E, Fraga MF, Frappart L, Esteller M, Guerin J-F, et al. Specific hypermethylation of LINE-1 elements during abnormal overgrowth and differentiation of human placenta. Oncogene [Internet]. 2007 Apr;26(17):2518–24. Available from: https://www.nature.com/articles/1210039

77. Schroeder DI, Blair JD, Lott P, Yu HOK, Hong D, Crary F, et al. The human placenta methylome. Proceedings of the National Academy of Sciences [Internet]. 2013 Apr 9;110(15):6037–42. Available from: https://www.pnas.org/content/110/15/6037

78. Renfree MB, Suzuki S, Kaneko-Ishino T. The origin and evolution of genomic imprinting and viviparity in mammals. Philosophical Transactions of the Royal Society B: Biological Sciences [Internet]. 2013 Jan 5;368(1609). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3539366/

79. Hanna CW, Peñaherrera MS, Saadeh H, Andrews S, McFadden DE, Kelsey G. Pervasive polymorphic imprinted methylation in the human placenta. Genome Res [Internet]. 2016;26. Available from: https://doi.org/10.1101/gr.196139.115

80. Court F, Tayama C, Romanelli V, Martin-Trujillo A, Iglesias-Platas I, Okamura K, et al. Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. Genome Research [Internet]. 2014 Apr;24(4):554–69. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3975056/

81. Hamada H, Okae H, Toh H, Chiba H, Hiura H, Shirane K, et al. Allele-Specific Methylome and Transcriptome Analysis Reveals Widespread Imprinting in the Human Placenta. The American Journal of Human Genetics [Internet]. 2016 Nov 3;99(5):1045–58. Available from: http://www.sciencedirect.com/science/article/pii/S0002929716303780

82. Sanchez-Delgado M, Court F, Vidal E, Medrano J, Monteagudo-Sánchez A, Martin-Trujillo A, et al. Human Oocyte-Derived Methylation Differences Persist in the Placenta Revealing Widespread Transient Imprinting. PLOS Genetics [Internet]. 2016 Nov 11;12(11):e1006427. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006427

83. Flanagan JM. Epigenome-Wide Association Studies (EWAS): Past, Present, and Future. In: Verma M, editor. New York, NY: Springer; 2015. p. 51–63. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-4939-1804-1_3

84. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: A review and recommendations. Epigenomics [Internet]. 2017 May;9(5):757–68. Available from: http://www.futuremedicine.com/doi/10.2217/epi-2016-0153

85. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. BMC Bioinformatics [Internet]. 2017 Dec;18(1):105–5. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1511-5

86. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics [Internet]. 2012;13(1):86–6. Available from: http://www.biomedcentral.com/1471-2105/13/86

87. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods [Internet]. 2015 May;12(5):453–7. Available from: http://www.nature.com/articles/nmeth.3337

88. Houseman EA, Kile ML, Christiani DC, Ince TA, Kelsey KT, Marsit CJ. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics [Internet]. 2016 Dec;17(1):259–9. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1140-4

89. Konwar C, Del Gobbo G, Yuan V, Robinson WP. Considerations when processing and interpreting genomics data of the placenta. Placenta [Internet]. 2019 Sep 1;84:57–62. Available from: http://www.sciencedirect.com/science/article/pii/S0143400418311470

90. Heiss JA, Just AC. Identifying mislabeled and contaminated DNA methylation microarray data: An extended quality control toolset with examples from GEO. Clinical Epigenetics [Internet]. 2018 Jun 1;10(1):73. Available from: https://doi.org/10.1186/s13148-018-0504-1

91. Turco MY, Moffett A. Development of the human placenta. Development [Internet]. 2019 Nov 15;146(22). Available from: https://dev.biologists.org/content/146/22/dev163428

92. Price EM, Peñaherrera MS, Casamar EP, Pavlidis P, Allen MIV, Mcfadden DE, et al. Profiling placental and fetal DNA methylation in human neural tube defects. Epigenetics & Chromatin. 2016;1–4.

93. Thomas JR, Appios A, Zhao X, Dutkiewicz R, Donde M, Lee CYC, et al. Phenotypic and functional characterization of first-trimester human placental macrophages, hofbauer cells. Journal of Experimental Medicine [Internet]. 2020 Oct 19;218(1). Available from: https://doi.org/10.1084/jem.20200891

94. Konwar C, Price EM, Wang LQ, Wilson SL, Terry J, Robinson WP. DNA methylation profiling of acute chorioamnionitis-associated placentas and fetal membranes: Insights into epigenetic variation in spontaneous preterm births. Epigenet Chromatin [Internet]. 2018;11. Available from: https://doi.org/10.1186/s13072-018-0234-9

95. Wan Q, Leemaqz SY-L, Pederson SM, McCullough D, McAninch DC, Jankovic-Karasoulos T, et al. Quality control measures for placental sample purity in DNA methylation array analyses. Placenta [Internet]. 2019 Dec;88:8–11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7705927/

96. Salhab A, Nordström K, Gasparoni G, Kattler K, Ebert P, Ramirez F, et al. A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. Genome Biology [Internet]. 2018 Sep 28;19(1):150. Available from: https://doi.org/10.1186/s13059-018-1510-5

97. Morin AM, Gatev E, McEwen LM, MacIsaac JL, Lin DTS, Koen N, et al. Maternal blood contamination of collected cord blood can be identified using DNA methylation at three CpGs. Clinical Epigenetics. 2017;9:75.

98. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res [Internet]. 2002;30. Available from: https://doi.org/10.1093/nar/30.1.207

99. Wilson SL, Leavey K, Cox BJ, Robinson WP. Mining DNA methylation alterations towards a classification of placental pathologies. Human molecular genetics [Internet]. 2018 Jan;27(1):135–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29092053

100. Leavey K, Wilson SL, Bainbridge SA, Robinson WP, Cox BJ. Epigenetic regulation of placental gene expression in transcriptional subtypes of preeclampsia. Clin Epigenet [Internet]. 2018;10. Available from: https://doi.org/10.1186/s13148-018-0463-6

101. Martin E, Ray PD, Smeester L, Grace MR, Boggess K, Fry RC. Epigenetics and preeclampsia: Defining functional epimutations in the preeclamptic placenta related to the TGF-? pathway. PLoS ONE [Internet]. 2015;10. Available from: https://doi.org/10.1371/journal.pone.0141294

102. Paquette AG, Houseman EA, Green BB, Lesseur C, Armstrong DA, Lester B. Regions of variable DNA methylation in human placenta associated with newborn neurobehavior. Epigenetics [Internet]. 2016;11. Available from: https://doi.org/10.1080/15592294.2016.1195534

103. Binder AM, LaRocca J, Lesseur C, Marsit CJ, Michels KB. Epigenome-wide and transcriptome-wide analyses reveal gestational diabetes is associated with alterations in the human leukocyte antigen complex. Clinical epigenetics [Internet]. 2015;7(1):79–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26244062

104. Price EM, Robinson WP. Adjusting for batch effects in DNA methylation microarray data, a lesson learned. Frontiers in genetics [Internet]. 2018;9:83–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29616078

105. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol [Internet]. 2013;14. Available from: https://doi.org/10.1186/gb-2013-14-10-r115

106. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Stat Sci [Internet]. 2003;18. Available from: https://doi.org/10.1214/ss/1056397488

107. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer genomics & proteomics [Internet]. 2018;15(1):41–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29275361

108. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software [Internet]. 2010;33(1):1–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20808728

109. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B [Internet]. 2005;67(2):301–20. Available from: https://web.stanford.edu/~hastie/Papers/B67.2%20(2005)%20301-320%20Zou%20&%20Hastie.pdf

110. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat [Internet]. 1992;46. Available from: https://doi.org/10.1080/00031305.1992.10475879

111. Cortes C, Vapnik V. Support-vector networks. Machine Learning [Internet]. 1995 Sep;20(3):273–97. Available from: http://link.springer.com/10.1007/BF00994018

112. Daca-Roszak P, Pfeifer A, Żebracka-Gala J, Rusinek D, Szybińska A, Jarząb B, et al. Impact of SNPs on methylation readouts by illumina infinium HumanMethylation450 BeadChip array: Implications for comparative population studies. BMC genomics [Internet]. 2015 Nov;16:1003–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26607064

113. Walser J-C, Furano AV. The mutational spectrum of non-CpG DNA varies with CpG content. Genome Res [Internet]. 2010;20. Available from: https://doi.org/10.1101/gr.103283.109

114. Del Gobbo GF, Price EM, Hanna CW, Robinson WP. No evidence for association of MTHFR 677C>t and 1298A>c variants with placental DNA methylation. Clinical epigenetics [Internet]. 2018;10:34–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29564022

115. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics [Internet]. 2010 Jul;11(7):459–63. Available from: http://www.nature.com/articles/nrg2813

116. Frichot E, François O. LEA : An r package for landscape and ecological association studies. O'Meara B, editor. Methods in Ecology and Evolution [Internet]. 2015 Aug;6(8):925–9. Available from: http://doi.wiley.com/10.1111/2041-210X.12382

117. Roslin NM, Li W, Paterson AD, Strug LJ. Quality control analysis of the 1000 genomes project Omni2.5 genotypes. 2016; Available from: http://www.1000genomes.org/faq/how-do-i-cite-1000-genomes-project.

118. Green BB, Karagas MR, Punshon T, Jackson BP, Robbins DJ, Houseman EA. Epigenome-wide assessment of DNA methylation in the placenta and arsenic exposure in the new hampshire birth cohort study (USA). Environ Health Perspect [Internet]. 2016;124. Available from: https://doi.org/10.1289/ehp.1510437

119. Chen Y, Wang K, Leach R. 5-aza-dC treatment induces mesenchymal-to-epithelial transition in 1st trimester trophoblast cell line HTR8/SVneo. Biochem Biophys Res Commun. 2013;432(1):116–22.

120. Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ. Additional annotation enhances potential for biologically-relevant analysis of the illumina infinium HumanMethylation450 BeadChip array. Epigenet Chromatin [Internet]. 2013;6. Available from: https://doi.org/10.1186/1756-8935-6-4

121. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet [Internet]. 2006;38. Available from: https://doi.org/10.1038/ng1847

122. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155.

123. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research [Internet]. 2009 Sep;19(9):1655–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19648217

124. Breton CV, Marsit CJ, Faustman E, Nadeau K, Goodrich JM, Dolinoy DC. Small-magnitude effect sizes in epigenetic end points are important in children's environmental health studies: The children's environmental health and disease prevention research center's epigenetics working group. Environ Health Perspect [Internet]. 2017;125. Available from: https://doi.org/10.1289/EHP595

125. Bryant AS, Worjoloh A, Caughey AB, Washington AE. Racial/ethnic disparities in obstetric outcomes and care: Prevalence and determinants. Am J Obstet Gynecol [Internet]. 2010;202. Available from: https://doi.org/10.1016/j.ajog.2009.10.864

126. Xiao J, Shen F, Xue Q, Chen G, Zeng K, Stone P, et al. Is ethnicity a risk factor for developing preeclampsia? An analysis of the prevalence of preeclampsia in china. Journal of Human Hypertension [Internet]. 2014 Nov;28(11):694–8. Available from: http://www.nature.com/articles/jhh2013148

127. Rosenberg TJ, Garbers S, Lipkind H, Chiasson MA. Maternal obesity and diabetes as risk factors for adverse pregnancy outcomes: Differences among 4 racial/ethnic groups. American journal of public health [Internet]. 2005 Sep;95(9):1545–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16118366

128. United states census bureau. Population and housing. 2018. Https://www.census.gov/en.html. Accessed 18 june 2018. Available from: https://www.census.gov/en.html

129. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of african americans, latinos, and european americans across the united states. American journal of human genetics. 2015 Jan;96(1):37–53.

130. Hastie T, Tibshirani R, Narasimhan B, Chu G. Impute: Imputation for microarray data. R package version 1560 [Internet]. 2018; Available from: http://bioconductor.org/packages/release/bioc/html/impute.html

131. Edgar RD, Jones MJ, Robinson WP, Kobor MS. An empirically driven data reduction method on the human 450K methylation array to remove tissue specific non-variable CpGs. Clin Epigenet [Internet]. 2017;9. Available from: https://doi.org/10.1186/s13148-017-0320-z

132. Triche TJ, Weisenberger DJ, Berg D, Laird PW, Siegmund KD, Siegmund KD. Low-level processing of illumina infinium DNA methylation BeadArrays. Nucleic Acids Res [Internet]. 2013;41. Available from: https://doi.org/10.1093/nar/gkt090

133. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450k DNA methylation data. Bioinformatics [Internet]. 2013;29. Available from: https://doi.org/10.1093/bioinformatics/bts680

134. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics [Internet]. 2014 May 15;30(10):1363–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4016708/

135. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing illumina 450K methylation array data. BMC Genomics [Internet]. 2013 May 1;14(1):293. Available from: https://doi.org/10.1186/1471-2164-14-293

136. Del Gobbo GF, Price EM, Hanna CW, Robinson WP. No evidence for association of MTHFR 677C>t and 1298A>c variants with placental DNA methylation. Clinical epigenetics [Internet]. 2018;10:34–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29564022

137. Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. Fast and efficient estimation of individual ancestry coefficients. Genetics [Internet]. 2014;196. Available from: https://doi.org/10.1534/genetics.113.160572

138. Phillips C, Aradas AF, Kriegel AK, Fondevila M, Bulbul O, Santos C, et al. Eurasiaplex: A forensic SNP assay for differentiating european and south asian ancestries. Forensic Science International: Genetics [Internet]. 2013 May;7(3):359–66. Available from: https://www.sciencedirect.com/science/article/pii/S1872497313000598?via%3Dihub

139. Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Álvarez-Dios J, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Science International: Genetics [Internet]. 2007 Dec;1(3-4):273–80. Available from: https://www.sciencedirect.com/science/article/pii/S1872497307001019?via%3Dihub

140. Fondevila M, Phillips C, Santos C, Freire Aradas A, Vallone PM, Butler JM, et al. Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. Forensic science international Genetics [Internet]. 2013 Jan;7(1):63–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22749789

141. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol [Internet]. 2013;14. Available from: https://doi.org/10.1186/gb-2013-14-10-r115

142. Kuhn M. Building predictive models in r using the caret package [Internet]. 2008. Available from: http://www.jstatsoft.org/v28/i05/

143. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw [Internet]. 2010;33. Available from: https://doi.org/10.18637/jss.v033.i01

144. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc B [Internet]. 2005;67. Available from: https://doi.org/10.1111/j.1467-9868.2005.00503.x

145. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Stat Sci [Internet]. 2003;18. Available from: https://doi.org/10.1214/ss/1056397488

146. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA [Internet]. 2002;99. Available from: https://doi.org/10.1073/pnas.082099299

147. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician [Internet]. 1992 Aug;46(3):175–85. Available from: http://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879

148. Kuhn M. Building predictive models in r using the caret package. J Stat Softw [Internet]. 2008;28. Available from: https://doi.org/10.18637/jss.v028.i05

149. McHugh ML. Interrater reliability: The kappa statistic. Biochemia medica [Internet]. 2012;22(3):276–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23092060

150. Hansen KD. IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for illumina's 450k methylation arrays. R package version 060. 2016;

151. Phipson B, Maksimovic J, Oshlack A. missMethyl: An r package for analyzing data from illumina's HumanMethylation450 platform. Bioinformatics [Internet]. 2015;32. Available from: https://doi.org/10.1093/bioinformatics/btv560

152. Zhou W, Triche TJ, Laird PW, Shen H. SeSAMe: Reducing artifactual detection of DNA methylation by infinium BeadChips in genomic deletions. Nucleic Acids Research [Internet]. 2018 Jul;46(20):e123–3. Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky691/5061974

153. GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards | database | oxford academic [Internet]. Available from: https://academic.oup.com/database/article/doi/10.1093/database/bax028/3737828

154. Das R, Lee YK, Strogantsev R, Jin S, Lim YC, Ng PY, et al. DNMT1 and AIM1 imprinting in human placenta revealed through a genome-wide screen for allele-specific DNA methylation. BMC Genomics [Internet]. 2013 Oct 5;14(1):685. Available from: https://doi.org/10.1186/1471-2164-14-685

155. Mayhew TM, Barker BL. Villous trophoblast: morphometric perspectives on growth, differentiation, turnover and deposition of fibrin-type fibrinoid during gestation. Placenta. 2001 Aug;22(7):628–38.

156. Liu Y, Fan X, Wang R, Lu X, Dang YL, Wang H, et al. Single-cell RNA-seq reveals the diversity of trophoblast subtypes and patterns of differentiation in the human placenta. Cell Research [Internet]. 2018 Aug;819–32. Available from: http://www.nature.com/articles/s41422-018-0066-y

157. Yabe S, Alexenko AP, Amita M, Yang Y, Schust DJ, Sadovsky Y, et al. Comparison of syncytiotrophoblast generated from human embryonic stem cells and from term placentas. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2016 May 10;113(19):E2598–607. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4868474/

158. Butler TM, Pater JA, MacPhee DJ. Integrin linked kinase regulates syncytialization of BeWo trophoblast cells. Biology of Reproduction [Internet]. 2017 Mar 1;96(3):673–85. Available from: https://academic.oup.com/biolreprod/article/96/3/673/2959520

159. Simmons DG, Rakoczy J, Jefferis J, Lourie R, McIntyre HD, Dawson PA. Human placental sulfate transporter mRNA profiling from term pregnancies identifies abundant SLC13A4 in syncytiotrophoblasts and SLC26A2 in cytotrophoblasts. Placenta [Internet]. 2013 Apr 1;34(4):381–4. Available from: http://www.sciencedirect.com/science/article/pii/S0143400413000726

160. Gervin K, Salas LA, Bakulski KM, van Zelm MC, Koestler DC, Wiencke JK, et al. Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. Clinical Epigenetics [Internet]. 2019 Aug 27;11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6712867/

161. Price EM, Cotton AM, Peñaherrera MS, McFadden DE, Kobor MS, Robinson W. Different measures of "genome-wide" DNA methylation exhibit unique properties in placental and somatic tissues. Epigenetics. 2012 Jun 1;7(6):652–63.

162. Zink F, Magnusdottir DN, Magnusson OT, Walker NJ, Morris TJ, Sigurdsson A, et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. Nature Genetics [Internet]. 2018 Nov;50(11):1542–52. Available from: https://www.nature.com/articles/s41588-018-0232-7

163. Reiss D, Zhang Y, Mager DL. Widely variable endogenous retroviral methylation levels in human placenta. Nucleic Acids Research. 2007;35(14):4743–54.

164. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. Nature [Internet]. 2018 Nov;563(7731):347. Available from: https://www.nature.com/articles/s41586-018-0698-6

165. Zheng SC, Breeze CE, Beck S, Dong D, Zhu T, Ma L, et al. EpiDISH web server: Epigenetic Dissection of Intra-Sample-Heterogeneity with online GUI. Bioinformatics [Internet]. Available from: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz833/5618716

166. Gamage TKJB, Schierding W, Hurley D, Tsai P, Ludgate JL, Bhoothpur C, et al. The role of DNA methylation in human trophoblast differentiation. Epigenetics [Internet]. 2018 Dec 2;13(12):1154–73. Available from: https://doi.org/10.1080/15592294.2018.1549462

167. Ehrlich M, Gama-Sosa MA, Huang L-H, Midgett RM, Kuo KC, McCune RA, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Research [Internet]. 1982 Apr 24;10(8):2709–21. Available from: https://academic.oup.com/nar/article/10/8/2709/1130953

168. Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics. 2013;8(3):290–302.

169. Fox H. Aging of the placenta. Archives of Disease in Childhood - Fetal and Neonatal Edition [Internet]. 1997 Nov 1;77(3):F171–5. Available from: https://fn.bmj.com/content/77/3/F171

170. Roland CS, Hu J, Ren C-E, Chen H, Li J, Varvoutis MS, et al. Morphological changes of placental syncytium and their implications for the pathogenesis of preeclampsia. Cellular and molecular life sciences : CMLS [Internet]. 2016 Jan;73(2):365–76. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4846582/

171. Bartels HC, Postle JD, Downey P, Brennan DJ. Placenta accreta spectrum: A review of pathology, molecular biology, and biomarkers. Disease Markers [Internet]. 2018 Jul 3;2018. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6051104/

172. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research [Internet]. 2015 Apr 20;43(7):e47–7. Available from: https://academic.oup.com/nar/article/43/7/e47/2414268

173. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology [Internet]. 2004;3(1):125. Available from: https://www.degruyter.com/view/j/sagmb.2004.3.issue-1/sagmb.2004.3.1.1027/sagmb.2004.3.1.1027.xml

174. Zheng Y, Joyce BT, Liu L, Zhang Z, Kibbe WA, Zhang W, et al. Prediction of genome-wide DNA methylation in repetitive elements. Nucleic Acids Research [Internet]. 2017 Sep 6;45(15):8697–711. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5587781/

175. Yuan V, Hui D, Yin Y, Peñaherrera MS, Beristain AG, Robinson WP. Cell-specific characterization of the placental methylome. BMC Genomics [Internet]. 2021 Jan 6;22(1):6. Available from: https://doi.org/10.1186/s12864-020-07186-6

176. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks: Challenges and recommendations. Genome Biology [Internet]. 2019 Nov 25;20(1):249. Available from: https://doi.org/10.1186/s13059-019-1824-y

177. Lee Y, Choufani S, Weksberg R, Wilson SL, Yuan V, Burt A, et al. Placental epigenetic clocks: Estimating gestational age using placental DNA methylation levels. Aging (Albany NY) [Internet]. 2019 Jun 24;11(12):4238–53. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6628997/

178. Cappelletti M, Presicce P, Kallapur SG. Immunobiology of acute chorioamnionitis. Frontiers in Immunology [Internet]. 2020 Apr 16;11:649. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7177011/

179. Fogarty NME, Mayhew TM, Ferguson-Smith AC, Burton GJ. A quantitative analysis of transcriptionally active syncytiotrophoblast nuclei across human gestation. Journal of Anatomy [Internet]. 2011;219(5):601–10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3222839/pdf/joa0219-0601.pdf

180. Dieckmann L, Cruceanu C, Lahti-Pulkkinen M, Lahti J, Kvist T, Laivuori H, et al. Reference-Based Versus Reference-Free Cell Type Estimation In DNA Methylation Studies Using Human Placental Tissue [Internet]. 2021 Sep. Available from: https://www.researchsquare.com/article/rs-848651/v1

181. Salavati N, Smies M, Ganzevoort W, Charles AK, Erwich JJ, Plösch T, et al. The Possible Role of Placental Morphometry in the Detection of Fetal Growth Restriction. Frontiers in Physiology [Internet]. 2019 [cited 2021 Nov 10];9:1884. Available from: https://www.frontiersin.org/article/10.3389/fphys.2018.01884

182. Kong Y, Rastogi D, Seoighe C, Greally JM, Suzuki M. Insights from deconvolution of cell subtype proportions enhance the interpretation of functional genomic data. PLOS ONE [Internet]. 2019 Apr [cited 2021 Nov 10];14(4):e0215987. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0215987

183. Filipski K, Scherer M, Zeiner KN, Bucher A, Kleemann J, Jurmeister P, et al. DNA methylation-based prediction of response to immune checkpoint inhibition in metastatic melanoma. Journal for ImmunoTherapy of Cancer [Internet]. 2021 Jul [cited 2021 Nov 10];9(7):e002226. Available from: https://jitc.bmj.com/content/9/7/e002226

184. Koestler DC, Usset J, Christensen BC, Marsit CJ, Karagas MR, Kelsey KT, et al. DNA Methylation-Derived Neutrophil-to-Lymphocyte Ratio: An Epigenetic Tool to Explore Cancer Inflammation and Outcomes. Cancer Epidemiology and Prevention Biomarkers [Internet]. 2017 Mar [cited 2021 Nov 9];26(3):328–38. Available from: https://cebp.aacrjournals.org/content/26/3/328

185. Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, L. McKenney S, et al. DNA methylation of cord blood cell types: Applications for mixed cell birth studies. Epigenetics [Internet]. 2016;11:354–62. Available from: http://dx.doi.org/10.1080/15592294.2016.1161875

186. Goede OM de, Lavoie PM, Robinson WP. Cord blood hematopoietic cells from preterm infants display altered DNA methylation patterns. Clinical epigenetics [Internet]. 2017;9:39–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28428831

187. Cell type specific DNA methylation in cord blood : A 450K-reference data set and cell count-based validation of estimated cell type composition. Epigenetics [Internet]. 2016;11(9):690–8. Available from: http://dx.doi.org/10.1080/15592294.2016.1214782

188. Lin X, Tan JYL, Teh AL, Lim IY, Liew SJ, MacIsaac JL, et al. Cell type-specific DNA methylation in neonatal cord tissue and cord blood: A 850K-reference panel and comparison of cell types. Epigenetics [Internet]. 2018 Sep [cited 2021 Nov 9];13(9):941–58. Available from: https://doi.org/10.1080/15592294.2018.1522929

189. Morin AM, Gatev E, McEwen LM, MacIsaac JL, Lin DTS, Koen N, et al. Maternal blood contamination of collected cord blood can be identified using DNA methylation at three CpGs. Clinical Epigenetics. 2017;9:75.

190. Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, et al. Additional annotation enhances potential for biologically-relevant analysis of the illumina infinium HumanMethylation450 BeadChip array. Epigenetics & Chromatin [Internet]. 2013 Mar;6(1):4–4. Available from: http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/1756-8935-6-4

191. Caniggia I, Winter J, Lye SJ, Post M. Oxygen and placental development during the first trimester: implications for the pathophysiology of pre-eclampsia. Placenta. 2000 Apr;21 Suppl A:S25–30.

192. Huppertz B, Gauster M, Orendi K, König J, Moser G. Oxygen as modulator of trophoblast invasion. Journal of Anatomy [Internet]. 2009 Jul;215(1):14–20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2714634/

193. Huppertz B. Biology of preeclampsia: Combined actions of angiogenic factors, their receptors and placental proteins. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease [Internet]. 2020 Feb 1;1866(2):165349. Available from: https://www.sciencedirect.com/science/article/pii/S0925443918304861

194. Masuzaki H, Miura K, Yoshiura K, Yoshimura S, Niikawa N, Ishimaru T. Detection of cell free placental DNA in maternal plasma: direct evidence from three cases of confined placental mosaicism. Journal of Medical Genetics [Internet]. 2004 Apr 1;41(4):289–92. Available from: https://jmg.bmj.com/content/41/4/289

195. Taglauer ES, Wilkins-Haug L, Bianchi DW. Review: Cell-free fetal DNA in the maternal circulation as an indication of placental health and disease. Placenta [Internet]. 2014 Feb 1;35:S64–8. Available from: https://www.sciencedirect.com/science/article/pii/S0143400413008345

196. Turco MY, Gardner L, Kay RG, Hamilton RS, Prater M, Hollinshead MS, et al. Trophoblast organoids as a model for maternal–fetal interactions during human placentation. Nature [Internet]. 2018 Dec;564(7735):263–7. Available from: https://www.nature.com/articles/s41586-018-0753-3

197. Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, et al. Inferring genetic ancestry: Opportunities, challenges, and implications. The American Journal of Human Genetics [Internet]. 2010 May;86(5):661–73. Available from: https://www.sciencedirect.com/science/article/pii/S0002929710001552#bib38

198. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature [Internet]. 2018 Mar;555(7697):469–74. Available from: http://www.nature.com/doifinder/10.1038/nature26000

199. Andrews SV, Yang IJ, Froehlich K, Oskotsky T, Sirota M. Large-scale placenta DNA methylation mega-analysis reveals fetal sex-specific differentially methylated CpG sites and regions [Internet]. 2021 Mar p. 2021.03.04.433985. Available from: https://www.biorxiv.org/content/10.1101/2021.03.04.433985v2

200. Pollheimer J, Vondra S, Baltayeva J, Beristain AG, Knöfler M. Regulation of placental extravillous trophoblasts by the maternal uterine environment. Frontiers in Immunology [Internet]. 2018;9:2597. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2018.02597

201. Loukeris K, Sela R, Baergen RN. Syncytial knots as a reflection of placental maturity: Reference values for 20 to 40 weeks' gestational age. Pediatric and Developmental Pathology: The Official Journal of the Society for Pediatric Pathology and the Paediatric Pathology Society. 2010 Aug;13(4):305–9.

202. Thomas JR, Appios A, Zhao X, Dutkiewicz R, Donde M, Lee CYC, et al. Phenotypic and functional characterization of first-trimester human placental macrophages, hofbauer cells. Journal of Experimental Medicine [Internet]. 2020 Oct 19;218(1). Available from: https://doi.org/10.1084/jem.20200891

203. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. Nature [Internet]. 2018 Nov;563(7731):347. Available from: https://www.nature.com/articles/s41586-018-0698-6

204. Tsang JCH, Vong JSL, Ji L, Poon LCY, Jiang P, Lui KO, et al. Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2017 Sep;114(37):E7786–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28830992

205. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Bisulfite sequencing for assessing epigenetic heterogeneity. 2014;11(8):6–11.

206. Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J, et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science [Internet]. 2017 Aug [cited 2021 Nov 10];357(6351):600–4. Available from: https://www.science.org/doi/10.1126/science.aan3351

Appendices

Appendix A - Supplementary information for Chapter 2

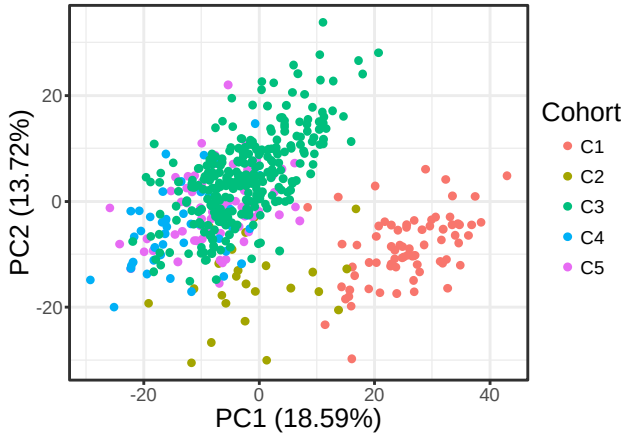

Figure A.1: Dataset-specific effects. PC1 by PC2 scatterplot from PCA computed on scaled and centered DNAme beta values from 499 samples and 319233 sites.

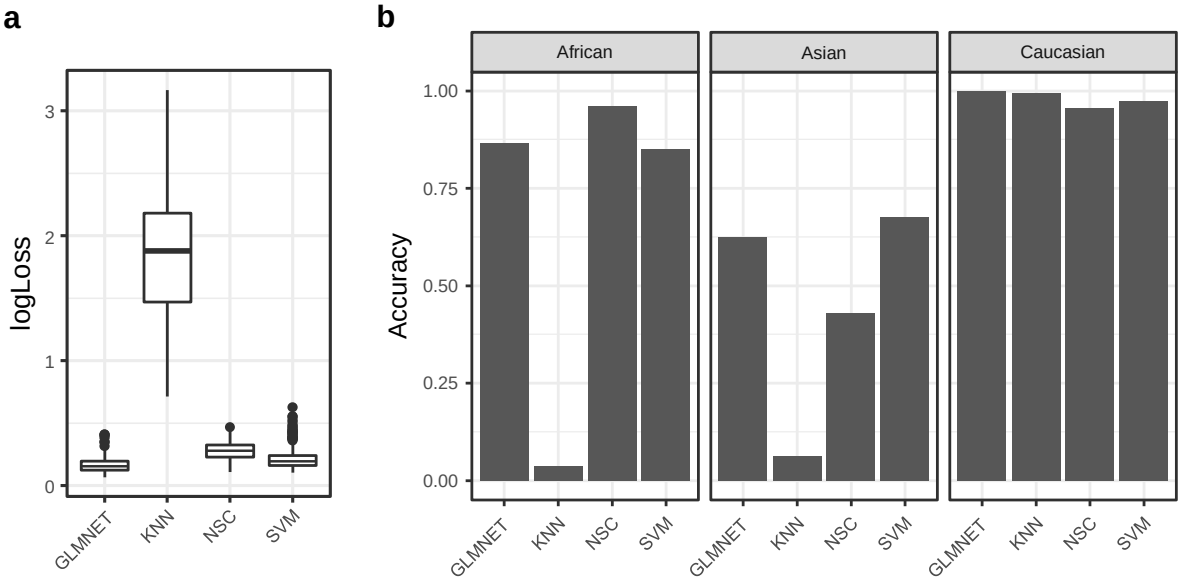

Figure A.2: Performance between machine learning algorithms in training. Resampling results for each machine learning algorithm. **a** performance (LogLoss) between machine learning algorithms in predicting ethnicity, and **b** class-specific accuracy.

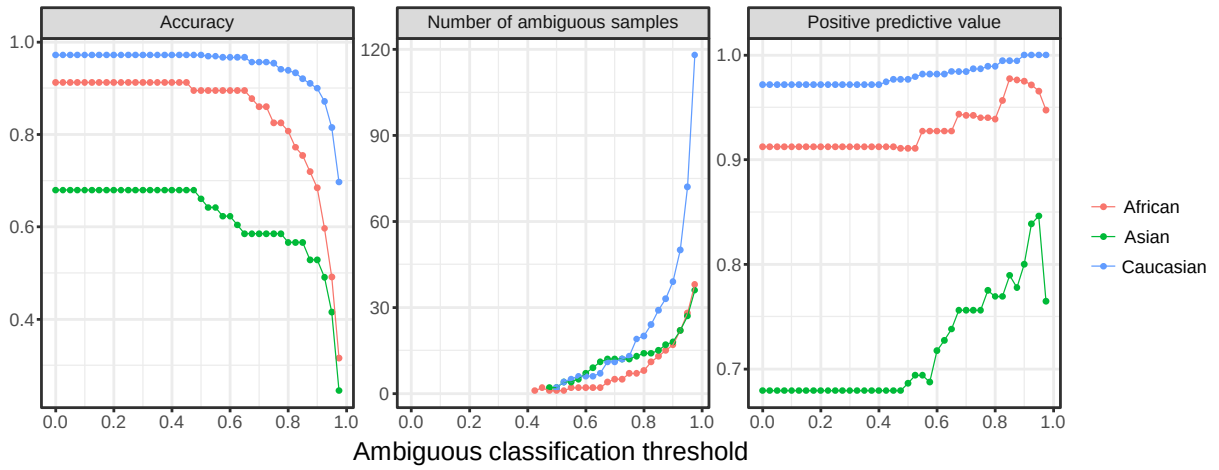

Figure A.3: Threshold analysis for determining "ambiguous" samples. Various cutoffs for predicted membership probabilities were compared with respect to changes in predictive performance.

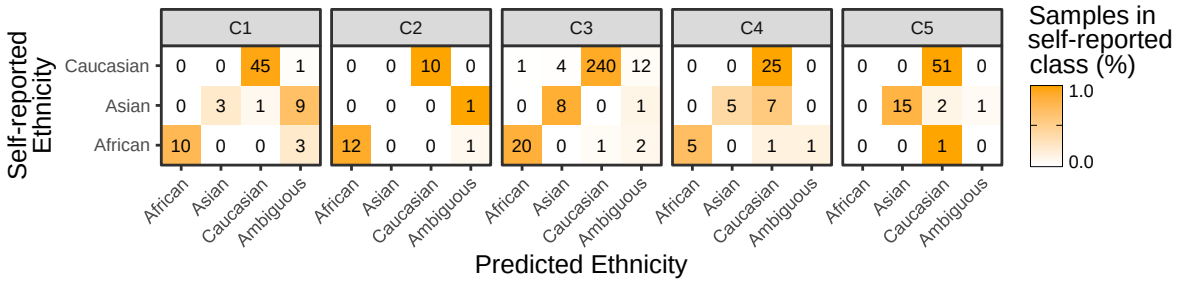

Figure A.4: Dataset-specific performance. PlaNET's classification performance was calculated for each dataset using a model trained to all other datasets.

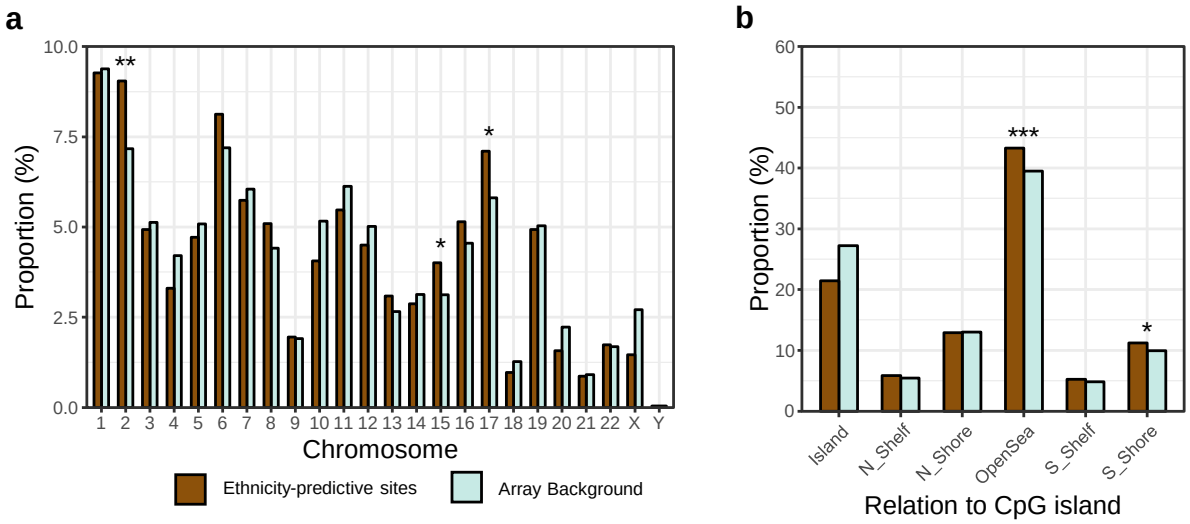

Figure A.5: Enrichment analysis on ethnicity-predictive HM450K sites. PlaNET's CpG sites used to predict ethnicity was tested for enrichment with respect to **a** chromosomal location, and **b** relation to CpG islands.

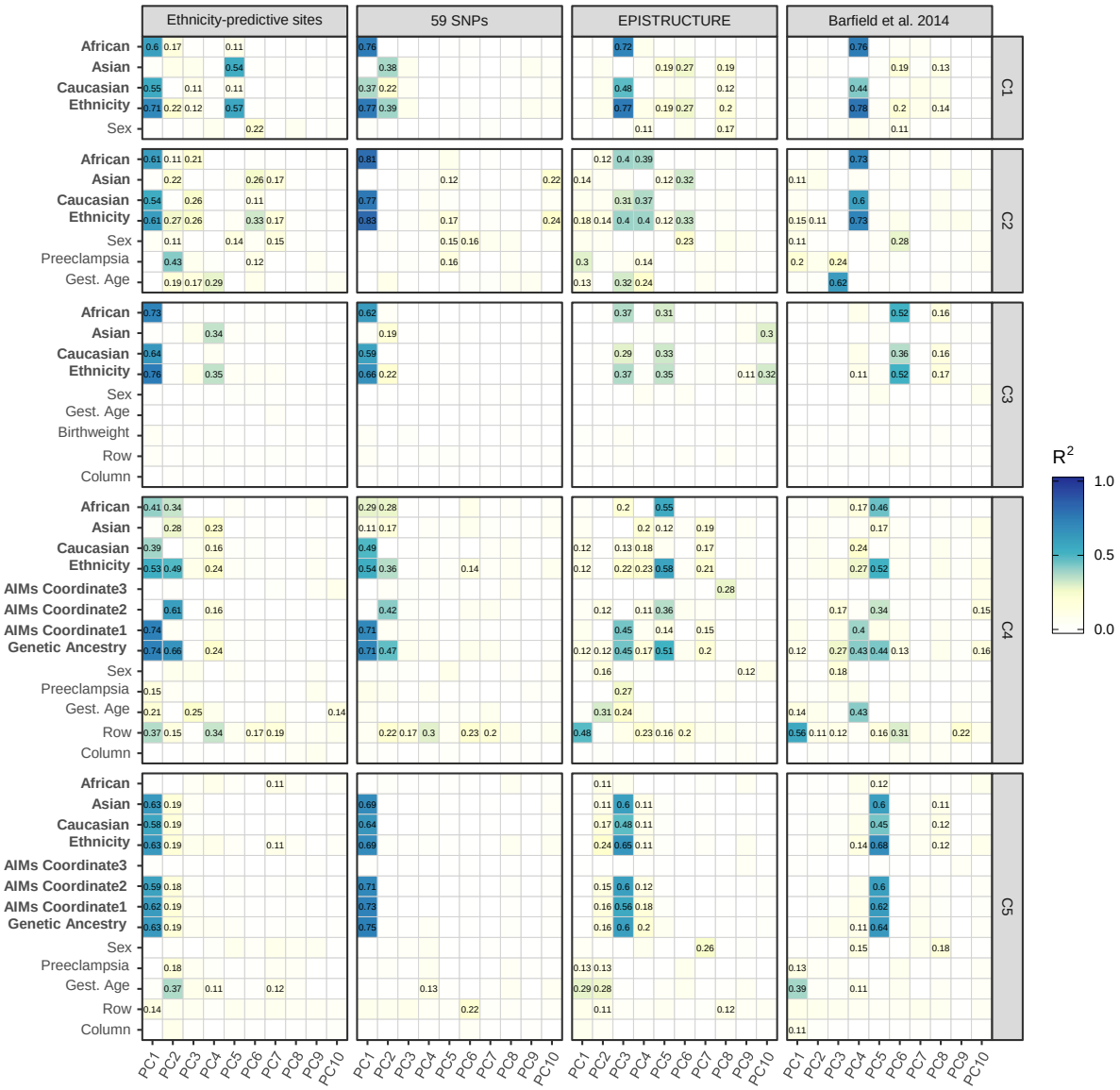

Figure A.6: Association of population structure PCs with technical and biological variables. PCs were computed on the following sets of HM450K sites: ethnicity-predictive sites, 59 SNPs, EPISTRUCTURE, and Barfield's method. Each PC was tested for their association with various cohort-specific technical and biological variables. For a given cohort (e.g. C1), ethnicity predictive sites from a classifier trained on all other cohorts (e.g. C2-C5) was used to avoid bias.

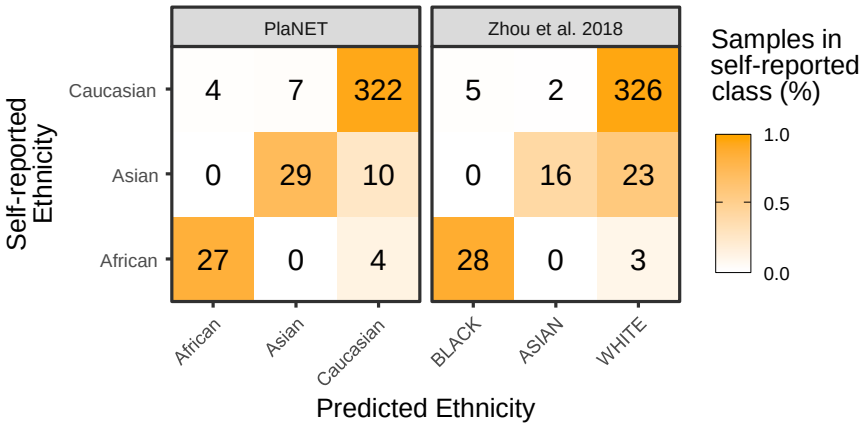

Figure A.7: PlaNET vs Zhou et al. 2017 snp-based classifier. PlaNET's ethnicity classification performance was compared to Zhou et al. 2017 (33) SNP-based ethnicity classifier in cohorts C3, C4, and C5.

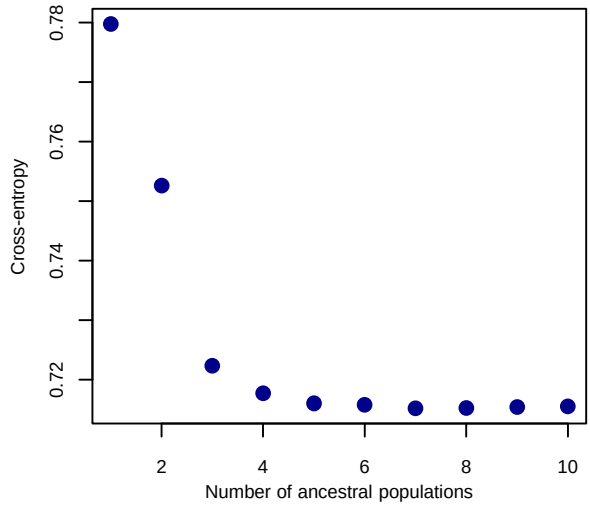

Figure A.8: Estimating k number of ancestral populations using in genetic admixture inference program LEA. The cross-entropy criterion was used to determine the number of ancestral populations for estimating genetic ancestry coefficients. The number of ancestral populations was chosen at the point k = 3, when the cross-entropy criterion decreases signficantly less with each integer-increase in k.

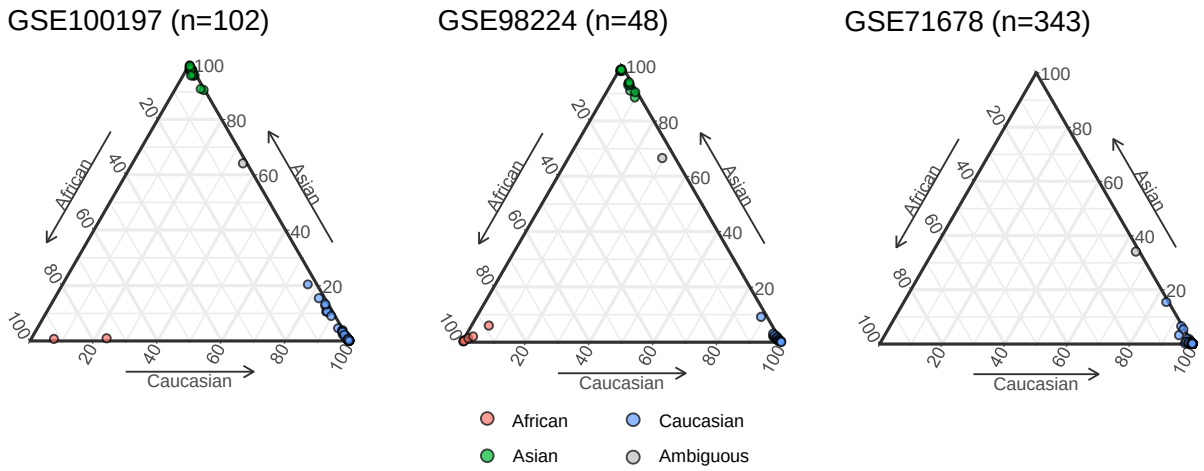

Figure A.9: Application of PlaNET to placental EWAS. Samples from three independent cohorts are plotted along three axes by their probability of belonging to each ethnicity class and colored by their final ethnicity classification determined by PlaNET.

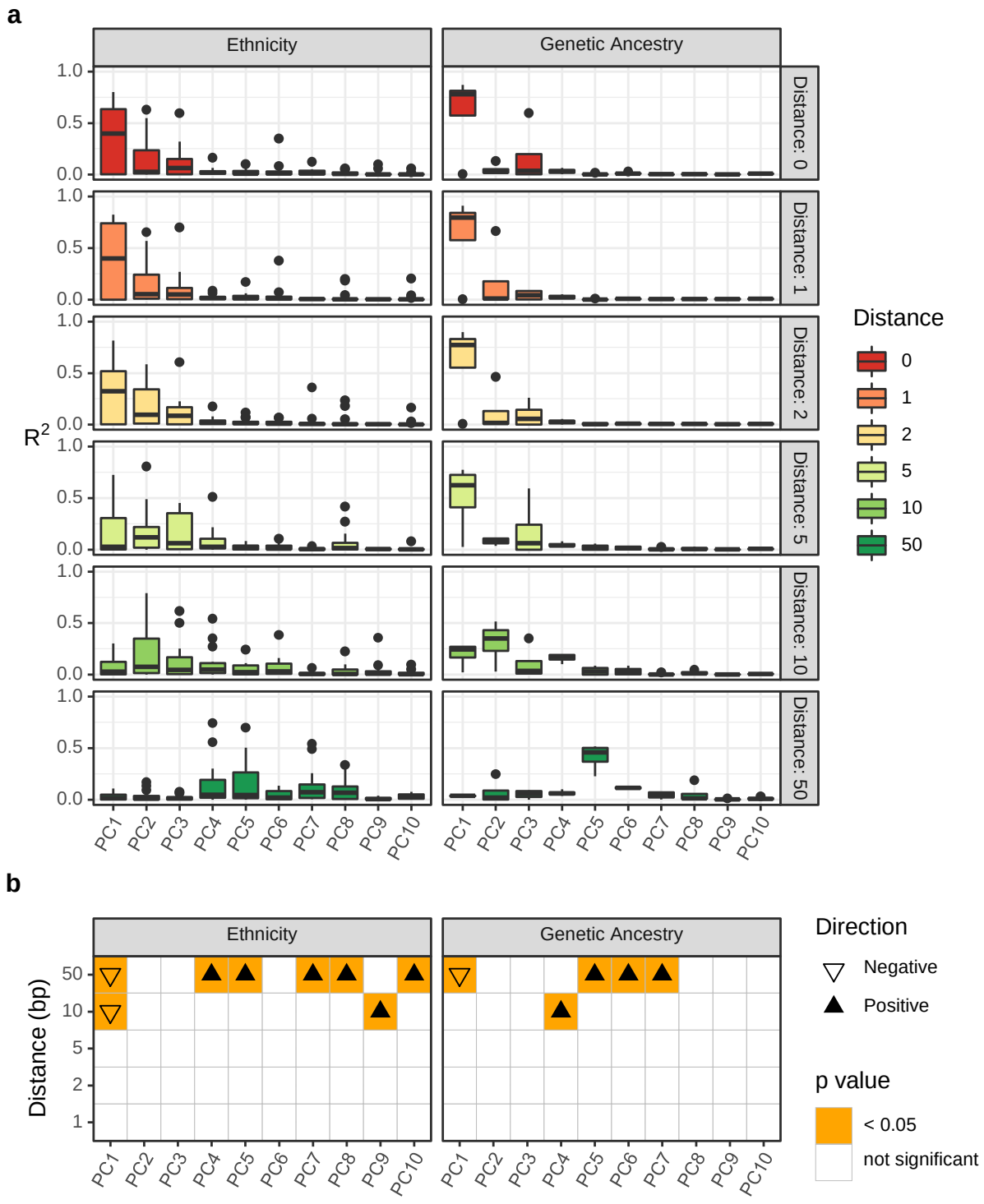

Figure A.10: Evaluation of Barfield's alternative location-based filtering approaches. The signal associated with ethnicity and genetic ancestry was measured in relation to the distance of which a genetic variant lies to a CpG site (0, 1, 2, 5, 10, 50 bp). **a** Amount of variance explained in PCi (i = 1, 2, 3, ..., 10) by either ethnicity or genetic ancestry. **b** Whether there was difference in the amount of ethnicity or genetic ancestry -associated variation in PCi, depending on distance to a genetic variant. Direction of association is indicated, where the reference group is the 0 bp set.

Table A.1: PlaNET's cohort-specific ethnicity classification performance. PlaNET's classification performance described across cohorts, assessed using LODOCV.

Cohort	Accuracy	Kappa
C1	0.8055556	0.6522939
C2	0.9166667	0.8500000
C3	0.9273356	0.7109036
C4	0.7954545	0.6102362
C5	0.9428571	0.8507463
C6	0.9259259	0.8125000

Table A.2: PlaNET's ethnicity classification performance described by class. PlaNET's classification performance described across ethnicity groups, assessed using LODOCV.

Predicted Ethnicity	Reported Ethnicity	Frequency	proportion of reported class	Cohort
African	African	10	0.7692308	C1
Asian	African	0	0.0000000	C1
Caucasian	African	0	0.0000000	C1
Ambiguous	African	3	0.2307692	C1
African	Asian	0	0.0000000	C1
Asian	Asian	3	0.2307692	C1
Caucasian	Asian	1	0.0769231	C1
Ambiguous	Asian	9	0.6923077	C1
African	Caucasian	0	0.0000000	C1
Asian	Caucasian	0	0.0000000	C1
Caucasian	Caucasian	45	0.9782609	C1
Ambiguous	Caucasian	1	0.0217391	C1
African	African	12	0.9230769	C2
Asian	African	0	0.0000000	C2
Caucasian	African	0	0.0000000	C2
Ambiguous	African	1	0.0769231	C2
African	Asian	0	0.0000000	C2

Asian	Asian	0	0.0000000	C2
Caucasian	Asian	0	0.0000000	C2
Ambiguous	Asian	1	1.0000000	C2
African	Caucasian	0	0.0000000	C2
Asian	Caucasian	0	0.0000000	C2
Caucasian	Caucasian	10	1.0000000	C2
Ambiguous	Caucasian	0	0.0000000	C2
African	African	20	0.8695652	C3
Asian	African	0	0.0000000	C3
Caucasian	African	1	0.0434783	C3
Ambiguous	African	2	0.0869565	C3
African	Asian	0	0.0000000	C3
Asian	Asian	8	0.8888889	C3
Caucasian	Asian	0	0.0000000	C3
Ambiguous	Asian	1	0.1111111	C3
African	Caucasian	1	0.0038911	C3
Asian	Caucasian	4	0.0155642	C3
Caucasian	Caucasian	240	0.9338521	C3
Ambiguous	Caucasian	12	0.0466926	C3
African	African	5	0.7142857	C4
Asian	African	0	0.0000000	C4
Caucasian	African	1	0.1428571	C4
Ambiguous	African	1	0.1428571	C4
African	Asian	0	0.0000000	C4
Asian	Asian	5	0.4166667	C4
Caucasian	Asian	7	0.5833333	C4
Ambiguous	Asian	0	0.0000000	C4
African	Caucasian	0	0.0000000	C4
Asian	Caucasian	0	0.0000000	C4
Caucasian	Caucasian	25	1.0000000	C4
Ambiguous	Caucasian	0	0.0000000	C4

African	African	0	0.0000000	C5
Asian	African	0	0.0000000	C5
Caucasian	African	1	1.0000000	C5
Ambiguous	African	0	0.0000000	C5
African	Asian	0	0.0000000	C5
Asian	Asian	15	0.8333333	C5
Caucasian	Asian	2	0.1111111	C5
Ambiguous	Asian	1	0.0555556	C5
African	Caucasian	0	0.0000000	C5
Asian	Caucasian	0	0.0000000	C5
Caucasian	Caucasian	51	1.0000000	C5
Ambiguous	Caucasian	0	0.0000000	C5
African	African	0	0.0000000	C6
Asian	African	0	0.0000000	C6
Caucasian	African	1	1.0000000	C6
Ambiguous	African	0	0.0000000	C6
African	Asian	0	0.0000000	C6
Asian	Asian	6	0.8571429	C6
Caucasian	Asian	1	0.1428571	C6
Ambiguous	Asian	0	0.0000000	C6
African	Caucasian	0	0.0000000	C6
Asian	Caucasian	0	0.0000000	C6
Caucasian	Caucasian	19	1.0000000	C6
Ambiguous	Caucasian	0	0.0000000	C6

Table A.3: PE linear modeling results while adjusting for predicted ethnicity. Results containing 651 PE-associated sites from linear modeling while adjusting for predicted ethnicity.

Closest_TSS_gene_name	cpg	Discovery Cohort p value	Validation Cohort p value
LAPTM4A	cg09394306	0.1743197	0.0960837
AX747766	cg17850498	0.1907809	0.1152059

ERGIC1	cg26813604	0.1733287	0.0754164
KRT15	cg26625897	0.2956765	0.1968648
ST3GAL1	cg00592695	0.2765183	0.1265308
INHBA	cg11079619	0.2106447	0.1121199
C10orf10	cg02334081	0.1225511	0.0747605
DENND2D	cg20317872	0.1762978	0.0943016
LIMCH1	cg03822934	0.1662338	0.0827407
TMEM242	cg05002580	0.1021112	0.0587353
FAM110A	cg16606561	0.1700186	0.0969453
TRIB3	cg15799353	0.1343169	0.0859857
PIGS	cg24569276	0.1131509	0.0563061
NCF4	cg08612539	0.1378323	0.1022450
ERN1	cg12779575	0.1880272	0.0841852
LOC400238	cg10108710	0.1661197	0.0980662
KIAA0182	cg16804825	0.1248011	0.0713625
CMIP	cg10246581	0.1763644	0.0670743
AMZ1	cg14605117	0.1859248	0.0877467
GDPD5	cg05521767	0.1709337	0.0999695
PIK3AP1	cg27479162	0.1408997	0.0659398
GALNT2	cg23677911	0.1686504	0.0941019
FAM105B	cg18190824	0.1659260	0.0926267
FHL2	cg13754437	0.2136664	0.0786833
SYDE1	cg18584265	0.1518324	0.0801390
PHLDA3	cg07950244	0.1906623	0.1125482
NDRG1	cg14143441	0.2069285	0.1425393
FLNB	cg23730027	0.1756390	0.0953749
STARD13	cg26651514	0.2002187	0.1195384
FLJ32224	cg26185836	0.1068574	0.0707043
C9orf46	cg14025883	0.1323388	0.0846568
IL1R2	cg24617203	0.1554369	0.0859868
SLCO2A1	cg15765546	0.1371885	0.0833081

ZNF395	cg01713086	0.1401826	0.0866359
TGM1	cg23696550	0.1451398	0.0751740
NCOR2	cg21626573	0.1356429	0.0738186
LOC644215	cg04869006	0.1105658	0.0592862
SNORA11B	cg05131266	0.1560187	0.0736005
CEBPA	cg22526990	0.1376826	0.0859620
MORC3	cg00622655	0.1657418	0.0700953
NCOR2	cg20352351	0.1902014	0.1056166
PHYHIP	cg26509870	0.1873843	0.1208976
CYP11A1	cg15903956	0.1813276	0.0988874
LINC00310	cg25032603	0.1625371	0.0752787
LOC100505839	cg19140548	0.1902205	0.1277972
Mir_548	cg26986443	0.1396634	0.0507085
LINC00284	cg05452692	0.1685425	0.0841893
IER5L	cg13908988	0.1267195	0.0664032
MIR3150A	cg07158065	0.1561546	0.0968439
HK1	cg15258080	0.1479864	0.0923529
DUSP16	cg15429134	0.1624628	0.0985726
PAPPA2	cg10994126	0.1649658	0.1056620
GRIP2	cg10586672	0.1596497	0.0827229
FN1	cg12436772	0.2697080	0.1578415
ST3GAL1	cg00736681	0.1648769	0.0813991
BCL6	cg17394304	0.1405970	0.0734435
LOC100507582	cg01180628	0.1624389	0.0813336
MSH4	cg17966362	0.1453088	0.0867641
RRBP1	cg12632411	0.1767792	0.0985087
KRT80	cg00822797	0.1412933	0.0788241
PARD6B	cg10187713	0.1898356	0.0939787
MLL5	cg07351322	0.1455964	0.0946419
KRT39	cg17464043	0.1556423	0.0830882
CORO1C	cg09182455	0.1722001	0.0982596

BET1	cg16568084	0.1527294	0.0772224
HSD3B1	cg16175792	0.1496510	0.1083410
C1orf98	cg19694404	0.1479230	0.0720177
POLE4	cg13547665	0.1499997	0.0659762
VIT	cg26638266	0.1079296	0.0500626
FLNB	cg24204017	0.1424196	0.0823461
SDC1	cg22344841	0.1310884	0.0725977
CALML3	cg17115419	0.1877563	0.1028913
SMARCB1	cg01912455	0.1285538	0.0800784
DUSP1	cg02029908	0.1668786	0.0789647
PDLIM2	cg26366616	0.1339218	0.0960141
EPAS1	cg08900316	0.1442795	0.0710788
LOC100507582	cg20971407	0.2013550	0.1149395
FOS	cg10565512	0.1153105	0.0764266
TANK	cg22060367	0.1183919	0.0906283
ALAD	cg01257194	0.1101688	0.0517497
STX1A	cg20663219	0.1404756	0.0727859
ZBTB4	cg07168214	0.1271096	0.0655327
AMN1	cg08198187	0.1190541	0.0874172
ARHGEF4	cg25500616	0.1444241	0.1063410
CMIP	cg02547035	0.1238429	0.0588209
FLJ32224	cg01924561	0.1713013	0.1021101
ZNF385A	cg09676376	0.1340021	0.0827121
SRPRB	cg16275903	0.1043673	0.0567156
VILL	cg06641593	0.1150055	0.0828387
ANO6	cg25162927	0.1334856	0.0862408
GRHL1	cg21560697	0.1232018	0.0707512
HSD17B8	cg17066452	0.1179542	0.0517812
VGLL4	cg22278433	0.1070755	0.0695603
DUSP1	cg04577249	0.1269161	0.0748897
DLG5	cg27328839	0.1528907	0.1086311

IFFO2	cg13173305	0.1100890	0.0563851
C1orf65	cg12458966	0.1053474	0.0564588
CTRC	cg17843487	0.1564664	0.0796085
TRNA_Arg	cg21545548	0.1403227	0.1288812
ZNF385A	cg12150931	0.1157498	0.0689389
PPFIA1	cg18564881	0.1338774	0.0778110
ADAM12	cg02494582	0.1470507	0.0923198
ZP3	cg00502662	0.1044143	0.0713550
AL109706	cg12290217	0.1206158	0.0435867
ITPRIP	cg16301004	0.1616608	0.1133359
TBC1D1	cg01723031	0.1454914	0.0732966
KAZALD1	cg02448743	0.1098813	0.0641338
LOC285954	cg14704980	0.1874618	0.0897959
IL22RA2	cg00415333	0.1563723	0.0759062
AMZ1	cg03653726	0.1594232	0.1211834
TNFSF18	cg22626683	0.1609253	0.0873861
MAFK	cg23843484	0.1614536	0.0821419
SLC45A1	cg11283860	0.1483595	0.0786322
DKFZp547K2416	cg05364179	0.1399604	0.0531375
FLNB	cg23812679	0.1127883	0.0739406
A2ML1	cg03490200	0.1056508	0.0643527
KRT80	cg27278470	0.1479152	0.1009707
BCL6	cg05663031	0.1226802	0.0660343
LOC100507091	cg05399718	0.1048972	0.0665540
GAS7	cg25379762	0.1019104	0.0519777
TIMP3	cg25245338	0.1926282	0.0896477
SSFA2	cg07835482	0.1034387	0.0719331
FLNB	cg02770406	0.2045698	0.1116934
DHX32	cg23997887	0.1093727	0.0789230
SIPA1L2	cg01432692	0.1028119	0.0577209
NPAT	cg12892243	0.1179568	0.0816362

KRT86	cg22193385	0.1870489	0.1002310
ACOX3	cg17055207	0.1023086	0.0716877
IRF8	cg09684264	0.1048425	0.0470659
PSG8	cg11387248	0.1220635	0.0690789
LIMS2	cg13795819	0.1212842	0.0822778
FAM160B2	cg20576064	0.1544276	0.0862692
NCOR2	cg09676622	0.1037704	0.0583922
NPB	cg21823502	0.1234261	0.0754399
TINAGL1	cg16103203	0.1039215	0.0796571
CLEC2L	cg01581050	0.1088211	0.0596923
SLC11A1	cg17010118	0.1436974	0.1029795
FAM18A	cg03777414	0.2031475	0.0892948
CLDN7	cg13724311	0.1086586	0.0657885
PAPPA2	cg25103772	0.1571566	0.0864519
MB21D2	cg20970886	0.1321369	0.0772116
BUB1B	cg25653839	0.1159770	0.0723955
KDELC1	cg06377626	0.1089064	0.0565417
CCL27	cg13562353	0.1279251	0.0659571
KIAA1614	cg13467459	0.1375105	0.1133960
SNORD54	cg02508743	0.1059206	0.0544641
LNPEP	cg24598187	0.1314568	0.0730223
PWWP2B	cg25961733	0.1118772	0.0762420
JHDM1D	cg26800802	0.1316227	0.0536699
SFT2D3	cg24925163	0.1545132	0.0894288
MYO7A	cg15433043	0.1172271	0.0648499
MBNL2	cg03099780	0.1377552	0.0849551
LOC90246	cg19234171	0.1176429	0.0731066
EPAS1	cg15129144	0.1367486	0.0952318
KIAA1211	cg08112737	0.1173152	0.0662634
TCF25	cg05412696	0.1003144	0.0708257
KRT86	cg05169499	0.1126673	0.0626144

FNIP2	cg11637968	0.1465957	0.0888873
C2	cg11049439	0.1048472	0.0747960
GATA3	cg01522692	0.1014322	0.0584117
CLN8	cg19366147	0.1027786	0.0435880
SAP30BP	cg20938708	0.1313375	0.0559659
AMBRA1	cg07819010	0.1266005	0.0710908
PAPPA2	cg18236464	0.2087063	0.1115184
CSGALNACT1	cg14854503	0.1949079	0.0894706
MSI2	cg04573500	0.1241051	0.0794701
DQX1	cg02034222	0.1007056	0.0654533
SOX6	cg19578175	0.1241422	0.0556661
TEAD3	cg10893014	0.2445801	0.1454212
MARCKS	cg16090790	0.1539983	0.0994383
MIR4499	cg23485627	0.1213576	0.0873941
GALNTL4	cg12511310	0.1067030	0.0547135
SPTLC3	cg16404259	0.1167299	0.0858556
AX747408	cg24233594	0.1047879	0.0439804
C22orf31	cg20080983	0.1187853	0.0675299
INSIG1	cg12979992	0.1624063	0.1184165
NFKBIZ	cg06056170	0.1052324	0.0858157
KLF5	cg26531076	0.1179189	0.1060172
YPEL5	cg02766770	0.1194489	0.0843706
CD47	cg23521980	0.1015381	0.0604525
FLJ43663	cg14343652	0.1115396	0.0571380
ZNF783	cg18783886	0.1100744	0.0815151
RRBP1	cg00990977	0.1438861	0.1005321
SLC19A3	cg04730276	0.1195528	0.0412091
ARID3A	cg25298189	0.1026607	0.0811279
PLEC	cg08161931	0.1011361	0.0633463
XIRP1	cg01127412	0.1121725	0.0681052
MFHAS1	cg12077460	0.1106951	0.0587861

DNASE1L3	cg24260359	0.1071034	0.0486690
U6	cg27378762	0.1051074	0.0582850
HTR1D	cg11811391	0.1598232	0.0762471
ZBTB38	cg14500070	0.1330913	0.0811329
STK24	cg19861486	0.1302341	0.0800659
ERICH1	cg13175060	0.1541380	0.0967391
PHF17	cg03078141	0.1103623	0.0742171
LIMCH1	cg25740652	0.1558911	0.0725388
EGFR	cg04156940	0.1171728	0.0522893
C14orf1	cg01284448	0.1428878	0.0583578
C10orf26	cg20340720	0.1288316	0.0834984
C2orf61	cg14327359	0.1342867	0.0601988
TNFAIP1	cg01257345	0.1438181	0.0723548
CSF3R	cg07285167	0.1013814	0.0722714
LTF	cg13672136	0.1304416	0.1023745
C11orf16	cg14402562	0.1105523	0.0823094
SNORA70	cg25103160	0.1107223	0.0272466
GLDN	cg07888040	0.1140518	0.0557577
FLNB	cg17338821	0.1003480	0.0748550
NEBL	cg00496126	0.1525362	0.0971906
ART4	cg10047173	0.1500502	0.0790102
IGF2BP2	cg03554286	0.1063125	0.0683139
CMIP	cg08946161	0.1034375	0.0639131
EPAS1	cg25589945	0.1312678	0.0929107
TMEM139	cg08261841	0.1145982	0.0821436
INHBA	cg18413237	0.1240939	0.0858612
SIAH2	cg21331845	0.1231915	0.0671214
MIR205	cg01334432	0.1042360	0.0573961
LOC285696	cg25503410	0.1186930	0.0675671
SLC30A2	cg26922451	0.1318192	0.0821282
POLD3	cg00453717	0.1205471	0.0726315

PWWP2B	cg26450254	0.1217395	0.0819728
CLIP4	cg17417693	0.1519117	0.0976928
SLC11A2	cg25493658	0.1297838	0.0964564
CORO1C	cg12647920	0.1681368	0.1266417
NAGLU	cg01515741	0.1136608	0.0917036
PLAC4	cg14867395	0.1194561	0.0766978
LOC728743	cg11034318	0.1301698	0.0575724
JUNB	cg22996170	0.2065297	0.1331657
UBR5	cg12517050	0.1037187	0.0684140
GATA3	cg07989490	0.1053503	0.0486932
GUCA2A	cg09278187	0.1085280	0.0617985
SNORA11B	cg21906866	0.1267400	0.0783527
C9orf46	cg23598352	0.1054446	0.0586757
LMNA	cg27182012	0.1714848	0.1257018
TCF25	cg07891440	0.1201064	0.1018249
ETV5	cg01519765	0.1264343	0.0733472
NCK2	cg03105244	0.1001367	0.0686878
DNASE1L3	cg10235741	0.1219790	0.0781700
RALGDS	cg14103123	0.1227093	0.0953999
C14orf181	cg20016914	0.1047761	0.0659774
MIR4284	cg06480942	0.1438235	0.1014500
NR3C1	cg23400056	0.1129706	0.0605575
CSRNP1	cg13062627	0.1481665	0.0934153
TRIM8	cg17782974	-0.1013934	-0.0604044
AK122764	cg18743287	0.1557075	0.0900433
XYLT1	cg00840341	0.1048457	0.0523269
LGALS8	cg07913153	0.1022334	0.0667470
GPR110	cg22572071	0.1280648	0.0641043
AK123450	cg26548682	0.1163168	0.0839074
CHSY1	cg12361046	0.1584981	0.1083621
BTG3	cg08875503	0.1203667	0.0621265

C10orf90	cg04507071	0.1045761	0.0577496
C8orf42	cg02341578	0.1222670	0.0837226
TECR	cg24143196	0.1089459	0.0822994
INHBA	cg12261055	0.1537256	0.1068337
CYP11A1	cg06285340	0.1032612	0.0721988
PMEL	cg08869883	0.1050266	0.0646207
PHF17	cg17233452	0.1080484	0.0736370
AX747368	cg11624780	0.1082280	0.0659777
ASAP1	cg00659129	0.1040518	0.0475575
IFITM3	cg20151221	-0.1173911	-0.0651992
RERE	cg27561786	0.1004465	0.0745782
CAMSAP3	cg15512156	0.1266321	0.1006686
PGRMC2	cg21491609	0.1008235	0.0473572
TRAM2	cg26301143	0.1140431	0.0467347
TRIM29	cg26247168	0.1289521	0.1017392
TTC7A	cg02286857	0.1055675	0.0667863
MAB21L3	cg24311182	0.1093634	0.0743636
CAB39L	cg01427300	0.1005192	0.0727096
HIST1H1T	cg19722391	0.1363634	0.1102061
TIMP3	cg27221424	0.1102110	0.0805264
U6	cg02275040	0.1131576	0.0824383
RDH13	cg20669049	0.1114053	0.1043323
BC035370	cg27307465	0.1488085	0.0923239
C15orf52	cg12732548	0.1052788	0.0390361
LOC145837	cg25175240	0.1307522	0.0620420
FLNB	cg02026180	0.1115888	0.0645682
CGA	cg07981495	0.1528186	0.0807794
CALD1	cg03188976	0.1156296	0.0515970
KLHL29	cg00886182	0.1030936	0.0627003
SLC4A3	cg18199208	0.1095529	0.0654031
LOC100131551	cg00490976	0.1023943	0.0181881

SOD3	cg22183373	0.1009750	0.0445267
FAM150B	cg13467628	0.1306837	0.0606305
MOB3A	cg07381806	0.1296334	0.0785332
LOC100505839	cg03880642	0.1193863	0.0696841
ZNF175	cg10668363	0.1777645	0.1330086
PPP2CA	cg18514949	0.1128646	0.0556066
SLC7A5	cg06665333	0.1254069	0.0837035
HS3ST3A1	cg09174601	0.1235391	0.0551345
MED13L	cg07060505	0.1376905	0.0749720
ITCH	cg09049982	0.1196241	0.0777704
MIR149	cg11479811	0.1153183	0.0723402
GNA12	cg08799766	0.1541177	0.0789335
SLC45A4	cg20555854	0.1209705	0.1032588
CLMN	cg16336066	0.1205331	0.0580211
BCL6	cg06070445	0.1239877	0.0734827
TRIM8	cg16654458	0.1288605	0.1083717
C2orf72	cg05647720	0.1035748	0.0670920
HIST3H3	cg04103490	0.1298763	0.0789691
COL17A1	cg13553455	0.1368860	0.1058703
GUCA2A	cg03730249	0.1282421	0.0735506
UBE2E1	cg03668982	0.1441141	0.0880851
GALM	cg22860917	0.1373979	0.0789049
AK125516	cg01144019	0.1160873	0.0824130
DL489896	cg18474072	0.1503834	0.1216858
DGCR8	cg20012247	0.1378600	0.0967745
MIR4711	cg19773937	0.1005233	0.0648673
LIMCH1	cg22204103	0.1157463	0.0600224
EPS8L2	cg27649971	0.1350507	0.0508469
FSTL1	cg11622516	0.1319501	0.0756418
CEP41	cg09507697	0.1116043	0.0608925
ERRFI1	cg14178899	0.1306096	0.0766995

CSF1R	cg07284261	0.1292712	0.0867693
LOC100507582	cg26269881	0.1014920	0.0782483
CPPED1	cg01323840	0.1151815	0.0604709
BC048114	cg10092779	0.1123170	0.0279655
PITRM1	cg18912160	-0.1232990	-0.0578447
TBXAS1	cg24431161	0.1699414	0.0791880
MIG7	cg15113123	0.1081305	0.0954038
RAD1	cg15003812	0.1078510	0.1043787
ZNF366	cg12508655	0.1211084	0.0769666
PWWP2B	cg05467828	0.1028622	0.0598790
PTPN3	cg13715502	0.1026235	0.0707956
RAB5C	cg20438445	0.1059765	0.0641981
SH3BP5	cg18444702	0.1256924	0.1079884
PNMA2	cg01862311	0.1212113	0.0522759
GAB1	cg12379954	0.1000227	0.0723453
CREB1	cg10440877	0.1016099	0.0492983
LIMCH1	cg23653457	0.1292172	0.0620651
SMOC2	cg02448805	0.1221783	0.0718814
TBCD	cg14605961	0.1148927	0.1179104
NRN1	cg05254646	0.1044441	0.0678944
RRM2B	cg20157339	0.1032087	0.0480085
BC044741	cg23751171	0.1121906	0.0867376
FLT4	cg10660844	0.1032464	0.0610380
MRVI1	cg00510149	0.1108835	0.0699342
LOC646324	cg20428989	0.1174069	0.0852193
EFCAB1	cg16100845	0.1035427	0.0769750
CUX1	cg05910443	-0.1016964	-0.0559258
MIR4708	cg01703196	0.1268545	0.1154066
MTA1	cg14377923	0.1053661	0.0806536
C17orf110	cg11619216	0.1018524	0.0817162
TBCD	cg24481782	0.1447773	0.0348348

Mir_633	cg04541146	0.1036194	0.0674555
PAQR8	cg07907670	0.1011613	0.0684461
RILPL1	cg23825057	0.1046165	0.0677902
CAMK1	cg20782816	0.1067637	0.0450816
LRP5	cg16365302	0.1079796	0.0637344
TNFAIP8L2-SCNM1	cg16565154	0.1117180	0.0508165
FRMD6	cg22800400	0.1041551	0.0745392
HECW1	cg03873694	0.1001198	0.0299642
PWWP2B	cg23249922	0.1076791	0.0674789
CSGALNACT1	cg11155735	0.1102377	0.0558490
BTBD3	cg20981848	-0.1097683	-0.0570965
USHBP1	cg14584702	0.1022332	0.1078845
FBLN1	cg05248804	0.1025982	0.0646462
ASAP1	cg13629652	0.1126874	0.0595073
NACC2	cg14350701	0.1277542	0.0664427
SLC16A12	cg26568031	0.1051765	0.0588520
TBC1D9	cg03407966	0.1103353	0.0809053
MAN1C1	cg12701302	0.1443433	0.0945245
SH3BP5	cg04858987	0.1442992	0.1328360
SCIN	cg18695259	0.1142665	0.0702054
AK056252	cg04365699	0.1000710	0.0677121
GNA12	cg04849508	0.1866925	0.0854574
5S_rRNA	cg11033588	0.1072733	0.0798587
FJX1	cg21913652	0.1314089	0.0787613
MGAT3	cg05541460	0.1024101	0.0628478
TAP2	cg03438552	0.1057367	0.0680240
DLC1	cg26244164	-0.1027766	-0.0447400
AK091866	cg05211068	0.1035508	0.0651143
ATP5G3	cg20497304	0.1188227	0.0491627
APOL4	cg11178302	0.1489536	0.1198913
PHC2	cg06953325	0.1378397	0.0885950

MIR3182	cg01750200	0.1047989	0.0635478
DIAPH3	cg00623826	0.1111177	0.0291128
ANKDD1A	cg18926409	0.1010817	0.0676463
LOC100288428	cg21908673	0.1048143	0.0823316
NRXN2	cg21660452	0.1200305	0.1055005
ZFP36L1	cg06617636	0.1215385	0.0943922
IMPG1	cg16423738	0.1076020	0.0459714
TRNA_Leu	cg20430841	0.1075810	0.0650972
C6orf186	cg17295389	0.1031338	0.0387928
MIR938	cg07437737	0.1166480	0.1107562
FBN2	cg02252421	0.1203530	0.0894573
BAIAP2L1	cg15059474	0.1068322	0.0740438
WDR69	cg09638264	0.1206870	0.1296782
SCOC	cg05347925	0.1010942	0.0958568
COL17A1	cg15715892	0.1232077	0.0984267
5S_rRNA	cg16710042	0.1033686	0.0524693
ADHFE1	cg10895168	0.1009293	0.0675347
AK127270	cg27547053	0.1191767	0.0564561
FAP	cg03506656	0.1053281	0.0609395
LOC100130275	cg00110654	-0.1129978	-0.0482570
PTPN14	cg06826449	0.1140376	0.0824830
PCAT1	cg25632577	0.1180590	0.0625875
PKP2	cg19677302	0.1113568	0.0762706
ERRFI1	cg00768179	0.1003417	0.0865060
TMEM184A	cg21368161	0.1084437	0.0797038
GTDC1	cg03678729	0.1002414	0.0727596
QSOX1	cg27056501	0.1232767	0.0804161
MIR29A	cg11370011	0.1012887	0.0327260
INHBA	cg00159987	0.1047375	0.0868566
RNF126P1	cg08805241	0.1011680	0.0592222
EIF4H	cg26311610	0.1053510	0.0872803

DUSP5	cg23608075	0.1135498	0.0928904
SERINC5	cg26075213	0.1025161	0.0973443
STK38L	cg03766264	0.1029616	0.0679093
EGFR	cg23757825	0.1109838	0.0768403
LOC100507582	cg16582517	0.1297090	0.0665812
SIX3	cg02657654	-0.1536524	-0.0434080
NUDT9P1	cg26276667	0.1221499	0.0494058
LOC100130987	cg11758697	0.1255752	0.0790006
LOC729799	cg20444525	0.1171663	0.0172050
AK097119	cg07502782	0.1261420	0.0831680
FAM196A	cg11503487	0.1039182	0.0642767
GPD1L	cg15318697	0.1113799	0.0517759
PKIG	cg19554235	0.1200539	0.0781402
TBCD	cg16518729	0.1489467	0.0229759
7SK	cg05510339	0.1045678	0.0541506
CTAGE1	cg20586124	0.1274574	0.0947300
HDAC4	cg27433031	0.1088437	0.0778873
TFAP2A	cg14993900	0.1085764	0.0810844
GNA12	cg13928313	0.1010088	0.0671306
U2	cg18751231	0.1298065	0.1414841
DNAJC6	cg18191867	0.1054921	0.0901807
IL12A	cg26187205	0.1004155	0.0702951
FMNL2	cg13923497	0.1059254	0.0708492
CUX1	cg14466759	0.1281262	0.0988289
ST3GAL4	cg12804791	0.1426295	0.0847261
BC048982	cg26947626	-0.1022112	-0.0057109
LINC00163	cg11327657	0.1493850	0.0763327
SLC39A10	cg05418915	0.1044852	0.0569594
TSPAN5	cg02304751	0.1005088	0.0582228
SAPCD1	cg10158997	0.1186199	0.0886067
DEPDC1B	cg12995421	0.1251520	0.0748640

ANKRD52	cg20059881	0.1084219	0.0753146
VNN2	cg13130224	0.1001317	0.0598166
LINC00085	cg22278033	0.1043312	0.0450697
C4orf26	cg10409560	0.1041299	0.0994824
PCDHB13	cg03631455	-0.1106805	-0.0003045
BC047484	cg21594328	0.1080598	0.0525451
RGS4	cg07835293	0.1019896	0.0629711
ANKRD13A	cg02927618	0.1011876	0.0771450
ZNF727	cg01176516	-0.1303795	-0.0612574
IDH2	cg19674091	0.1563880	0.1115881
PXDN	cg26137290	0.1016939	0.0559345
EDNRB	cg23648516	0.1018177	0.0752002
GNA12	cg01839603	0.1328468	0.0713629
GPC6	cg26530275	0.1033434	0.0227496
ABLIM2	cg07296849	0.1023259	0.0649949
EPHA1	cg05385805	0.1039086	0.0673299
LIFR	cg08219241	0.1056169	0.1149519
Mir_633	cg15015143	0.1018028	0.0588296
EXTL2	cg04098985	0.1269597	0.1114514
PDPN	cg24671344	-0.1186792	-0.0539037
AK125212	cg15448894	0.1011026	0.0718488
IQCE	cg10688297	0.1108185	0.0810201
AK098012	cg27543214	-0.1161200	-0.0344788
FRZB	cg02625481	0.1049542	0.0602340
DDR1	cg14279856	0.1150937	0.1008434
RBPMS	cg07575193	-0.1080789	-0.0223641
MTUS1	cg05799058	0.1014948	0.0788375
ENPP2	cg01874183	0.1138941	0.0600458
SHFM1	cg01534613	0.1071936	0.0889347
CCDC140	cg25596297	-0.1139630	-0.0495734
SIX3	cg22884656	-0.1545105	-0.0781166

FOXP1	cg20891481	-0.1000147	-0.0739368
LRRFIP1	cg21708130	0.1382708	0.0836199
ADAMTS6	cg21033632	0.1092575	0.1055919
HOXD13	cg07802350	-0.1509524	-0.0880126
MEIS2	cg09205538	-0.1060544	-0.0084153
FGF2	cg07095252	0.1057079	0.0527533
KRT7	cg18689240	0.1084419	0.0487002
SFSWAP	cg02495552	0.1102770	0.0832737
AK055631	cg03326762	-0.1122530	-0.0457710
PIM1	cg21950107	0.1132253	0.0890689
INHBA	cg19323059	0.1088729	0.0938924
ETS1	cg21121082	0.1009837	0.0459503
EPHA2	cg05797770	0.1096481	0.0813233
FOXB1	cg00970361	-0.1200376	-0.0101300
BTBD3	cg15582126	0.1074323	0.0951913
TSPAN8	cg04950931	0.1109606	0.0330566
FAM150B	cg04709035	-0.1394304	-0.0664550
GPR39	cg24964130	0.1095576	0.0277418
PTPRK	cg20639396	-0.1252792	-0.0161954
RASA3	cg00458564	0.1096073	0.0672541
NUP160	cg16282339	0.1514743	0.0903178
MLLT1	cg06633438	-0.1039302	-0.0688399
GJA3	cg21554895	0.1068586	0.0994095
AP4E1	cg24311373	0.1268986	0.1395025
FOXA1	cg11760593	-0.1104974	-0.0486145
CACNA1C	cg13939602	-0.1086387	-0.1403057
WIBG	cg08338281	0.1136902	0.0874294
OLIG3	cg12744820	-0.1402366	-0.0207167
CD200	cg13671536	0.1049106	0.1008612
HOXD3	cg05864326	-0.1360797	-0.0363825
HDAC4	cg17939889	-0.1104113	-0.0761938

MIR4750	cg05213896	0.1081023	0.0898183
HES1	cg22305268	0.1500393	0.1213995
CCDC81	cg01663953	0.1083143	0.0681647
PML	cg01947066	-0.1061350	-0.0552786
LOC100130872	cg17227257	-0.1200439	-0.0799425
SLC39A1	cg25064552	0.1109112	0.0984771
MTMR10	cg18120790	0.1026694	0.1090240
LOC728640	cg26159090	0.1044333	0.0744315
AK127787	cg25102293	0.1029914	0.0343238
HLX	cg17108958	0.1032405	0.0449790
SRPK2	cg02813710	0.1131485	0.1250638
PCDHA13	cg25027798	-0.1016931	-0.0447850
KCND2	cg15536401	0.1125496	0.0413750
DCDC5	cg21758962	0.1298383	0.0509047
ALX4	cg25363445	-0.1687768	-0.0679886
ZFHX3	cg05340094	-0.1002702	-0.0212161
SLC5A12	cg05862393	0.1194612	0.0406921
DACT2	cg04680393	-0.1292147	-0.0297330
BC025350	cg26104143	-0.1025550	-0.0402475
CTXN3	cg23309843	0.1051087	0.0606455
AK092048	cg25161868	0.1107768	0.0585350
SALL1	cg00683332	-0.1436805	-0.0045126
NKX2-6	cg09618933	-0.1093871	-0.0631415
TET3	cg14702570	-0.1149805	-0.0137546
HS3ST1	cg05521150	0.1049740	0.0653794
SLC5A12	cg26210364	0.1067078	0.0466491
MIR340	cg11860760	0.1080024	0.1130163
SHANK2	cg07579831	0.1103508	0.0316507
PPIL2	cg27512565	0.1031035	0.1191705
DKFZp547J0510	cg12750431	0.1040196	0.0195708
TYRP1	cg25989745	0.1084904	0.0639991

KCNB1	cg26709285	-0.1000245	-0.0130675
POU3F3	cg03928875	-0.1453285	-0.0876868
CCDC140	cg06916239	-0.1242459	-0.0608652
OBSL1	cg04447708	0.1033495	0.0988812
SLC7A5	cg01829163	0.1105368	0.0944486
BC047484	cg01491071	0.1011019	0.0310873
PDX1	cg17200768	-0.1143192	-0.0893012
ZIC4	cg00334063	-0.1205338	-0.0531493
TNFRSF11B	cg03489427	0.1004217	0.0479695
ADCY4	cg25556905	-0.1100087	-0.0797854
ADCY4	cg23179456	-0.1400501	-0.0656607
BARHL2	cg00088183	-0.1097612	-0.0073072
SIX3	cg13905258	-0.1045044	-0.0598212
MIR183	cg21743907	-0.1036419	-0.0452294
NKX2-3	cg03711485	-0.1086765	-0.0255120
mir-108-1	cg03986418	0.1079746	0.1318765
SIX3	cg08696165	-0.1046349	-0.0392362
LOC255480	cg16559598	-0.1178727	-0.0210367
ADCYAP1	cg14200170	-0.1013491	-0.0036157
EBF3	cg16589299	-0.1021720	-0.0316996
SLC5A12	cg20092728	0.1146445	0.0453458
PAX3	cg16529477	-0.1096832	-0.0815705
NPC1	cg13421439	-0.1194738	-0.0487079
RHOJ	cg13199429	-0.1057039	-0.0454063
AK024936	cg14393923	0.1038193	0.0905213
FAM207A	cg02930963	0.1035153	0.0848348
MIR1243	cg06254768	0.1040927	0.0024179
CBLN1	cg06919440	-0.1194981	-0.1349869
PAX3	cg18077971	-0.1579233	-0.1261241
BC040304	cg15844438	0.1122886	0.0817321
ETS1	cg23774988	-0.1016640	-0.0686372

TXNIP	cg10713626	-0.1037301	-0.0744196
SCCPDH	cg14094521	0.1067686	0.1089175
SP9	cg08813062	-0.1240808	-0.0365725
NR4A2	cg18786593	-0.1175889	-0.0604366
CNR2	cg03611151	-0.1055246	-0.0148517
LOC100128239	cg16082695	0.1040723	0.0791492
ISM2	cg02034328	0.1042289	0.1019556
ADCY4	cg13631572	-0.1174449	-0.0351177
DPP6	cg10552126	-0.1122217	-0.0003559
ADCY4	cg16215203	-0.1172913	-0.0597037
POPDC3	cg16734734	0.1039432	0.0533165
DMRTA2	cg23097402	-0.1051854	-0.0363167
RALYL	cg25757598	-0.1077858	-0.0372598
RGS20	cg24645214	-0.1150925	-0.0483674
AX747372	cg18448949	-0.1187160	-0.0392604
LGALS8	cg18322510	0.1062610	0.2014020
ABCA13	cg02845274	-0.1062262	-0.0050174
BARHL2	cg24453699	-0.1216496	-0.0131754
MIR124-1	cg18246262	-0.1043957	-0.0352377
TRIM36	cg14132888	-0.1225024	-0.0156796
PDGFRA	cg06973595	-0.1238537	-0.0082039
LOC154860	cg27393010	-0.1051439	-0.0223084
FOXB1	cg21253459	-0.1166186	-0.0074233
FOXA1	cg19578835	-0.1022915	-0.0412477
TAC1	cg01287975	-0.1174638	-0.0438560
PAX3	cg09424526	-0.1035105	-0.0236433
GPR39	cg12157646	0.1164163	0.0475691
AGPAT4	cg26872907	-0.1014030	-0.0553293
LRRC4C	cg17949440	-0.1189794	-0.1094734
MEIS1	cg07809589	-0.1126549	-0.0703398
ARHGEF16	cg10328768	0.1001278	0.0725527

HDGFL1	cg26230039	0.1085959	0.0822853
FLJ23152	cg24071719	-0.1220715	-0.0866560
MIR4300	cg20219053	0.1021268	0.0036703
T	cg23302682	-0.1106680	-0.0748138
FAM3B	cg09179211	-0.1280524	-0.1927753
FOXG1	cg25078444	-0.1081361	-0.0862568
SLC5A7	cg26001902	-0.1280932	-0.0789488
LOC286184	cg02197192	-0.1314968	-0.0727589
LOC145845	cg04640920	-0.1042300	-0.0247391
CCDC140	cg01841641	-0.1145781	-0.0490557
ZBTB20	cg10911004	0.1022476	0.0562652
LRRC4C	cg19849428	-0.1079825	-0.0829319
EBF2	cg08283882	-0.1289306	-0.0508039
CD38	cg15994026	-0.1129467	-0.0650444
LOC100130155	cg15715477	-0.1202482	-0.0937751
CNR2	cg26404511	-0.1335413	-0.0252860
STRA8	cg23903035	0.1006868	0.0937747
FAM207A	cg10007534	0.1225716	0.1120182
CASR	cg11008866	-0.1017786	-0.0059483
SIM2	cg21697851	-0.1027212	-0.0038089
PAX6	cg20014398	-0.1236693	-0.0711978
CBLN1	cg02809746	-0.1206720	-0.0541757
RNF175	cg18355902	-0.1271932	-0.0729838
SATB2	cg10168149	-0.1009922	-0.0334369
FOXD1	cg18063312	-0.1237103	-0.1122693
C13orf33	cg19863655	-0.1145676	-0.0079773
LOC100506274	cg06720768	-0.1106381	-0.0176291
ADCY8	cg13912117	-0.1024663	-0.0130179
TBR1	cg22794704	-0.1469631	-0.0355343
DSCR6	cg16886987	-0.1178744	-0.0411601
EVX2	cg15133351	-0.1493974	-0.0731471

GBA3	cg24512731	-0.1093054	-0.0023675
PRDM6	cg02081006	-0.1285646	-0.0618576
LOC286189	cg21479226	-0.1269771	-0.0076465
EVX2	cg14118515	-0.1610922	-0.0403519
HTR4	cg12825070	-0.1009211	-0.0231072
NR4A2	cg11856078	-0.1170576	-0.0050271
FAM3B	cg10054197	-0.1137068	-0.1407630
PRDM14	cg11229513	-0.1050604	-0.0588580
DMRT3	cg19291576	-0.1396551	-0.0563173
PKNOX2	cg03419885	-0.1017826	-0.0281104
NBLA00301	cg22517735	-0.1011068	-0.0329391
MARCH11	cg16150752	-0.1071953	-0.0324181
SLC6A5	cg02027945	-0.1212780	-0.0282470
BC031238	cg04597985	0.1000892	0.0255522
LOC100130992	cg18794404	-0.1132725	-0.0217928
EOMES	cg11642106	-0.1014266	-0.0324487
PKDCC	cg14882311	0.1064891	0.1379679
OCA2	cg04803843	-0.1310316	-0.0117010
DMRTA2	cg26560222	-0.1073123	-0.0411774
SALL1	cg08526074	-0.1042057	-0.0675537
TBX4	cg00037457	-0.1130234	-0.1103478
LINC00461	cg24804195	-0.1026727	-0.1258471
LHX9	cg06371502	-0.1184641	-0.0015204
BC062758	cg13336662	0.1132549	0.0614202
AF071167	cg21404045	-0.1065988	-0.0075464
EVX2	cg11359133	-0.1107262	-0.0419957
SLIT2	cg10947633	0.1040768	0.0445162
LHX9	cg16353957	-0.1101648	-0.0321444
OTOP1	cg03160466	-0.1025783	-0.0511647

Table A.4: GO enrichment analysis results. GO terms specifically enriched based on the 651 PE-associated sites obtained from linear modeling.

Closest_TSS_gene_name	cpg	Discovery Cohort p value	Validation Cohort p value
LAPTM4A	cg09394306	0.1762488	0.1155985
ERGIC1	cg26813604	0.1793689	0.1018817
AX747766	cg17850498	0.1924544	0.1370919
KRT15	cg26625897	0.2981361	0.2393815
INHBA	cg11079619	0.2159444	0.1154296
NCF4	cg08612539	0.1489004	0.1278609
C10orf10	cg02334081	0.1270129	0.0764630
DENND2D	cg20317872	0.1824586	0.1115317
ST3GAL1	cg00592695	0.2622026	0.1675079
TGM1	cg23696550	0.1571351	0.0918194
STARD13	cg26651514	0.2159330	0.1720341
LOC100505839	cg19140548	0.2106330	0.1345067
MIR3150A	cg07158065	0.1649417	0.1087298
LIMCH1	cg03822934	0.1697386	0.0898849
ZNF395	cg01713086	0.1398992	0.0993220
CEBPA	cg22526990	0.1490031	0.1122649
PCDH1	cg23044186	0.1053694	0.0980231
GDPD5	cg05521767	0.1756805	0.1253889
FAM110A	cg16606561	0.1726153	0.1265220
ANO6	cg25162927	0.1492231	0.0879258
AMN1	cg08198187	0.1293568	0.1163599
FAM105B	cg18190824	0.1724711	0.1077959
CMIP	cg16353318	0.1328936	0.0970200
AMZ1	cg14605117	0.1968040	0.1150958
HK1	cg15258080	0.1559307	0.1034492
PIK3AP1	cg27479162	0.1393245	0.0965035
DLG5	cg27328839	0.1679715	0.1232817
FLJ32224	cg26185836	0.1118404	0.0820114

NDRG1	cg14143441	0.2167347	0.1654555
AMZ1	cg03653726	0.1661188	0.1341008
FN1	cg12436772	0.2704475	0.1908631
PHYHIP	cg26509870	0.1914165	0.1480857
IL1R2	cg24617203	0.1575016	0.1111186
IER5L	cg13908988	0.1310268	0.0690969
C1orf98	cg19694404	0.1588839	0.0793082
CMIP	cg01033642	0.1346521	0.1117363
CYP11A1	cg15903956	0.1834619	0.1159416
MSH4	cg17966362	0.1547178	0.0852058
TANK	cg22060367	0.1319618	0.0933415
TRIB3	cg15799353	0.1310048	0.1028805
CALML3	cg17115419	0.2037900	0.1361790
SRPRB	cg16275903	0.1164611	0.0582654
DHX32	cg23997887	0.1180493	0.0825463
GRIP2	cg10586672	0.1640134	0.0986082
ZNF385A	cg09676376	0.1343436	0.1137443
NCOR2	cg21626573	0.1257014	0.1026060
ZP3	cg16772533	0.1093900	0.0878149
LINC00284	cg05452692	0.1720094	0.1091923
CHI3L2	cg26366091	0.1718769	0.0843807
CMIP	cg10246581	0.1560846	0.1085283
PHLDA3	cg07950244	0.1843856	0.1242228
FLNB	cg23730027	0.1653212	0.1054604
BUB1B	cg25653839	0.1324543	0.0792319
GALNT2	cg23677911	0.1656313	0.1393207
SLCO2A1	cg15765546	0.1299143	0.0823860
ARHGEF4	cg25500616	0.1479462	0.1090096
CORO1C	cg09182455	0.1719971	0.1022355
LOC400238	cg10108710	0.1520188	0.1303691
ST3GAL1	cg00736681	0.1553735	0.0947189

RRBP1	cg12632411	0.1794180	0.1030725
BSCL2	cg07094785	0.1152064	0.0585061
TINAGL1	cg16103203	0.1141249	0.0848134
LIMS2	cg13795819	0.1218462	0.1014617
BET1	cg16568084	0.1477034	0.1025327
PAPPA2	cg10994126	0.1511426	0.1173865
SLC11A1	cg17010118	0.1565582	0.1221889
NPAT	cg12892243	0.1319349	0.0816390
LINC00310	cg25032603	0.1521295	0.0810892
TIMP3	cg25245338	0.1984709	0.1535695
ZBTB4	cg07168214	0.1238338	0.0844725
PARD6B	cg10187713	0.1883095	0.1071141
FLJ32224	cg01924561	0.1720685	0.1200934
KRT86	cg22193385	0.1976942	0.1139649
BCL6	cg17394304	0.1408706	0.0766436
CHI3L2	cg10045881	0.2356582	0.1208264
PAPPA2	cg18236464	0.2148526	0.1231496
PDLIM2	cg26366616	0.1253932	0.1109417
SNORD54	cg02508743	0.1042322	0.0879402
ZNF385A	cg12150931	0.1175044	0.0793333
IFITM3	cg20151221	-0.1416310	-0.0662126
RALGDS	cg14103123	0.1399157	0.0972403
LTF	cg13672136	0.1306179	0.1425736
PHF17	cg03078141	0.1264031	0.0675889
TRIM29	cg26247168	0.1449313	0.1125867
VILL	cg06641593	0.1151523	0.1097128
FLNB	cg24204017	0.1304899	0.0934378
TRNA_Arg	cg21545548	0.1435637	0.1662567
LOC285954	cg14704980	0.2042223	0.0730101
MMP15	cg04211309	0.1067482	0.0645321
PHF17	cg17233452	0.1202395	0.0723816

NR2F2	cg17661642	0.1185910	0.1190850
MBNL2	cg03099780	0.1443609	0.1221243
CSF3R	cg07285167	0.1071080	0.0951602
PWWP2B	cg25961733	0.1186476	0.0944467
FAM160B2	cg20576064	0.1498508	0.0786766
PPFIA1	cg18564881	0.1300447	0.0840991
C11orf16	cg14402562	0.1144046	0.0720752
MIR4284	cg06480942	0.1539992	0.0906186
MLL5	cg07351322	0.1352379	0.1085833
LIMCH1	cg22204103	0.1482763	0.0672620
DQX1	cg24544105	0.1027483	0.0775145
XIRP1	cg01127412	0.1267807	0.0749353
JUNB	cg22996170	0.2145465	0.1404922
SNORA11B	cg21906866	0.1403945	0.0744550
RDH13	cg20669049	0.1260385	0.1152781
LNPEP	cg24598187	0.1386856	0.0854524
A2ML1	cg03490200	0.1053874	0.0656080
SIPA1L2	cg01432692	0.1047955	0.0633543
TEAD3	cg10893014	0.2468720	0.1912331
AMBRA1	cg07819010	0.1327321	0.0785149
MSI2	cg04573500	0.1344202	0.0788883
RBM47	cg06332621	0.1067688	0.0745359
MYO7A	cg15433043	0.1140219	0.0859117
PKM2	cg22234930	0.2032501	0.0994724
GALNTL4	cg12511310	0.1063327	0.0816910
ADAM12	cg02494582	0.1425002	0.0880331
C22orf31	cg20080983	0.1202626	0.0873451
C2	cg11049439	0.1041528	0.0946484
LEP	cg06987369	0.1052829	0.0877503
FLNB	cg02770406	0.2004214	0.0934495
POLD3	cg00453717	0.1249826	0.0917825

LOC100507091	cg05399718	0.1035536	0.0893499
CLEC2L	cg01581050	0.1018998	0.0900954
DKFZp547K2416	cg05364179	0.1385653	0.0742379
DQX1	cg02034222	0.1045459	0.0720015
NFKBIZ	cg06056170	0.1115525	0.0945533
FNIP2	cg11637968	0.1517075	0.1030839
TRIM8	cg16654458	0.1457415	0.1243255
U6atac	cg10189029	0.1392404	0.0946701
PLEKHA6	cg11131599	0.1100419	0.0707025
MIR4499	cg23485627	0.1191803	0.1076308
KIAA1614	cg13467459	0.1354345	0.1129899
FAM150B	cg13467628	0.1283506	0.0746500
AP1G1	cg10407113	0.1192888	0.0701779
CMIP	cg08946161	0.1083718	0.0756837
SLC45A1	cg11283860	0.1457413	0.1234598
KIAA1211	cg08112737	0.1141088	0.0767745
DGCR8	cg20012247	0.1559105	0.0977748
KAZALD1	cg02448743	0.1012339	0.0792401
PCID2	cg10350215	0.1080837	0.0838975
ITPRIP	cg16301004	0.1514493	0.1319952
PSG11	cg07920195	0.1003350	0.0901322
EPAS1	cg15129144	0.1376505	0.1049498
ARID5A	cg05933789	0.1087142	0.0621318
CAPNS1	cg18709710	0.1074846	0.0971703
CORO1C	cg12647920	0.1775886	0.1179099
ZSWIM4	cg25722029	0.1251416	0.0654344
LPP	cg17317338	0.1052743	0.0759132
CSRNP1	cg13062627	0.1540136	0.0866311
EGFR	cg04156940	0.1185266	0.0896936
GBX2	cg26454433	0.1026525	0.0676502
FAM18A	cg03777414	0.2038463	0.1079243

NR3C1	cg23400056	0.1214372	0.0741881
LMNA	cg27182012	0.1762920	0.1649320
C10orf90	cg04507071	0.1104607	0.0689047
GNA12	cg08799766	0.1596030	0.1189175
BCL6	cg05663031	0.1195386	0.0689620
TIMP3	cg27221424	0.1132492	0.1196439
C10orf26	cg20340720	0.1228038	0.0926289
PEX14	cg22715764	0.1011310	0.0550781
EPAS1	cg25589945	0.1249127	0.0897729
CLIP4	cg17417693	0.1537936	0.1072281
FLNB	cg17338821	0.1010904	0.0815451
U6	cg02275040	0.1191463	0.1066391
INHBA	cg18413237	0.1255024	0.0932958
OLFML2A	cg00992055	0.1117773	0.0897939
STON1-GTF2A1L	cg26837192	0.1032263	0.0693357
COL17A1	cg13553455	0.1557471	0.1194540
RRBP1	cg00990977	0.1480687	0.1274827
A2LD1	cg20198393	0.1032439	0.0807160
HIST1H1T	cg19722391	0.1444891	0.1910316
LOC100505839	cg03880642	0.1158724	0.1152257
CMIP	cg04897892	0.1475595	0.1327736
ENSA	cg11147155	0.1121804	0.0724865
SCOC	cg05347925	0.1201696	0.0960729
LOC646324	cg20428989	0.1186111	0.0801354
UBE2E1	cg03668982	0.1554643	0.1039143
TRAM2	cg26301143	0.1107192	0.0853976
FSTL1	cg11622516	0.1442870	0.0818265
NPB	cg21823502	0.1085694	0.0945396
MGAT3	cg05541460	0.1058957	0.0673778
TCF25	cg07891440	0.1229996	0.1183821
PMEL	cg08869883	0.1048622	0.0667430

PLA2G4E	cg12452386	0.1576734	0.1111678
KLF5	cg26531076	0.1152145	0.1050242
EPS8L2	cg27649971	0.1448863	0.0976509
STK24	cg19861486	0.1233168	0.1020152
SH3BP5	cg04858987	0.1667636	0.1539031
LINC00163	cg11327657	0.1726848	0.1114610
MOB3A	cg07381806	0.1319074	0.1250795
PNMA2	cg01862311	0.1361850	0.0826204
PWWP2B	cg26450254	0.1109619	0.1021787
DL489896	cg18474072	0.1599509	0.1675220
RANBP3L	cg15169286	0.1124261	0.0798795
ZNF783	cg18783886	0.1046545	0.1042132
NACC2	cg14350701	0.1470443	0.0855510
C2orf61	cg14327359	0.1252225	0.0673201
ANKDD1A	cg18926409	0.1090146	0.0803986
CHSY1	cg12361046	0.1527898	0.1329554
ITCH	cg09049982	0.1150276	0.0829737
FAM124B	cg00392155	0.1020767	0.0630411
CEP41	cg09507697	0.1067296	0.0845174
EPHA1	cg05385805	0.1209984	0.0694392
FLJ13197	cg24279243	0.1034128	0.0581950
INSIG1	cg12979992	0.1504744	0.1465894
GRHL3	cg13987674	0.1012415	0.0862911
CREB1	cg10440877	0.1104541	0.0669938
CPPED1	cg01323840	0.1284753	0.0538420
USHBP1	cg14584702	0.1080895	0.1357364
AX748283	cg15676500	0.1149958	0.0728982
AK122764	cg18743287	0.1433035	0.1097513
OR52K2	cg12250761	0.1234940	0.0707486
LOC100130275	cg00110654	-0.1177342	-0.0621028
C14orf1	cg01284448	0.1257140	0.1195566

PABPC3	cg14920808	0.1281780	0.0800569
CGA	cg07981495	0.1522790	0.1054230
SLC45A4	cg20555854	0.1199058	0.0930851
C8orf42	cg02341578	0.1178303	0.0881730
STC2	cg25592413	0.1408410	0.0302803
CTAGE1	cg20586124	0.1447415	0.1233495
GALM	cg22860917	0.1407808	0.0837119
SLC7A5	cg06665333	0.1203619	0.1322104
TBCD	cg14605961	0.1121029	0.1552277
PWWP2B	cg05467828	0.1067891	0.0772865
JPH2	cg18942298	-0.1034076	-0.0929964
ZNF175	cg10668363	0.1674458	0.1949214
GNA12	cg01839603	0.1469256	0.0912815
NEBL	cg00496126	0.1381810	0.1058340
TNFAIP8L2-SCNM1	cg16565154	0.1153559	0.0470728
ZHX2	cg15531512	0.1018891	0.0843642
EIF4H	cg26311610	0.1087352	0.1188633
5S_rRNA	cg11033588	0.1055937	0.0879960
C4orf26	cg10409560	0.1062920	0.1158302
ZFP36L1	cg06617636	0.1245903	0.1194213
STRA8	cg06825631	0.1367064	0.1419325
SLC11A2	cg25493658	0.1048224	0.0581554
CALD1	cg03188976	0.1030026	0.0675175
RAD1	cg15003812	0.1069456	0.1099495
MIR4708	cg01703196	0.1299448	0.1240728
EXTL2	cg04098985	0.1445115	0.1167884
BC044741	cg23751171	0.1093071	0.0908032
SH3BP5	cg18444702	0.1166323	0.1097695
KCP	cg00139092	0.1008959	0.0789227
WDR69	cg09638264	0.1196858	0.1317754
SLC20A2	cg22855020	0.1127744	0.1207485

MTMR10	cg18120790	0.1265137	0.1179497
LOC152225	cg03538833	-0.1024445	-0.0525828
AK091866	cg05211068	0.1003265	0.0851387
SERTAD4	cg11214507	-0.1043236	-0.0664014
GPR132	cg21510995	0.1209976	0.0852175
RYBP	cg10805254	0.1090491	0.0817135
ERRFI1	cg00768179	0.1033485	0.0958330
PKP2	cg19677302	0.1109476	0.0834394
DNAJC6	cg18191867	0.1072833	0.0745990
SYDE1	cg00713022	0.1431890	0.1056350
PCAT1	cg25632577	0.1166528	0.0833044
PTPN3	cg13715502	0.1016954	0.0855104
INHBA	cg00159987	0.1041149	0.1010672
MIR340	cg11860760	0.1272976	0.1164016
FNIP2	cg03227611	0.1117606	0.1039792
CABLES1	cg01288184	0.1277758	0.0387460
APOL4	cg11178302	0.1573320	0.1100429
EN1	cg01165776	-0.1143348	-0.0750388
ACSL6	cg27097034	0.1045707	0.0611043
LOC727677	cg12560931	-0.1036006	-0.0502981
TRNA_Leu	cg20430841	0.1037816	0.0831308
MFSD10	cg13785473	0.1072274	0.0948475
AX747853	cg14792548	0.1124007	0.0911584
HIPK4	cg09412728	0.1041352	0.0882879
COL17A1	cg15715892	0.1305339	0.1473990
ZFP42	cg00469814	-0.1070747	-0.0483893
GALNT2	cg17737409	-0.1011384	-0.0686482
UNC79	cg10532364	-0.1251421	-0.1095131
ST3GAL4	cg12804791	0.1439781	0.1188859
SHFM1	cg01534613	0.1155720	0.0884691
LOC100507582	cg16582517	0.1269999	0.0772773

DDR1	cg14279856	0.1221406	0.1070260
IDH2	cg19674091	0.1565227	0.1687085
SLC20A2	cg10806146	0.1092828	0.1477129
AK125212	cg15448894	0.1093031	0.0687104
LINC00310	cg19118951	0.1135028	0.0915168
TFAP2A	cg14993900	0.1045150	0.0998829
SAPCD1	cg10158997	0.1246066	0.1111205
RCBTB2	cg18731055	-0.1017133	-0.0912764
DUSP5	cg23608075	0.1066479	0.1336834
NRXN2	cg21660452	0.1092307	0.1008692
MIR938	cg07437737	0.1091087	0.1270809
PDPN	cg24671344	-0.1144931	-0.0636435
ANKRD52	cg20059881	0.1112127	0.0308736
CTBP2	cg17191109	0.1374573	0.1681589
GNA12	cg04849508	0.1472580	0.1525895
LRRFIP1	cg21708130	0.1387597	0.1133808
AF339817	cg22223119	0.1013930	0.1162137
NEDD4	cg20612128	-0.1016248	-0.0957753
EGFR	cg23757825	0.1044328	0.0984762
SIX3	cg13905258	-0.1238786	-0.0717545
TMEM174	cg22120063	-0.1014325	-0.1078645
mir-108-1	cg03986418	0.1237650	0.1492405
LOC157627	cg21762788	-0.1199716	-0.0596098
U1	cg24967553	0.1080177	0.0885870
TRIM40	cg12612406	0.1276182	0.1442272
RHOJ	cg13199429	-0.1322748	-0.0933939
STRA8	cg23903035	0.1390035	0.1620199
ACPL2	cg18802720	-0.1258434	-0.0370503
AK097686	cg08530838	-0.1264907	-0.1020174
NR2F2	cg07379949	-0.1137175	-0.0419030
PXK	cg24175188	-0.1079104	-0.0764821

LOC401022	cg10159630	-0.1314751	-0.1108394
LEP	cg19594666	0.1332098	0.0920887
PDX1	cg17200768	-0.1193134	-0.0598558
MIR603	cg21396949	-0.1008504	-0.0532213
AJAP1	cg13056990	0.1066677	0.1146372
CAST	cg12866104	0.1011459	0.0930617
AK024936	cg14393923	0.1245521	0.1159032
GPR39	cg24964130	0.1088200	0.0359574
DIO3AS	cg05369857	0.1071648	0.1268123
SLC39A1	cg25064552	0.1190307	0.1763146
HOXA1	cg18557185	-0.1229324	-0.0076852
USP34	cg03589282	0.1095716	0.1908809
SRGAP1	cg01099855	0.1004787	0.0821632
TBXAS1	cg24431161	0.1313699	0.1203989
IQCG	cg21915659	-0.1002350	-0.0425345
MED30	cg00592871	0.1055723	0.0470123
MIR4750	cg05213896	0.1107362	0.1218362
TBCD	cg16094026	-0.1071515	-0.0055439
SLC1A3	cg21050001	0.1425513	0.1451981
AP4E1	cg24311373	0.1164390	0.1230426
Mir_548	cg13007502	0.1354421	0.1597983
CADPS	cg15253604	-0.1029813	-0.0089923
EPHA2	cg05797770	0.1093776	0.1002196
ALX4	cg06245037	-0.1105070	-0.0043119
CCDC81	cg01663953	0.1041392	0.0780464
GPR39	cg17091301	0.1095570	0.0282611
ODZ4	cg02980047	-0.1135685	-0.1353950
WIBG	cg08338281	0.1000529	0.1429603
LOC100128239	cg16082695	0.1144923	0.1070973
PML	cg01947066	-0.1024238	-0.0617708
IGFBP2	cg06271720	-0.1064898	-0.0663750

NRN1	cg04187403	-0.1029897	-0.1057090
ATP5G3	cg07298772	-0.1038753	-0.0522010
CFLAR	cg23681745	0.1251774	0.1919709
LGALS8	cg18322510	0.1141797	0.1961030
FOXD1	cg18063312	-0.1559150	-0.1052904
MNX1	cg13345957	-0.1001399	-0.0286152
BARHL2	cg00088183	-0.1106928	-0.0089216
TBX4	cg00037457	-0.1549207	-0.1067988
HIVEP3	cg19619028	-0.1000616	-0.1464566
ZFYVE28	cg07042007	-0.1010364	-0.0841700
WRAP73	cg19135761	-0.1055765	-0.0935807
LOC401022	cg02746725	-0.1037036	-0.0242127
LOC154860	cg27393010	-0.1176741	-0.0304895
Mir_548	cg25029529	0.1001944	0.1692182
UBE2MP1	cg01947695	0.1031825	0.1467774
CD38	cg15994026	-0.1230982	-0.0249218
FAM207A	cg02930963	0.1071846	0.1367691
KCNIP2	cg03557857	0.1053333	0.0991780
LINC00085	cg16854917	0.1148057	0.1477937
LOC1720	cg05779458	-0.1103677	-0.0202628
Mir_548	cg05391892	0.1015469	0.1404382

Appendix B - Supplementary information for Chapter 3

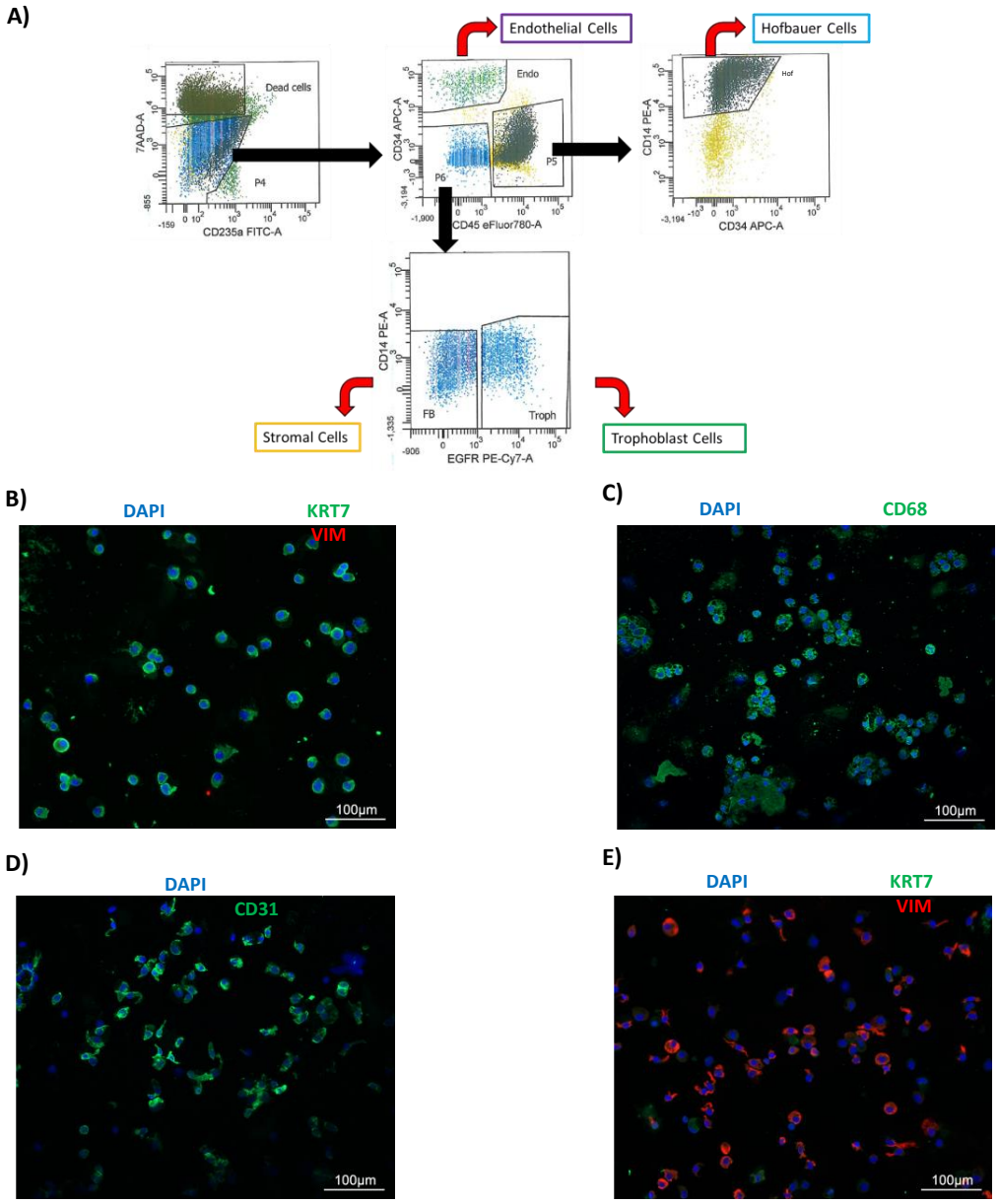

Figure B.1: Fluorescence-activated cell-sorting and immune fluorescence staining. **A)** Fluorescence-activated cell-sorting (FACS) workflow schematic. **B-E)** Immunofluorescence staining (IF) of term cell-sorted sample with known characteristic cell type markers that were not selected for in the FACS procedure. Nuclei are shown via DAPI staining (blue). Scale bars: 100µm. **B)** Trophoblasts (KRT7: green, VIM: red). **C)** Hofbauer cells (CD68: green). **D)** Endothelial cells (CD31: green). **E)** Stromal cells (VIM: red).

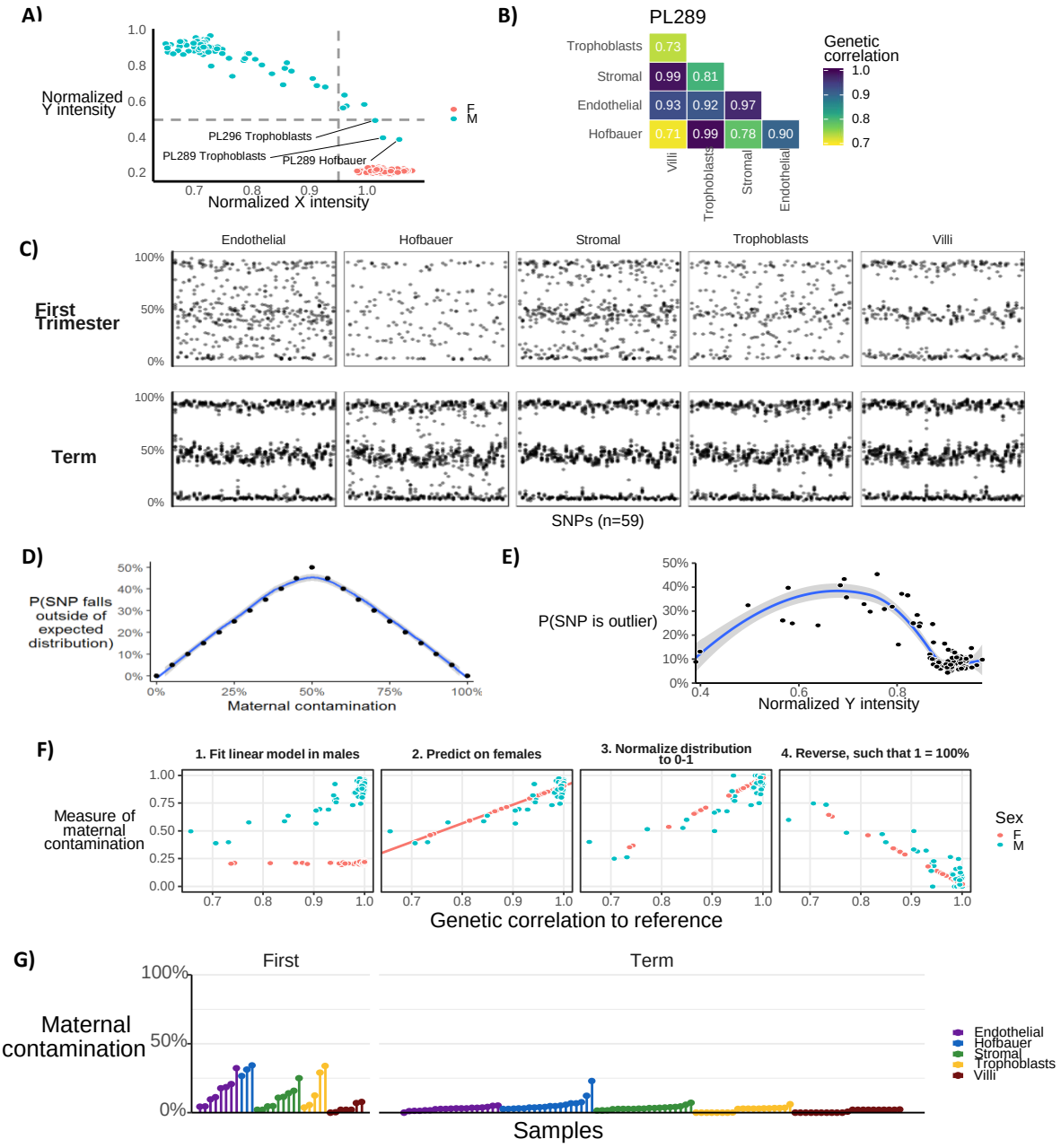

Figure B.2: Identifying maternal contamination. **A)** Total intensity over all probes from X and Y chromosomes normalized to total autosomal intensity can be used to determine sex. **B)** Within-donor sample-sample correlation on SNP probes. **C)** SNP distributions (n = 59 probes). **D)** Theoretical relationship between the average probability SNP is an outlier from the expected distribution, and maternal contamination. **E)** Empirically observed relationship between the average probability a SNP is an outlier, and normalized Y intensity, in male samples. Normalized Y intensity is a quantifiable measure of maternal contamination in male samples. **F)** Training a linear predictor of maternal contamination in male samples, then applying it to female samples. **G)** Estimated maternal contamination (y-axis) across first trimester and term samples.

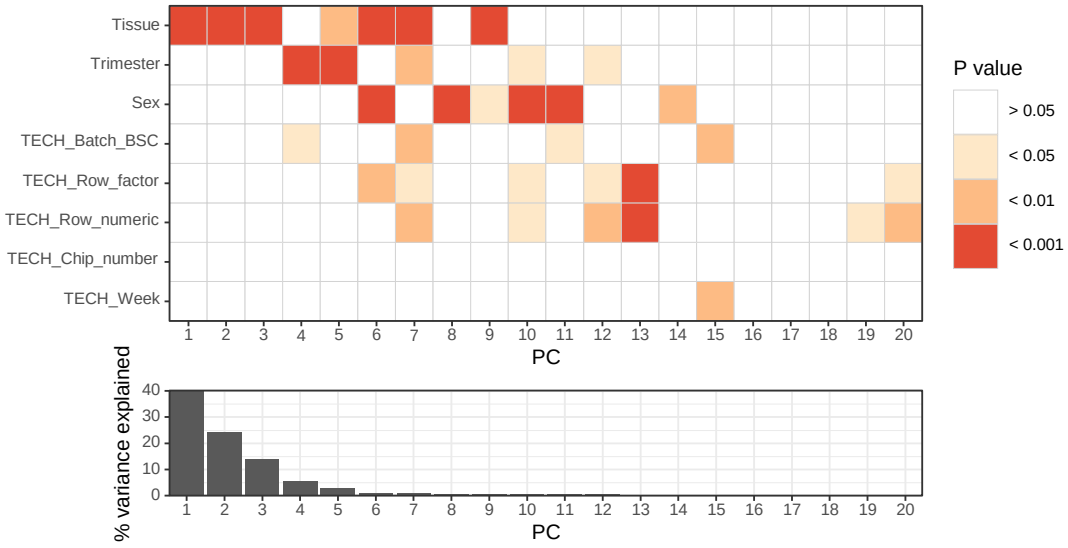

Figure B.3: Principal component (PC) associations with phenotype variables. Principal components were tested for their association with various biological and technical sample variables. Each PC was tested individually in a simple linear model with each sample variable.

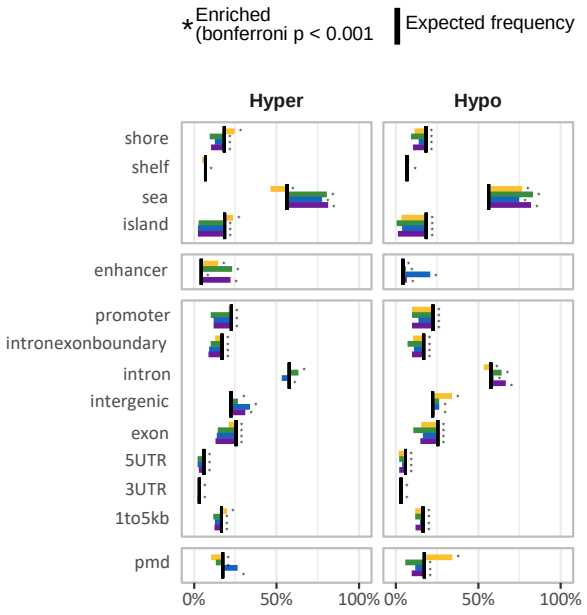

Figure B.4: First trimester differentially methylated CpGs enrichment for genomic location. First trimester differentially methylated CpGs were tested for enrichment at various genomic features (e.g. CpG island, enhancers, gene transcripts, PMDs).

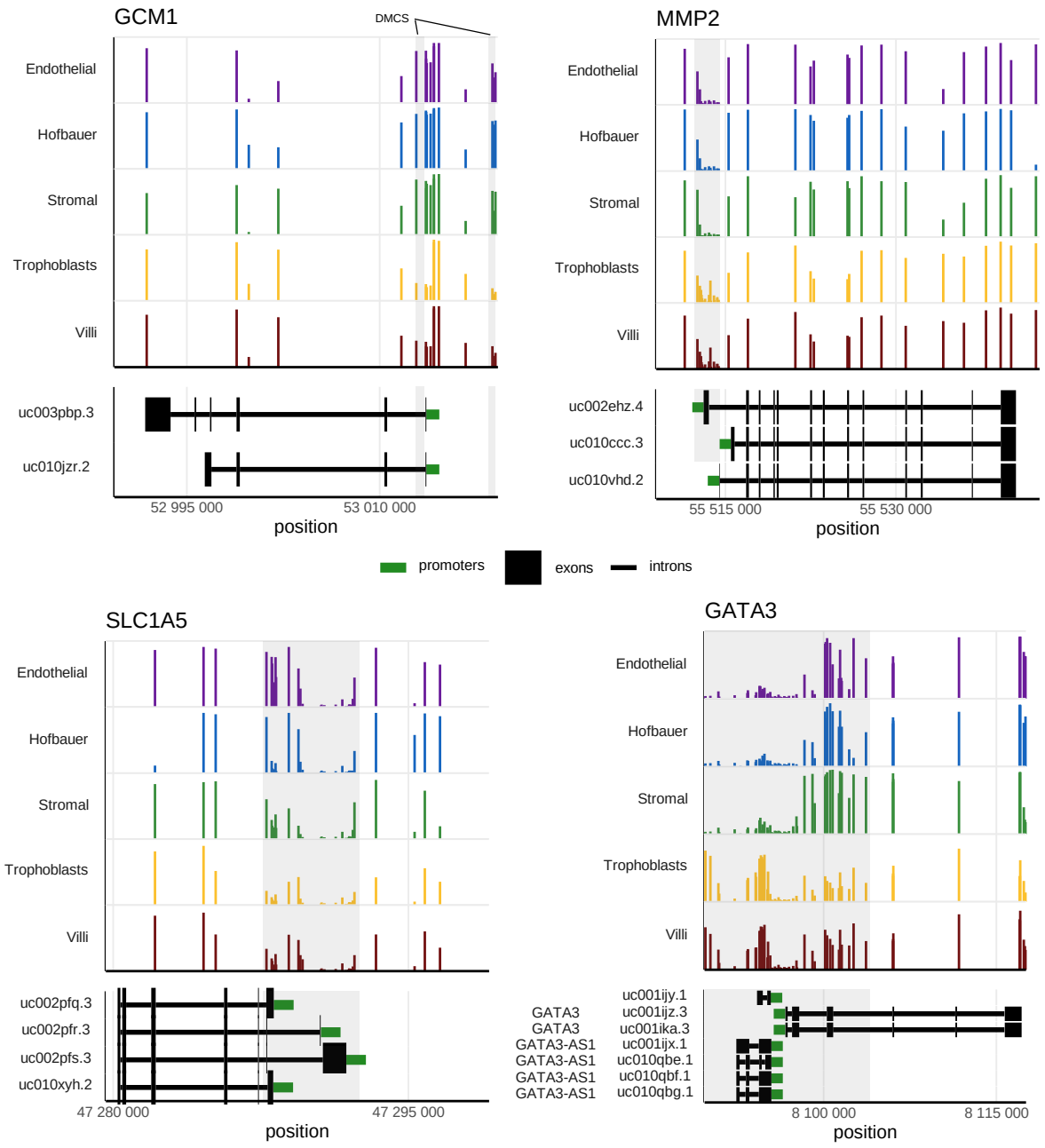

Figure B.5: Mean DNAm for each cell type across CpGs in selected functionally-relevant genes. Average term placental cell-specific DNA methylation across select genes. Differentially methylated regions (defined as regions with a high density of differentially methylated CpGs), are highlighted with a grey background.

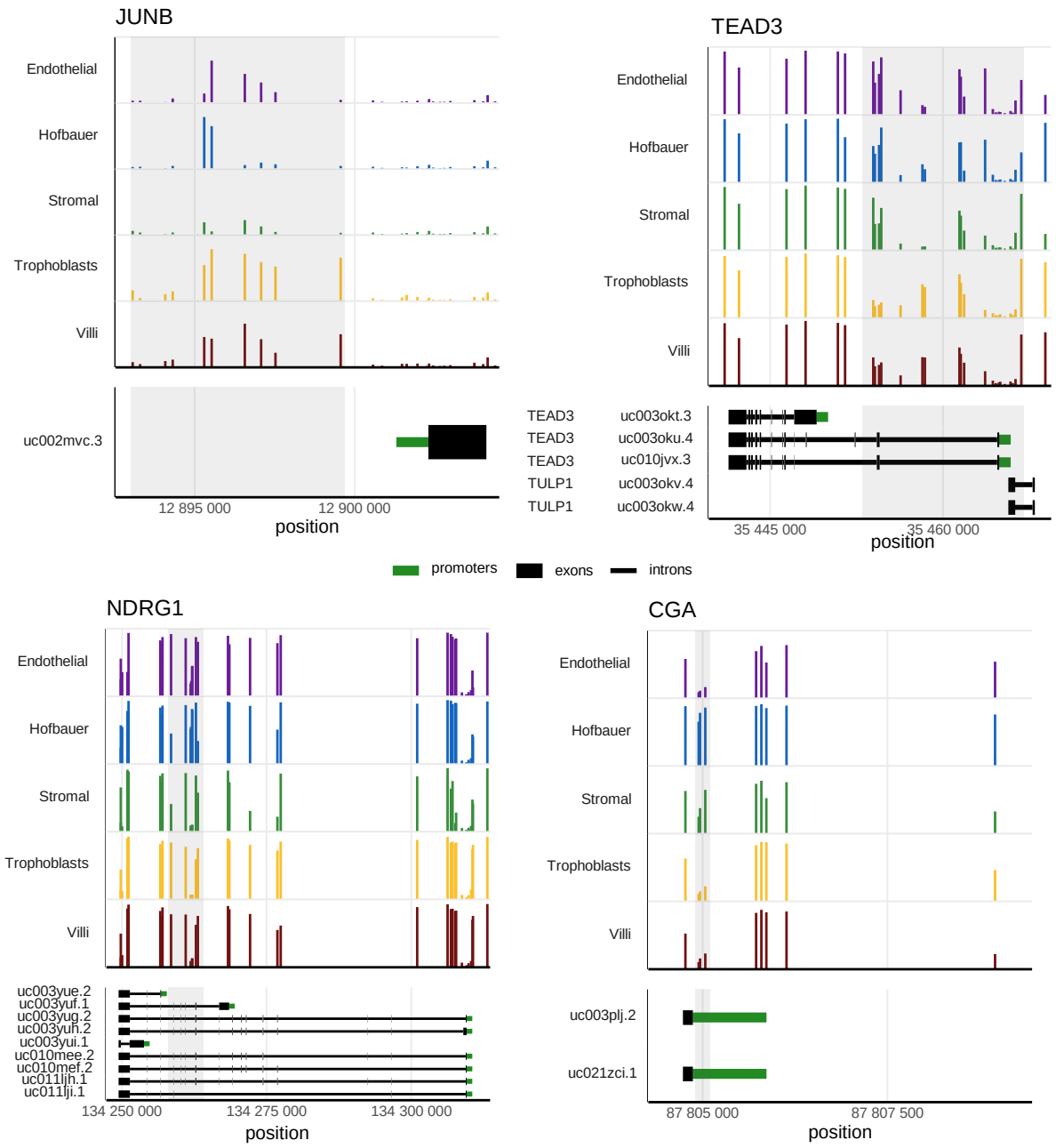

Figure B.6: Mean DNAm for each cell type across CpGs in selected preeclampsia genes. Average term placental cell-specific DNA methylation across select genes. Differentially methylated regions (defined as regions with a high density of differentially methylated CpGs), are highlighted with a grey background.

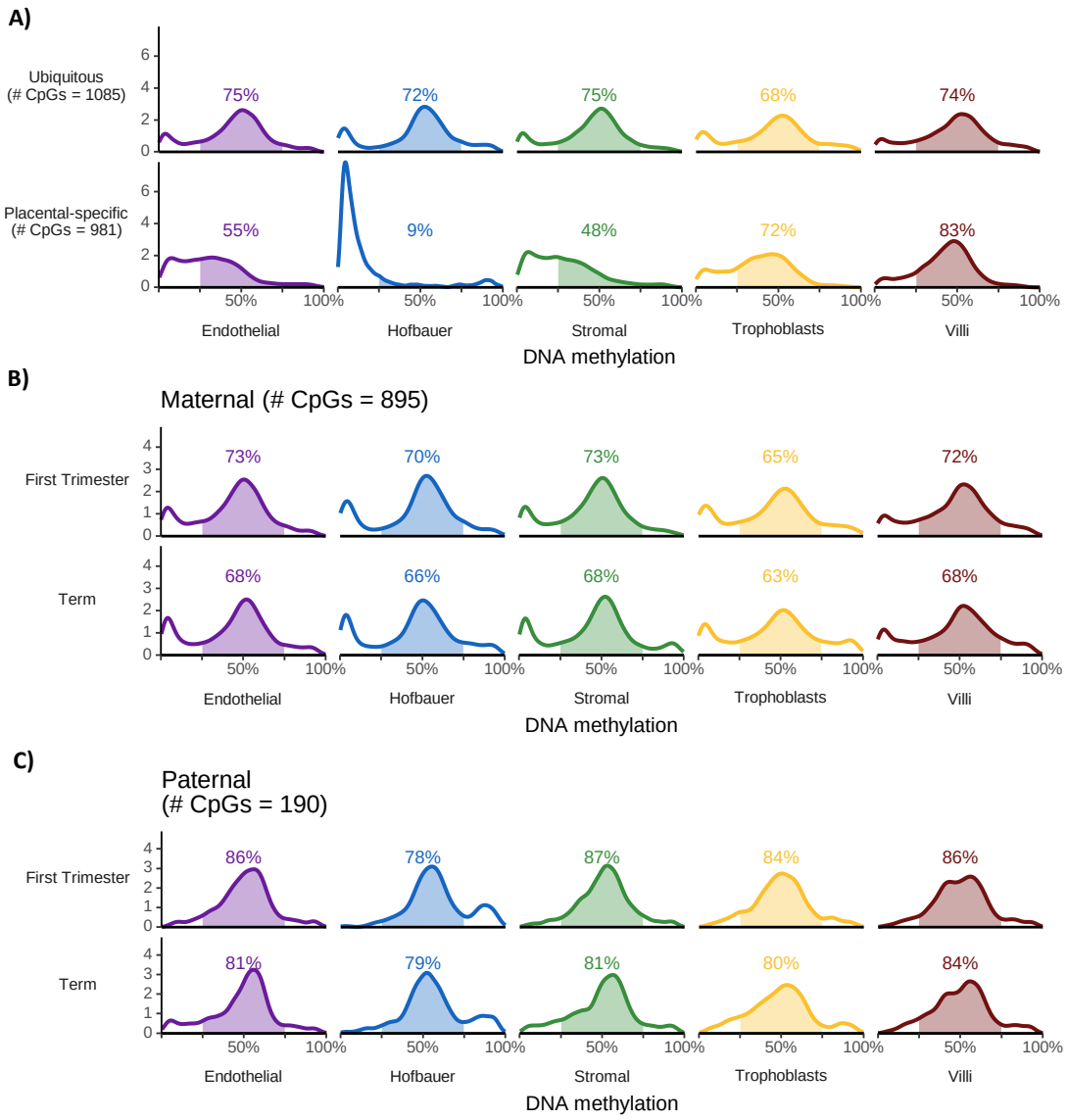

Figure B.7: Density graphs of CpGs in imprinted regions. **A)** Density plots (y-axis) of imprinted regions divided into those that are imprinted in more than one tissue (top) and placental-specific (bottom). The percentage of CpGs falling within 25%-75% is labelled in each plot. First trimester samples are shown. **B)** Maternal imprinted regions. Density of DNAm at CpGs in maternally imprinted regions. The total percentage of CpGs that have 25% - 75% DNAm are shown in each plot. **C)** Paternally imprinted regions.

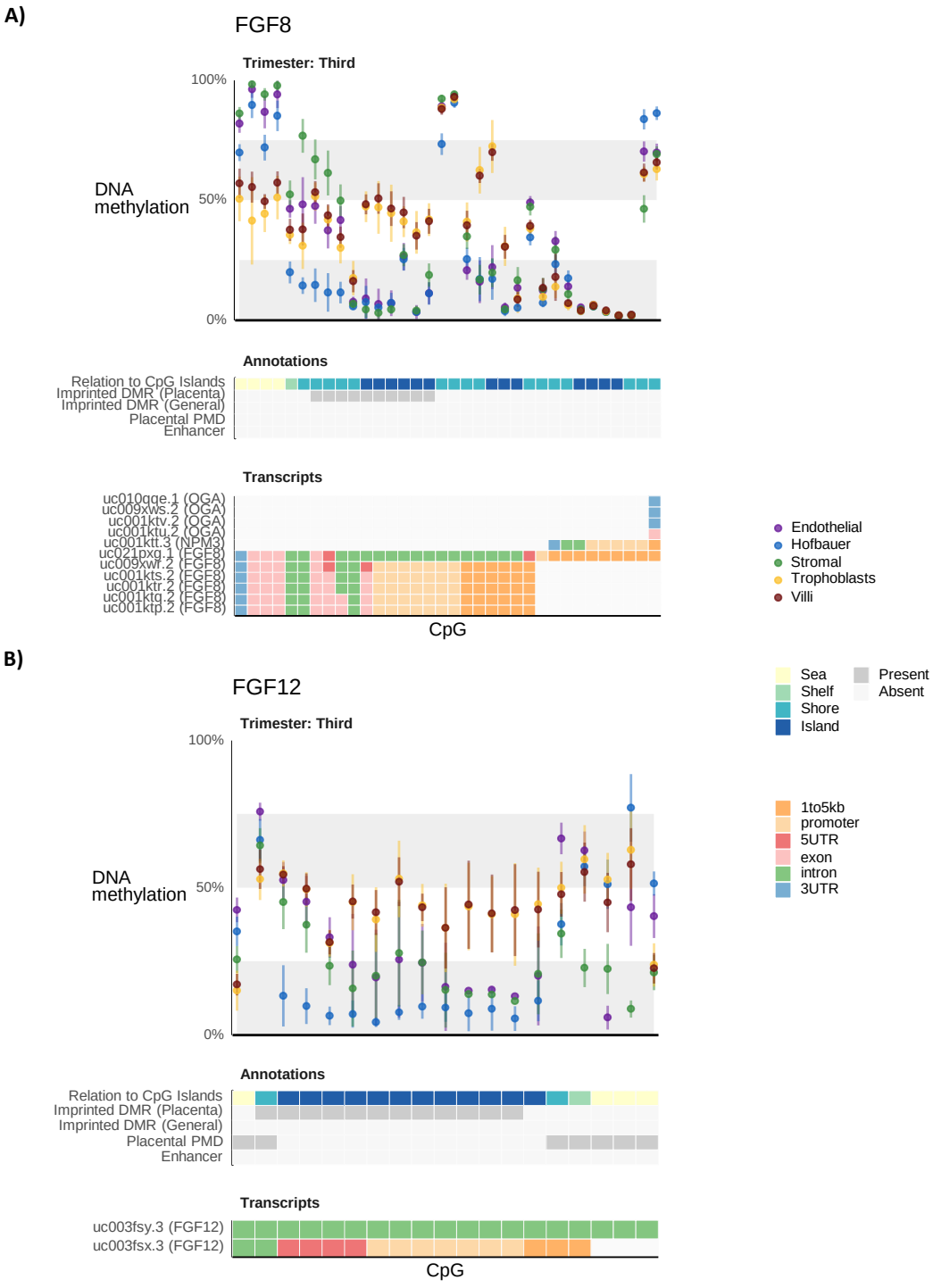

Figure B.8: DNAm at imprinted regions for specific genes. **A)** Cell-specific DNAm at placental-specific imprinted regions for genes FGF8 and **B)** FGF12.

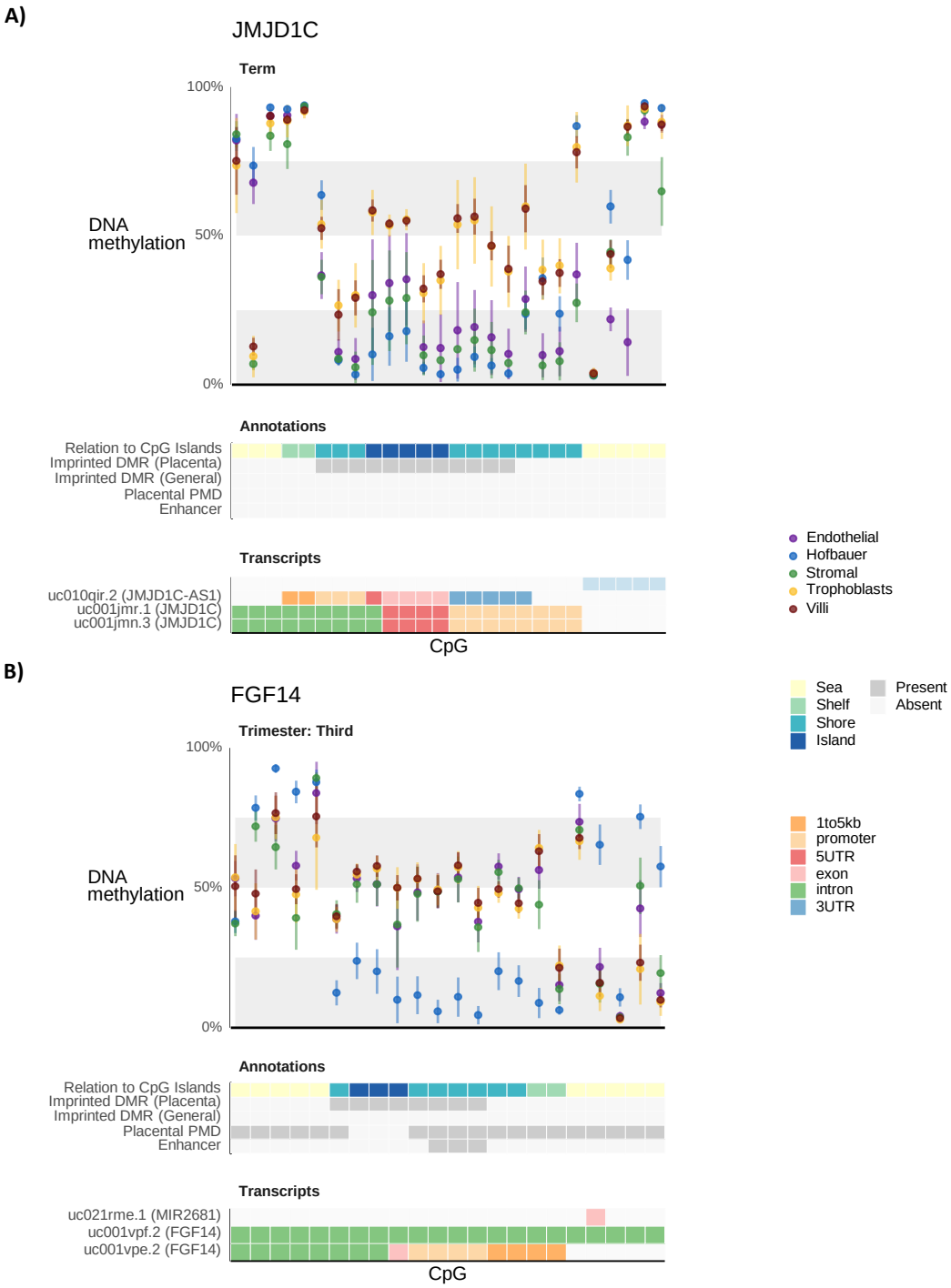

Figure B.9: DNAm at imprinted regions for specific genes. **A)** Cell-specific DNAm at placental-specific imprinted regions for genes JMJD1C and **B)** FGF14.

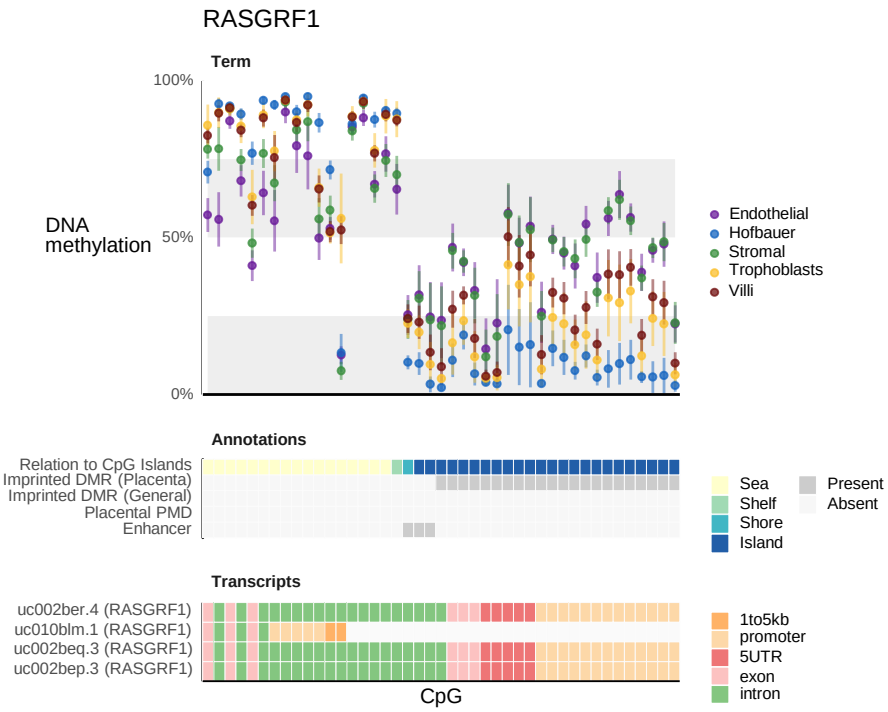

Figure B.10: DNAm at imprinted regions for specific genes. Cell-specific DNAm at placental-specific imprinted regions for genes RASGRF1.

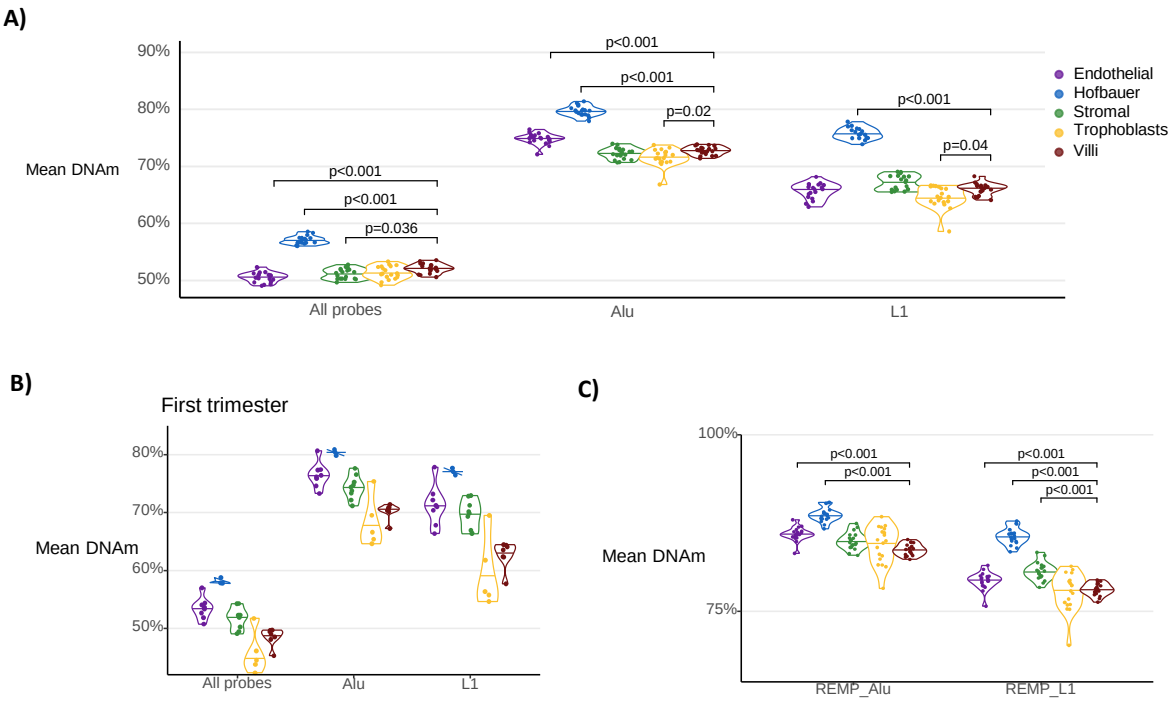

Figure B.11: DNAm summarized over repetitive elements. **A)** Repetitive element DNA methylation. CpG sites overlapping Alu and Line 1 (L1) elements were determined using the 'rmsk' track from UCSC. Mean DNAm over these CpGs was calculated for each sample. **B)** First trimester mean DNAm across repetitive elements and all 850k CpGs. **C)** REMP-predicted repetitive element DNAm in third trimester samples.

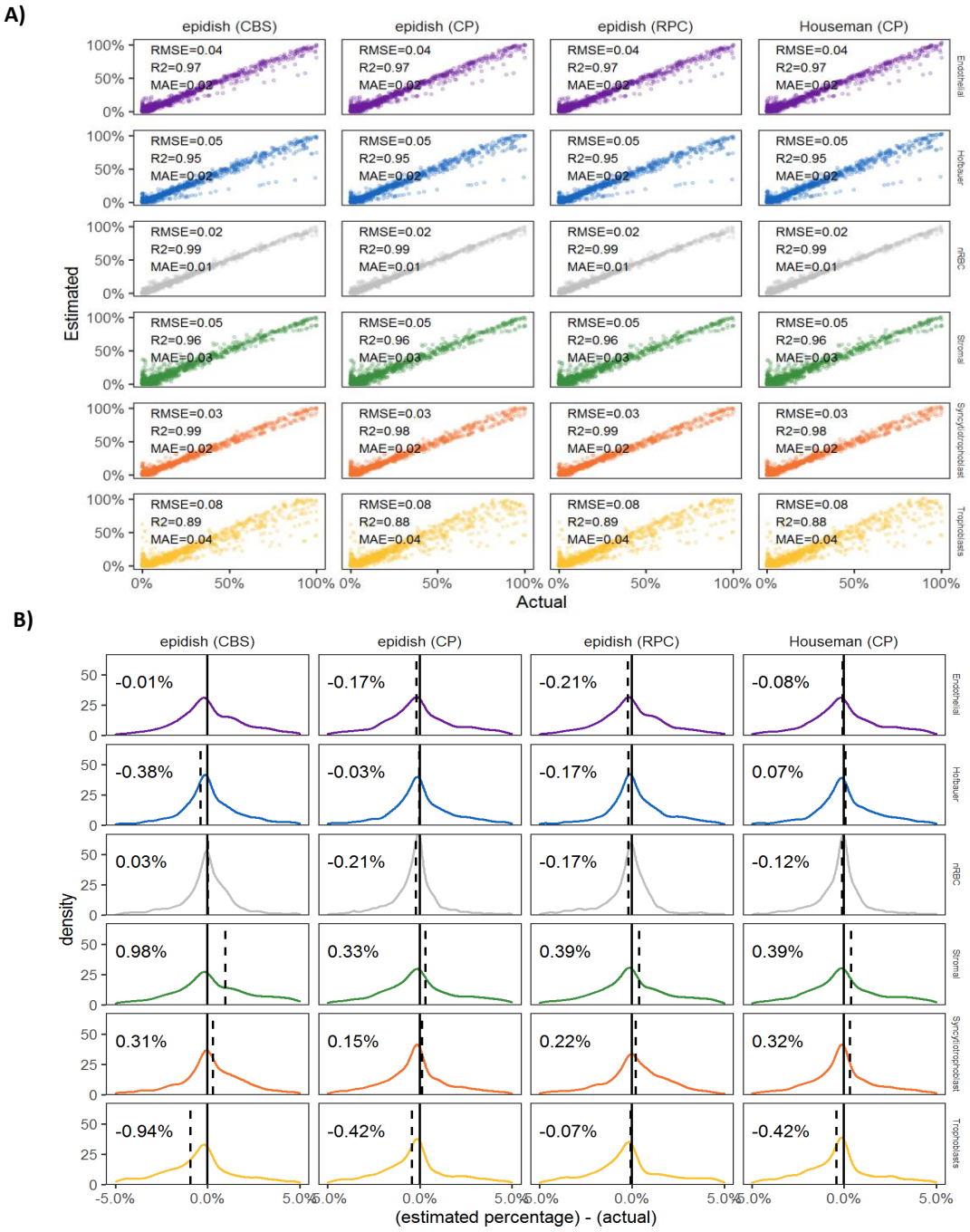

Figure B.12: Comparison of cell deconvolution algorithms. **A)** Estimated percentage by deconvolution (y-axis) by actual percentage used to construct in silico mixtures (x-axis). Performance metrics are shown for each algorithm and cell type. RMSE, root mean squared error; R2, R squared; MAE, mean absolute error. **B)** Distribution of deviations from deconvolution estimates and actual percentages for in silico mixtures. The mean deviation (estimated minus actual) is labelled in each panel as text, and as the dotted vertical line.

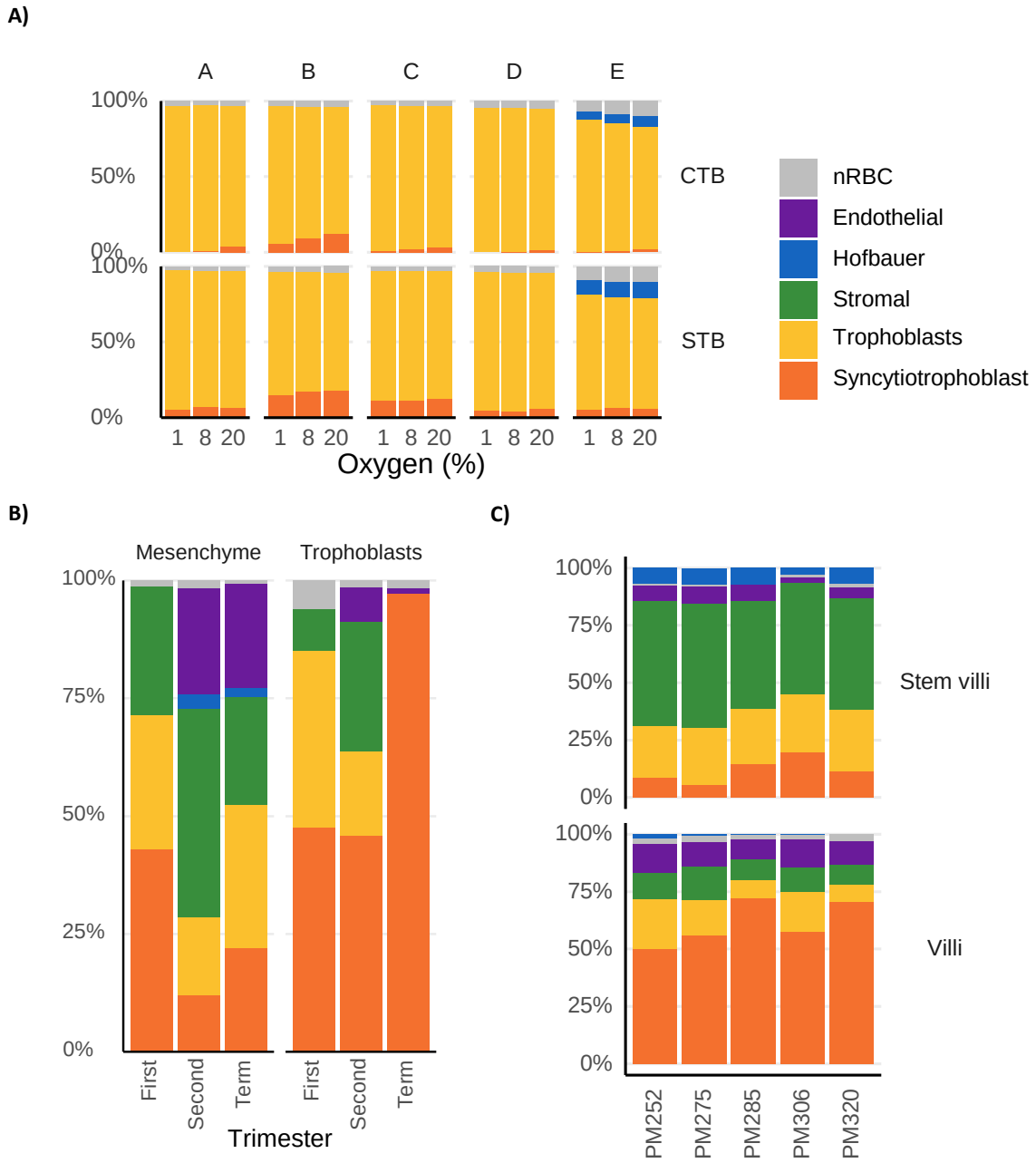

Figure B.13: Validating cell composition estimates. **A)** Cell deconvolution was applied to n=5 (labelled A-E) cultured trophoblast samples from Yuen et al. 2011 produced trophoblast-dominant samples. Trophoblast samples were treated in varying oxygen levels (1%, 8%, 20%). Half were maintained as CTB (top) and the other half was cultured for 48 hours (bottom), which promotes syncytialization. **B)** Enzymatic treatment to separate chorionic villi samples into inner mesenchyme and outer trophoblast layer samples. Both types of samples are heterogeneous in cell composition but mesenchymal samples are enriched from endothelial and stromal cells, whereas the outer chorionic villi samples are mostly trophoblast. **C)** Chorionic villi was processed to isolate large stem villi, produced samples that resulted in mainly stromal in proportion compared to normally processed villi. CTB: cytotrophoblast; STB: syncytiotrophoblast.

Table B.1: Top 10 GO enrichment for cell DMCs.

Group	Celltype	TERM	P.DE	FDR	Generatio
First	Endothelial	immune response	0.0e+00	0.0000000	0.2715691
First	Endothelial	immune system process	0.0e+00	0.0000000	0.2657495
First	Endothelial	leukocyte activation	0.0e+00	0.0000000	0.3001723
First	Endothelial	cell activation	0.0e+00	0.0000000	0.2942748
First	Endothelial	cell activation involved in immune response	0.0e+00	0.0000000	0.3151694
First	Endothelial	leukocyte activation involved in immune response	0.0e+00	0.0000000	0.3125926
First	Endothelial	immune effector process	0.0e+00	0.0000000	0.2852650
First	Endothelial	myeloid leukocyte activation	0.0e+00	0.0000000	0.3160000
First	Endothelial	cytokine production	0.0e+00	0.0000002	0.2987128
First	Endothelial	actin cytoskeleton organization	0.0e+00	0.0000004	0.3779528
First	Hofbauer	immune response	0.0e+00	0.0000000	0.2992444
First	Hofbauer	leukocyte activation	0.0e+00	0.0000000	0.3323285
First	Hofbauer	immune system process	0.0e+00	0.0000000	0.2855419
First	Hofbauer	cell activation	0.0e+00	0.0000000	0.3243003
First	Hofbauer	cell activation involved in immune response	0.0e+00	0.0000000	0.3532155
First	Hofbauer	leukocyte activation involved in immune response	0.0e+00	0.0000000	0.3508642
First	Hofbauer	vesicle-mediated transport	0.0e+00	0.0000000	0.3203248
First	Hofbauer	immune effector process	0.0e+00	0.0000000	0.3083258
First	Hofbauer	regulation of immune system process	0.0e+00	0.0000000	0.3007246
First	Hofbauer	myeloid leukocyte activation	0.0e+00	0.0000000	0.3429333
First	Stromal	actin filament-based process	0.0e+00	0.0000006	0.3862069
First	Stromal	cytosol	0.0e+00	0.0000006	0.2695285
First	Stromal	actin cytoskeleton organization	0.0e+00	0.0000010	0.3937008
First	Stromal	enzyme binding	0.0e+00	0.0000076	0.3065025
First	Stromal	small GTPase mediated signal transduction	0.0e+00	0.0000222	0.3928571
First	Stromal	regulation of small GTPase mediated signal transduction	0.0e+00	0.0000340	0.4492063
First	Stromal	vesicle-mediated transport	0.0e+00	0.0001380	0.2893312
First	Stromal	guanyl-nucleotide exchange factor activity	1.0e-07	0.0001730	0.4733010
First	Stromal	GTPase binding	1.0e-07	0.0001730	0.3895582
First	Stromal	nucleoside-triphosphatase regulator activity	1.0e-07	0.0001730	0.4150641
Term	Endothelial	cytosol	0.0e+00	0.0000061	0.2534909
Term	Endothelial	enzyme binding	0.0e+00	0.0000722	0.2891020
Term	Endothelial	intracellular signal transduction	1.0e-07	0.0005720	0.2658375
Term	Endothelial	adherens junction	2.0e-07	0.0007310	0.3707224
Term	Endothelial	nucleoplasm	2.0e-07	0.0007310	0.2526930
Term	Endothelial	anchoring junction	2.0e-07	0.0007310	0.3659889
Term	Endothelial	macromolecule modification	4.0e-07	0.0012195	0.2480502
Term	Endothelial	regulation of GTPase activity	7.0e-07	0.0020470	0.3678161
Term	Endothelial	cellular protein modification process	1.7e-06	0.0038387	0.2502562
Term	Endothelial	protein modification process	1.7e-06	0.0038387	0.2502562
Term	Hofbauer	immune response	0.0e+00	0.0000000	0.2648041
Term	Hofbauer	leukocyte activation	0.0e+00	0.0000001	0.2941430
Term	Hofbauer	regulation of immune system process	0.0e+00	0.0000001	0.2819293
Term	Hofbauer	cell activation	0.0e+00	0.0000002	0.2889313
Term	Hofbauer	immune system process	0.0e+00	0.0000002	0.2590514
Term	Hofbauer	positive regulation of immune system process	0.0e+00	0.0000006	0.2936275
Term	Hofbauer	myeloid leukocyte activation	0.0e+00	0.0000024	0.3096000
Term	Hofbauer	positive regulation of cell adhesion	0.0e+00	0.0000027	0.3814433
Term	Hofbauer	intracellular signal transduction	0.0e+00	0.0000035	0.2766643
Term	Hofbauer	regulation of cell adhesion	0.0e+00	0.0000035	0.3482549
Term	Stromal	actin filament-based process	0.0e+00	0.0004610	0.3489655
Term	Stromal	collagen-containing extracellular matrix	0.0e+00	0.0005570	0.3428571
Term	Stromal	extracellular matrix	5.0e-07	0.0034671	0.3153693
Term	Stromal	collagen binding	9.0e-07	0.0050212	0.5447761
Term	Stromal	actin cytoskeleton organization	1.8e-06	0.0081906	0.3448819

Term	Stromal	actin cytoskeleton	4.9e-06	0.0184389	0.3512873
Term	Stromal	anchoring junction	9.1e-06	0.0296833	0.3505853

Table B.2: KEGG enrichment for cell DMCs.

Group	Celltype	Description	P.DE	FDR
First	Endothelial	Cytokine-cytokine receptor interaction	0.0041570	0.0483066
First	Endothelial	Pathways in cancer	0.0035835	0.0483066
First	Endothelial	Shigellosis	0.0002610	0.0147357
First	Endothelial	Natural killer cell mediated cytotoxicity	0.0029434	0.0472351
First	Endothelial	Chemokine signaling pathway	0.0003500	0.0147357
First	Endothelial	Malaria	0.0043265	0.0486011
First	Endothelial	Proteoglycans in cancer	0.0040161	0.0483066
First	Endothelial	TGF-beta signaling pathway	0.0025529	0.0441775
First	Endothelial	Chagas disease	0.0010931	0.0263131
First	Endothelial	Relaxin signaling pathway	0.0044760	0.0486581
First	Endothelial	Th17 cell differentiation	0.0046636	0.0491133
First	Endothelial	Osteoclast differentiation	0.0005920	0.0166272
First	Endothelial	HIF-1 signaling pathway	0.0041166	0.0483066
First	Endothelial	Autophagy - animal	0.0015393	0.0324209
First	Endothelial	Rap1 signaling pathway	0.0005400	0.0165576
First	Endothelial	PD-L1 expression and PD-1 checkpoint pathway in cancer	0.0019094	0.0378519
First	Endothelial	Focal adhesion	0.0021166	0.0396271
First	Endothelial	Regulation of actin cytoskeleton	0.0000474	0.0053291
First	Endothelial	Leukocyte transendothelial migration	0.0002970	0.0147357
First	Endothelial	Phospholipase D signaling pathway	0.0038427	0.0483066
First	Endothelial	Inflammatory bowel disease	0.0004180	0.0156348
First	Endothelial	Platelet activation	0.0003410	0.0147357
First	Endothelial	Fc gamma R-mediated phagocytosis	0.0041153	0.0483066
First	Endothelial	Parathyroid hormone synthesis, secretion and action	0.0039164	0.0483066
First	Endothelial	Cholinergic synapse	0.0034570	0.0483066
First	Endothelial	AGE-RAGE signaling pathway in diabetic complications	0.0006890	0.0178643
First	Endothelial	Colorectal cancer	0.0013794	0.0309896
First	Endothelial	Fc epsilon RI signaling pathway	0.0026218	0.0441775
First	Endothelial	Yersinia infection	0.0000064	0.0021490
First	Endothelial	Pancreatic cancer	0.0005070	0.0165576
First	Endothelial	Acute myeloid leukemia	0.0001620	0.0136244
First	Endothelial	Adherens junction	0.0000199	0.0033455
First	Hofbauer	Chemokine signaling pathway	0.0005350	0.0257466
First	Hofbauer	Human T-cell leukemia virus 1 infection	0.0014401	0.0478661
First	Hofbauer	Osteoclast differentiation	0.0010920	0.0408882
First	Hofbauer	Apelin signaling pathway	0.0015624	0.0478661
First	Hofbauer	Yersinia infection	0.0001490	0.0167683
First	Hofbauer	T cell receptor signaling pathway	0.0004130	0.0231995
First	Hofbauer	Autophagy - animal	0.0000555	0.0093599
First	Hofbauer	B cell receptor signaling pathway	0.0003330	0.0224440
First	Hofbauer	Acute myeloid leukemia	0.0010745	0.0408882
First	Hofbauer	Bacterial invasion of epithelial cells	0.0002930	0.0224440
First	Hofbauer	Fc gamma R-mediated phagocytosis	0.0000210	0.0070602
First	Stromal	Shigellosis	0.0001850	0.0172432
First	Stromal	Chemokine signaling pathway	0.0004410	0.0297071
First	Stromal	Ras signaling pathway	0.0006990	0.0336638
First	Stromal	Endocytosis	0.0000622	0.0104790
First	Stromal	Rap1 signaling pathway	0.0002050	0.0172432
First	Stromal	Focal adhesion	0.0012680	0.0427305
First	Stromal	AGE-RAGE signaling pathway in diabetic complications	0.0009770	0.0365750
First	Stromal	Yersinia infection	0.0000312	0.0104790

First	Stromal	Longevity regulating pathway	0.0005360	0.0300840
First	Stromal	EGFR tyrosine kinase inhibitor resistance	0.0008950	0.0365750
First	Trophoblasts	ECM-receptor interaction	0.0000133	0.0044964
Term	Endothelial	Adherens junction	0.0000515	0.0173604
Term	Hofbauer	Cytokine-cytokine receptor interaction	0.0004400	0.0494750
Term	Hofbauer	Apelin signaling pathway	0.0002600	0.0437646
Term	Hofbauer	Hematopoietic cell lineage	0.0000158	0.0053238
Term	Trophoblasts	Regulation of actin cytoskeleton	0.0000278	0.0093774

Table B.3: List of imprinted genes and regions from multiple studies.

position	Methylated allele	Tissue	Gene	publication
1:6684860-6685996	M	placental-specific	THAP3	hanna
1:7827139-7827709	M	other		zink
1:19614429-19615702	M	placental-specific	AKR7A3	hanna
1:36184400-36184863	M	placental-specific	C1orf216	hanna
1:38200920-38201123	M	other	EPHA10	hanna
1:39559602-39559980	M	other	PPIEL	zink
1:40024626-40026520	M	other	PPIEL	court, hanna
1:54940382-54941170	M	placental-specific	ACOT11	hanna
1:55504848-55506512	M	placental-specific	PCSK9	hanna
1:67772896-67773725	M	placental-specific	IL12RB2	hanna
1:68050532-68051790	M	other	DIRAS3/GNG12-AS1	zink
1:68511835-68513486	M	other	DIRAS3	court, hanna
1:68515433-68517545	M	other	DIRAS3	court, hanna
1:175568216-175568710	M	placental-specific	TNR	hanna
1:177032657-177032837	M	other		zink
1:181286640-181287967	M	placental-specific	CACNA1E	hanna, hamada
1:209848306-209849445	M	placental-specific	G0S2	hanna, hamada, hanna, hamada
1:211589678-211590292	M	placental-specific	LINC00467	hanna
2:27484942-27488313	M	placental-specific	SLC30A3	hanna, hamada
2:42067938-42068648	M	placental-specific	C2orf91	hanna, hamada
2:45231226-45232888	M	placental-specific	SIX2	hanna
2:46655894-46657384	M	placental-specific	TMEM247	sanchez delgado, hamada
2:74345587-74348298	M	placental-specific	TET3	sanchez delgado, hamada
2:94871304-94872050	P	other		zink
2:130037382-130038260	M	other		zink
2:131300169-131300522	M	other		zink
2:136766706-136767019	M	other		zink
2:196933266-196934154	M	placental-specific	DNAH7	hanna
2:206249787-206253341	P	other	ZDBF2/GPR1-AS	zink
2:206254404-206255173	P	other	ZDBF2	zink
2:206255756-206258686	P	other	ZDBF2	zink
2:206259834-206260336	P	other	ZDBF2	zink
2:206261783-206265613	P	other	ZDBF2/GPR1-AS	zink
2:206266861-206268649	P	other	ZDBF2	zink
2:206269197-206270970	P	other	ZDBF2	zink
2:206271479-206272728	P	other	ZDBF2	zink
2:206273392-206274051	P	other	ZDBF2	zink

2:207066967-207069445	M	placental-specific	GPR1-AS	court, hanna
2:207114583-207136544	P	other		court
2:209208321-209209864	M	other		zink
2:229043530-229047055	M	placental-specific	SPHKAP	hanna, sanchez delgado, hamada, hanna, sanchez delgado, hamada
2:232351651-232352005	M	other		zink
2:240557132-240557566	M	other		zink
3:11704930-11705437	M	other		zink
3:21790873-21792537	M	placental-specific	ZNF385D	sanchez delgado, hamada
3:30894288-30894468	M	other		zink
3:30936070-30936531	M	placental-specific	GADL1	hanna
3:39501848-39502216	M	other		zink
3:49314155-49314920	M	placental-specific	C3orf62	hanna, hamada
3:51740741-51741473	M	other	GRM2	hanna
3:128336483-128337044	M	placental-specific	RPN1	hanna, hamada
3:128564782-128565090	M	other	RAB7A	hanna
3:182815725-182817627	M	placental-specific	MCCC1	court, hanna
3:192124589-192127457	M	placental-specific	FGF12	hanna, sanchez delgado, hamada
3:192571291-192571834	M	other		zink
3:196756621-196756875	M	other	MFI2	hanna
4:4576220-4577911	M	placental-specific	STX18-AS1	sanchez delgado
4:6104931-6106089	M	other	JAKMIP1	zink
4:6107021-6107791	M	other	JAKMIP1	hanna
4:8580973-8581784	M	other		zink
4:17641982-17642265	M	other		zink
4:88697379-88698217	M	other	NAP1L5/HERC3	zink
4:89617925-89619237	M	other	NAP1L5	court, hanna
4:93226245-93227270	M	placental-specific	GRID2	hanna, hamada, hanna, hamada
4:102711702-102712397	M	placental-specific	BANK1	hanna
4:119455534-119455914	M	other		zink
4:121932574-121933160	M	other		zink
4:154709200-154715220	M	placental-specific	SFRP2	sanchez delgado, hamada, sanchez delgado, hamada
4:169774385-169774935	M	other		zink
4:187065417-187066505	M	placental-specific	FAM149A	hanna
5:1594021-1595048	M	placental-specific	SDHAP3	hanna, hamada
5:15271459-15271770	M	other		zink
5:15384409-15384881	M	other		zink
5:58333774-58336554	M	placental-specific	PDE4D	court, hanna
5:95066568-95068092	M	placental-specific	RHOBTB3	hanna

5:135767915-135768919	M	other		zink
5:135769815-135771319	M	other		zink
5:135778208-135778521	M	other		zink
5:136078784-136080957	M	other	VTRNA2-1	zink
5:178593785-178594990	M	other	ZNF354C	hanna
6:3848512-3850359	M	other	FAM50B	court, hanna, zink
6:14117480-14118415	M	placental-specific	CD83	hanna
6:18387077-18387809	M	other	RNF144B	hanna
6:31627653-31628935	M	other	C6orf47	hanna
6:37616410-37617124	M	placental-specific	MDGA1	hanna
6:39901897-39902693	M	placental-specific	MOCS1	hanna
6:103332287-103333035	M	other		zink
6:105400631-105402559	M	placental-specific	LIN28B	court, hanna
6:106957945-106961974	M	placental-specific	AIM1	court, hanna
6:144007163-144009025	M	other	PLAGL1	zink
6:144328078-144329922	M	other	PLAGL1	court, hanna
6:160005255-160006715	M	other	IGF2R	zink
6:160426558-160427561	M	other		court
6:161188022-161188822	M	placental-specific	PLG	hanna
6:169577181-169577797	M	other		zink
6:169654367-169655912	M	other	WDR27	zink
6:170054504-170055618	M	other		court
7:12609907-12610833	M	placental-specific	SCIN	hanna
7:16850625-16851508	M	other		zink
7:22122473-22123315	M	placental-specific	RAPGEF5	hanna, hamada
7:23490380-23491301	M	other	RPS2P32	zink
7:24323128-24325371	M	placental-specific	NPY	hanna, hamada, hanna, hamada
7:42856558-42857371	M	other		zink
7:43151828-43153950	M	placental-specific	HECW1	hanna
7:50781399-50783614	M	other	GRB10	zink
7:50848726-50851312	M	other		court
7:64575091-64575739	M	other		zink
7:81240257-81240667	M	other	HGF	hanna
7:94285501-94287960	M	other	PEG10/SGCE	court, hanna
7:100091181-100091786	M	placental-specific	NYAP1	hanna
7:101006052-101006963	M	placental-specific	EMID2	hanna, hamada
7:106300098-106302548	M	placental-specific	CCDC71L	hanna, hamada, hanna, hamada
7:130130122-130134388	M	other	MEST/MESTIT1	court, hanna
7:130489758-130494000	M	other	MEST	zink
7:134671024-134672011	M	placental-specific	AGBL3	court, hanna
7:134831752-134832178	M	other		zink
7:138664014-138664236	M	other	SVOPL	zink

7:149389444-149389941	M	other	KRBA1	hanna, hamada
7:154585539-154586375	M	other	DPP6	hanna
7:154861569-154863382	M	other	HTR5A	court, hanna
7:155070814-155072164	M	other	HTR5A	zink
8:8702335-8703302	M	other		zink
8:11659497-11660209	M	other	FDFT1	hanna
8:22010755-22011096	M	other		zink
8:23145610-23146931	M	placental-specific	R3HCC1	sanchez delgado
8:27182871-27183342	M	other	PTK2B	hanna
8:37604992-37606088	M	other	ERLIN2	court, hanna
8:61626185-61627281	M	other	CHD7	hanna, hamada
8:94119077-94120749	M	other		zink
8:102527935-102529486	M	other		zink
8:135707227-135710114	M	placental-specific		court
8:140035407-140035811	P	other		zink
8:140038516-140038940	P	other		zink
8:140085725-140086791	P	other		zink
8:140088744-140088969	P	other		zink
8:140097560-140101293	M	other	TRAPPC9	zink
8:141107717-141111081	M	other	TRAPPC9	court, hanna
8:142215227-142216514	M	placental-specific	DENND3	sanchez delgado, hamada, sanchez delgado, hamada
9:73568-73835	M	placental-specific	PGM5P3-AS1	hanna
9:4297279-4300432	M	placental-specific	GLIS3	court, hanna
9:34989434-34989605	M	placental-specific	DNAJB5	hanna
9:37800140-37802937	M	placental-specific	DCAF10	court, hanna
9:86136501-86137778	M	placental-specific	FRMD3	sanchez delgado
9:86151350-86154260	M	placental-specific	FRMD3	sanchez delgado, hamada
9:137416835-137418376	M	other		zink
9:140301079-140302117	M	placental-specific	EXD3	hanna
10:11936672-11937255	M	placental-specific	PROSER2-AS1	hanna, hamada, hanna, hamada
10:15761192-15762312	M	placental-specific	ITGA8	hanna, hamada
10:27702309-27703547	M	other	PTCHD3	hanna, hamada, hanna, hamada
10:65224441-65225999	M	placental-specific	JMJD1C	hanna, hamada
10:103534501-103536348	M	placental-specific	FGF8	hanna
10:121577530-121578846	M	other	INPP5F	court, hanna
10:128993405-128995242	M	placental-specific	FAM196A/DOCK1	court, hanna
10:130190919-130191634	M	other		zink
10:135278717-135279147	M	other	SPRN	hanna
10:135341528-135343280	M	placental-specific	CYP2E1	hanna, hamada
11:2016513-2024740	P	other	H19	court, hanna

11:2132351-2133881	P	other	IGF2	zink
11:2145821-2149292	P	other	IGF2	zink
11:2153991-2155112	P	other		court
11:2168333-2169768	P	other		court
11:2490964-2492685	M	placental-specific	KCNQ1	sanchez delgado
11:2699100-2700636	M	other	KvDMR1	zink
11:2719948-2722440	M	other	KCNQ1OT1	court, hanna
11:3662967-3663842	M	other	ART5	hanna
11:7088843-7089383	M	other		zink
11:10562070-10563302	M	placental-specific	RNF141	hanna
11:19366443-19368277	M	placental-specific	NAV2	hanna
11:45201741-45202557	M	placental-specific	PRDM11	hanna
11:45921134-45922184	M	placental-specific	MAPK8IP1	hanna
11:60647012-60647445	M	other		zink
11:68451396-68452097	M	placental-specific	GAL	hanna
11:109962727-109964976	M	placental-specific	ZC3H12C	court, hanna
12:203429-204151	M	other		zink
12:2800562-2800919	M	placental-specific	CACNA1C	sanchez delgado
12:4433587-4433983	M	placental-specific	C12orf5	hanna, hamada
12:22487219-22488465	M	placental-specific	ST8SIA1	hanna
12:34218926-34219575	P	other		zink
12:34362974-34363820	M	other		zink
12:34372206-34374112	M	other		zink
12:65121923-65122279	M	other		zink
12:65218069-65218869	M	placental-specific	TBC1D30	hanna
13:33000694-33002597	M	placental-specific	N4BP2L1	court, hanna
13:48317165-48320825	M	other	RB1	zink
13:48892341-48895763	M	other		court
13:51417469-51418614	M	other	DLEU7	hanna
13:60267520-60269245	M	other		zink
13:70680712-70683111	M	other	KLHL1	hanna
13:80654825-80655272	M	other		zink
13:102568126-102569981	M	placental-specific	FGF14	sanchez delgado, hamada
14:24563095-24564067	M	placental-specific	NRL	hanna, hamada
14:33799741-33800646	M	other		zink
14:52734156-52736420	M	placental-specific	PTGDR	hanna, hamada
14:68874777-68874992	M	other		zink
14:100727514-100728411	M	other		zink
14:100807670-100811737	P	other	IG-DMR	zink
14:100823704-100825381	P	other	MEG3	zink
14:100826012-100828230	P	other	MEG3	zink
14:100835999-100836272	M	other		zink

14:100900401-100901267	M	other		zink
14:100904519-100905081	M	other	MEG8	zink
14:101067986-101069282	M	other		zink
14:101072233-101073637	M	other		zink
14:101151414-101151997	P	other		zink
14:101275427-101278058	P	other		court
14:101290524-101293978	P	other	MEG3	court, hanna
14:101370741-101371419	M	other		court
14:101635099-101635431	P	other		zink
14:105830606-105830859	M	other	PACS2	hanna
15:23364409-23364521	M	other		zink
15:23534531-23535315	P	other		zink
15:23606638-23609456	P	other		zink
15:23629039-23629213	M	other		zink
15:23634077-23634289	M	other		zink
15:23642878-23643103	M	other		zink
15:23647776-23648881	M	other	MAGEL2	zink
15:23661255-23662817	P	other		zink
15:23674115-23675360	P	other		zink
15:23686304-23688131	M	other	NDN	zink
15:23769176-23769945	P	other		zink
15:23782845-23783768	P	other		zink
15:23797680-23798290	M	other		zink
15:23807086-23812495	M	other		court
15:23829311-23829706	P	other		zink
15:23854644-23855506	M	other		zink
15:23857016-23857880	M	other		zink
15:23858683-23861887	M	other		zink
15:23869240-23869921	M	other		zink
15:23877454-23878654	M	other		zink
15:23883075-23883432	P	other		zink
15:23892425-23894029	M	other		court
15:23896280-23898594	M	other	PWRN4	zink
15:23908972-23909688	P	other		zink
15:23914065-23915807	P	other		zink
15:23922214-23923177	P	other		zink
15:23931257-23933138	P	other		court, zink
15:23939735-23940870	P	other		zink
15:23967138-23967845	P	other		zink
15:23968767-23970583	P	other		zink
15:23971255-23971960	P	other		zink
15:23975861-23977507	P	other		zink

15:23984693-23986673	P	other		zink
15:23999391-24001886	P	other		zink
15:24004599-24005919	P	other		zink
15:24009020-24009903	P	other		zink
15:24019161-24019660	P	other		zink
15:24029218-24029630	P	other		zink
15:24050087-24051502	P	other		zink
15:24086139-24087500	P	other		zink
15:24101056-24101995	M	other	SNRPN	zink
15:24111944-24113625	P	other		zink
15:24115509-24117169	P	other		zink
15:24132683-24133507	P	other		zink
15:24156357-24157105	P	other		zink
15:24163134-24163879	P	other		zink
15:24174217-24175175	P	other		zink
15:24225156-24227501	P	other		zink
15:24239642-24241406	P	other		zink
15:24252735-24254128	P	other		zink
15:24274859-24275806	P	other		zink
15:24300755-24301849	P	other		zink
15:24346736-24347142	M	other		court
15:24406235-24407327	P	other		zink
15:24416085-24417692	P	other		zink
15:24426478-24427389	M	other	SNRPN	zink
15:24535702-24537396	P	other		zink
15:24551067-24552042	P	other		zink
15:24552829-24553614	P	other		zink
15:24561446-24562639	P	other		zink
15:24566428-24567551	P	other		zink
15:24576113-24576926	P	other		zink
15:24578494-24579011	P	other		zink
15:24606385-24608581	P	other		zink
15:24648973-24649668	P	other		zink
15:24671872-24672679	M	other		court
15:24689074-24690306	P	other		zink
15:24722753-24723071	M	other		court
15:24733127-24735542	P	other		zink
15:24747566-24748890	P	other		zink
15:24768838-24769120	P	other		zink
15:24772760-24773758	M	other	SNRPN	zink
15:24784260-24785474	M	other		zink
15:24809111-24810320	P	other		zink

15:24816493-24817798	P	other		zink
15:24823417-24824759	M	other	SNRPN	zink
15:24828587-24829569	P	other		zink
15:24833609-24834203	P	other		zink
15:24846843-24848681	M	other	SNRPN	zink
15:24856474-24859220	M	other	SNRPN	zink
15:24877552-24880264	M	other	SNRPN	zink
15:24885960-24886572	P	other		zink
15:24887240-24888535	P	other		zink
15:24901543-24903906	P	other		zink
15:24910643-24911281	P	other		zink
15:24949168-24949368	P	other		zink
15:24954493-24957248	M	other	SNURF/SNRPN	zink
15:24996577-24998201	P	other		zink
15:25000817-25000975	P	other		zink
15:25005051-25005302	P	other		zink
15:25015223-25015809	P	other		zink
15:25017924-25018886	M	other		court
15:25039354-25039564	P	other		zink
15:25048423-25049425	P	other		zink
15:25060537-25061269	P	other		zink
15:25062432-25064422	P	other		zink
15:25064951-25066060	P	other		zink
15:25066776-25067078	P	other		zink
15:25068564-25069481	M	other		court
15:25075739-25076861	P	other	SNORD116-10,SNORD116-11,SNORD116-12,SNORD116-13,SNORD116-14,SNORD116-15,SNORD116-16,SNORD116-17,SNORD116-19,SNORD116-19,SNORD116-18,SNORD116-20,SNORD116-21,SNORD116-22,SNORD116-23,SNORD116-24	zink

15:25080754-25081544	P	other	SNORD116-10,SNORD116-11,SNORD116-12,SNORD116-13,SNORD116-14,SNORD116-15,SNORD116-16,SNORD116-17,SNORD116-19,SNORD116-19,SNORD116-18,SNORD116-20,SNORD116-21,SNORD116-22,SNORD116-23,SNORD116-24	zink
15:25083328-25085575	P	other	SNORD116-10,SNORD116-11,SNORD116-12,SNORD116-13,SNORD116-14,SNORD116-15,SNORD116-16,SNORD116-17,SNORD116-19,SNORD116-19,SNORD116-18,SNORD116-20,SNORD116-21,SNORD116-22,SNORD116-23,SNORD116-24	zink
15:25093008-25093829	M	other		court
15:25123027-25123905	M	other		court
15:25200004-25201976	M	other	SNURF/SNRPN	court, hanna
15:40486055-40486903	M	other		zink
15:45314789-45315642	M	placental-specific	SORD	hanna
15:50909603-50909875	M	other		zink
15:76030565-76031591	M	other	DNM1P35	hanna
15:79382548-79383980	M	placental-specific	RASGRF1	hanna
15:93614758-93616859	M	placental-specific	RGMA	court, hanna
15:98865575-98867104	M	other	IGF1R	zink
15:99408496-99409650	M	other		court
15:101626335-101626824	M	placental-specific	LRRK1	hanna
16:806879-808764	M	other		zink
16:817075-818443	M	other		zink

16:863240-863884	M	placental-specific	PRR25	hanna
16:3254405-3254770	P	other		zink
16:3364797-3366488	P	other	MTRNR2L4	zink
16:3413769-3414262	P	other		zink
16:3431450-3432387	M	other	ZNF597	zink
16:3434920-3435926	M	other		zink
16:3442978-3444462	P	other	ZNF597/NAA60/ZNF597,NAA60	zink
16:3481801-3482388	M	other		court
16:3492724-3494463	P	other	ZNF597/NAA60	court, hanna
16:10707754-10708438	M	other		zink
16:30816719-30817779	M	placental-specific	ZNF629	hanna, hamada
16:48399731-48400475	M	placental-specific	SIAH1	hanna
16:58534681-58535556	M	placental-specific	NDRG4	hanna
16:66637919-66639593	M	placental-specific	CMTM3	hanna
16:68572892-68573971	M	placental-specific	ZFP90	hanna
16:74734230-74734885	M	other	MLKL	hanna, hamada
16:78079569-78080193	M	other	CLEC3A	hanna
17:259426-260589	M	placental-specific	C17orf97	hanna
17:409552-410094	M	other		zink
17:4900360-4902310	M	other		zink
17:56609082-56609687	M	placental-specific	12:00 AM	hanna
17:66596155-66597643	M	placental-specific	FAM20A	court, hanna
18:32956510-32957683	M	placental-specific	ZNF396	court, hanna
18:47667786-47668339	M	other		zink
18:60051870-60052464	M	other	TNFRSF11A	hanna
18:79899277-79899991	M	other		zink
18:80147168-80149255	M	other		zink
19:1324834-1325348	M	other	MUM1	hanna
19:10303506-10306415	M	placental-specific	DNMT1	court, hanna
19:11784246-11785337	M	placental-specific	ZNF833P	hanna
19:12075601-12076549	M	other	ZNF763	hanna
19:13614882-13618186	M	placental-specific	CACNA1A	sanchez delgado
19:17438249-17439339	M	other	ANO8	hanna
19:21482472-21484491	M	other		zink
19:36266285-36266855	M	other		zink
19:38527504-38528578	M	other		zink
19:38543710-38544472	M	other		zink
19:53536955-53539153	M	other	ZNF331	zink
19:53553367-53555638	M	other	ZNF331	zink
19:54040510-54042212	M	other	ZNF331	court, hanna
19:54057086-54058425	M	other		court
19:54150515-54155608	M	placental-specific	C19MC	court, hanna

19:56725960-56726738	M	other		zink
19:56829669-56829913	P	other		zink
19:56837802-56840915	M	other	PEG3	zink
19:56864809-56865037	M	other		zink
19:57348493-57353271	M	other	PEG3	court, hanna
19:58055058-58055675	P	other		zink
20:30134663-30135933	M	other	MCTS2P/HM13	court, hanna
20:30618874-30619244	M	placental-specific	C20orf160	hanna
20:31547246-31548401	M	other	MCTS2P/HM13/HM13	zink
20:34638489-34639686	M	other	LINC00657	hanna
20:36147042-36150528	M	other	BLCAP/NNAT	court, hanna
20:37519780-37521655	M	other	BLCAP/NNAT/BLCAP,NNAT	zink
20:42142005-42144040	M	other	L3MBTL1	court, hanna
20:43514638-43515399	M	other	L3MBTL/L3MBTL1	zink
20:48384392-48385349	M	other		zink
20:52789646-52791472	M	placental-specific	CYP24A1	hanna, hamada
20:57413694-57418612	P	other	GNAS	court, hanna
20:57425157-57428033	M	other	GNAS-AS1	court, hanna
20:57428905-57431463	M	other	GNAS XL	court, hanna
20:57463265-57465201	M	other		court
20:58838984-58843556	P	other	GNAS/GNAS-AS1	zink
20:58846210-58846670	P	other		zink
20:58850116-58853207	M	other	NESP-AS/GNAS-AS1/GNAS-AS1,GNAS	zink
20:58854498-58856228	M	other	GNAS/GNAS-AS1,GNAS	zink
20:58888210-58889085	M	other	GNAS/LOC101927932,GNAS	zink
20:58889838-58890443	M	other	GNAS/LOC101927932,GNAS	zink
20:59526001-59526478	M	other		zink
20:60540388-60541082	M	placental-specific	TAF4	hanna
20:63938417-63938769	M	other		zink
21:39385677-39386349	M	other	WRB	zink
21:39387837-39388701	M	other		zink
21:40757510-40758276	M	other		court
21:42218551-42219853	M	placental-specific	TMPRSS3	hanna, hamada
21:46661264-46661774	M	other		zink
21:48087452-48088150	M	placental-specific	PRMT2	hanna
22:19973978-19974866	M	placental-specific	ARVCF	hanna
22:32026380-32026975	M	other	PISD	hanna
22:37464839-37465279	M	placental-specific	KCTD17	hanna
22:38412407-38412798	M	other		zink
22:40057496-40061223	M	placental-specific	CACNA1I	sanchez delgado

22:41681734-41682868	M	other	NHP2L1/SNU13	zink
22:42077774-42078873	M	other	NHP2L1	court, hanna

Table B.4: Linear modelling results of repetitive element methylation and cell type.

Trimester	test	Tissue	estimate	p.value	p_value	p_value_adj	p_value_adj_label
Third	Alu	Trophoblasts cs	-0.0117235	0.0008400	<0.001	0.0336012	0.034
Third	Alu	Stromal cs	-0.0051651	0.1313044	0.131	1.0000000	>0.999
Third	Alu	Hofbauer cs	0.0696154	0.0000000	<0.001	0.0000000	<0.001
Third	Alu	Endothelial cs	0.0207889	0.0000000	<0.001	0.0000009	<0.001
Third	L1	Trophoblasts cs	-0.0144791	0.0016805	0.002	0.0672202	0.067
Third	L1	Stromal cs	0.0109213	0.0164964	0.016	0.6598560	0.66
Third	L1	Hofbauer cs	0.0970748	0.0000000	<0.001	0.0000000	<0.001
Third	L1	Endothelial cs	-0.0033859	0.4506288	0.451	1.0000000	>0.999
Third	REMP_Alu	Trophoblasts cs	0.0071025	0.1415773	0.142	1.0000000	>0.999
Third	REMP_Alu	Stromal cs	0.0123476	0.0115689	0.012	0.4627559	0.463
Third	REMP_Alu	Hofbauer cs	0.0495099	0.0000000	<0.001	0.0000000	<0.001
Third	REMP_Alu	Endothelial cs	0.0221681	0.0000124	<0.001	0.0004980	<0.001
Third	REMP_L1	Trophoblasts cs	0.0005080	0.9248994	0.925	1.0000000	>0.999
Third	REMP_L1	Stromal cs	0.0425323	0.0000000	<0.001	0.0000000	<0.001
Third	REMP_L1	Hofbauer cs	0.0896848	0.0000000	<0.001	0.0000000	<0.001
Third	REMP_L1	Endothelial cs	0.0296470	0.0000003	<0.001	0.0000134	<0.001
Third	all_probes	Trophoblasts cs	-0.0076146	0.0109680	0.011	0.4387200	0.439
Third	all_probes	Stromal cs	-0.0096135	0.0014818	0.001	0.0592724	0.059
Third	all_probes	Hofbauer cs	0.0496159	0.0000000	<0.001	0.0000000	<0.001
Third	all_probes	Endothelial cs	-0.0160625	0.0000004	<0.001	0.0000156	<0.001
First	Alu	Trophoblasts cs	-0.0190449	0.1848773	0.185	1.0000000	>0.999
First	Alu	Stromal cs	0.0413275	0.0019516	0.002	0.0780648	0.078
First	Alu	Hofbauer cs	0.1017086	0.0000014	<0.001	0.0000547	<0.001
First	Alu	Endothelial cs	0.0626632	0.0000256	<0.001	0.0010229	0.001
First	L1	Trophoblasts cs	-0.0308210	0.1400921	0.14	1.0000000	>0.999
First	L1	Stromal cs	0.0687587	0.0005180	<0.001	0.0207215	0.021
First	L1	Hofbauer cs	0.1439364	0.0000020	<0.001	0.0000794	<0.001
First	L1	Endothelial cs	0.0855743	0.0000557	<0.001	0.0022290	0.002
First	REMP_Alu	Trophoblasts cs	-0.0125097	0.2254359	0.225	1.0000000	>0.999
First	REMP_Alu	Stromal cs	0.0299786	0.0018370	0.002	0.0734782	0.073
First	REMP_Alu	Hofbauer cs	0.0767463	0.0000006	<0.001	0.0000256	<0.001
First	REMP_Alu	Endothelial cs	0.0472673	0.0000135	<0.001	0.0005400	<0.001
First	REMP_L1	Trophoblasts cs	-0.0337078	0.0363391	0.036	1.0000000	>0.999
First	REMP_L1	Stromal cs	0.0573872	0.0001710	<0.001	0.0068439	0.007
First	REMP_L1	Hofbauer cs	0.1254843	0.0000002	<0.001	0.0000071	<0.001
First	REMP_L1	Endothelial cs	0.0717158	0.0000135	<0.001	0.0005420	<0.001
First	all_probes	Trophoblasts cs	-0.0283173	0.0301032	0.03	1.0000000	>0.999
First	all_probes	Stromal cs	0.0326958	0.0048211	0.005	0.1928456	0.193
First	all_probes	Hofbauer cs	0.0961799	0.0000004	<0.001	0.0000178	<0.001
First	all_probes	Endothelial cs	0.0494214	0.0001100	<0.001	0.0044187	0.004

Table B.5: Performance metrics for deconvolution algorithms on in-silico mixtures.

algorithm	component	.estimate_rmse	.estimate_rsq	.estimate_mae
epidish (CBS)	Endothelial	0.0388516	0.9737324	0.0212150
epidish (CBS)	Hofbauer	0.0519962	0.9536757	0.0223685
epidish (CBS)	nRBC	0.0205382	0.9927641	0.0120923
epidish (CBS)	Stromal	0.0487289	0.9615058	0.0288258
epidish (CBS)	Syncytiotrophoblast	0.0282039	0.9862831	0.0178662
epidish (CBS)	Trophoblasts	0.0800551	0.8896896	0.0387900
epidish (CP)	Endothelial	0.0400325	0.9721659	0.0221737
epidish (CP)	Hofbauer	0.0532140	0.9510740	0.0244164
epidish (CP)	nRBC	0.0201766	0.9930678	0.0116109
epidish (CP)	Stromal	0.0474398	0.9615222	0.0280432
epidish (CP)	Syncytiotrophoblast	0.0324855	0.9816575	0.0199726
epidish (CP)	Trophoblasts	0.0821854	0.8837051	0.0411788
epidish (RPC)	Endothelial	0.0389847	0.9737806	0.0208709
epidish (RPC)	Hofbauer	0.0528783	0.9515670	0.0225016
epidish (RPC)	nRBC	0.0205564	0.9930958	0.0118227
epidish (RPC)	Stromal	0.0459021	0.9638079	0.0271040
epidish (RPC)	Syncytiotrophoblast	0.0274391	0.9869849	0.0178080
epidish (RPC)	Trophoblasts	0.0794303	0.8925289	0.0392630
Houseman (CP)	Endothelial	0.0400965	0.9720581	0.0222576
Houseman (CP)	Hofbauer	0.0533663	0.9508613	0.0247614
Houseman (CP)	nRBC	0.0197231	0.9932896	0.0113504
Houseman (CP)	Stromal	0.0474257	0.9616552	0.0280482
Houseman (CP)	Syncytiotrophoblast	0.0325610	0.9817015	0.0198326
Houseman (CP)	Trophoblasts	0.0821157	0.8843876	0.0410057

Table B.6: Statistical testing results for cell composition versus sex, ethnicity, and gestational age. Each cell type proportion was tested against each sample variable.

component	r.squared	p.value	adj_p	p_value	testing_variable
Trophoblasts	0.0414868	0.4175567	0.4462719	0.45	within Third trimester-reported - ga
Trophoblasts	0.1622335	0.0873036	0.1455060	0.15	within Third trimester-estimated - ga
Stromal	0.1051078	0.1893443	0.2524591	0.25	within Third trimester-reported - ga
Stromal	0.2097722	0.0486107	0.1240732	0.12	within Third trimester-estimated - ga
Hofbauer	0.3260707	0.0133178	0.0697651	0.07	within Third trimester-reported - ga
Hofbauer	0.3034171	0.0145180	0.0697651	0.07	within Third trimester-estimated - ga
Endothelial	0.1801355	0.0791734	0.1439517	0.14	within Third trimester-reported - ga
Endothelial	0.2081063	0.0496293	0.1240732	0.12	within Third trimester-estimated - ga
Syncytiotrophoblast	0.1843137	0.0754120	0.1439517	0.14	within Third trimester-reported - ga
Syncytiotrophoblast	0.3964107	0.0038684	0.0697651	0.07	within Third trimester-estimated - ga
Trophoblasts	0.7091501	0.0174413	0.0697651	0.07	within First trimester-reported - ga
Trophoblasts	0.2963436	0.2064526	0.2580657	0.26	within First trimester-estimated - ga
Stromal	0.6666725	0.0250298	0.0834328	0.08	within First trimester-reported - ga
Stromal	0.4579729	0.0949887	0.1461365	0.15	within First trimester-estimated - ga
Hofbauer	0.0409975	0.6632646	0.6632646	0.66	within First trimester-reported - ga
Hofbauer	0.2777403	0.2242003	0.2637651	0.26	within First trimester-estimated - ga
Endothelial	0.7158684	0.0163998	0.0697651	0.07	within First trimester-reported - ga
Endothelial	0.5375077	0.0608190	0.1351534	0.14	within First trimester-estimated - ga
Syncytiotrophoblast	0.1315468	0.4239583	0.4462719	0.45	within First trimester-reported - ga
Syncytiotrophoblast	0.3604387	0.1540597	0.2200853	0.22	within First trimester-estimated - ga
Trophoblasts	0.0102904	0.6794432	1.0000000	>0.999	Sex
Stromal	0.0552617	0.3326524	1.0000000	>0.999	Sex
Hofbauer	0.1991480	0.0554656	0.2773280	0.277	Sex
Endothelial	0.0548712	0.3344163	1.0000000	>0.999	Sex
Syncytiotrophoblast	0.0789358	0.2439496	1.0000000	>0.999	Sex
Trophoblasts	0.0594006	0.3458482	1.0000000	>0.999	ancestry
Stromal	0.0080693	0.7317058	1.0000000	>0.999	ancestry
Hofbauer	0.0009310	0.9074792	1.0000000	>0.999	ancestry
Endothelial	0.0285027	0.5171480	1.0000000	>0.999	ancestry
Syncytiotrophoblast	0.0531260	0.3734547	1.0000000	>0.999	ancestry

Appendix C - Supplementary information for Chapter 4

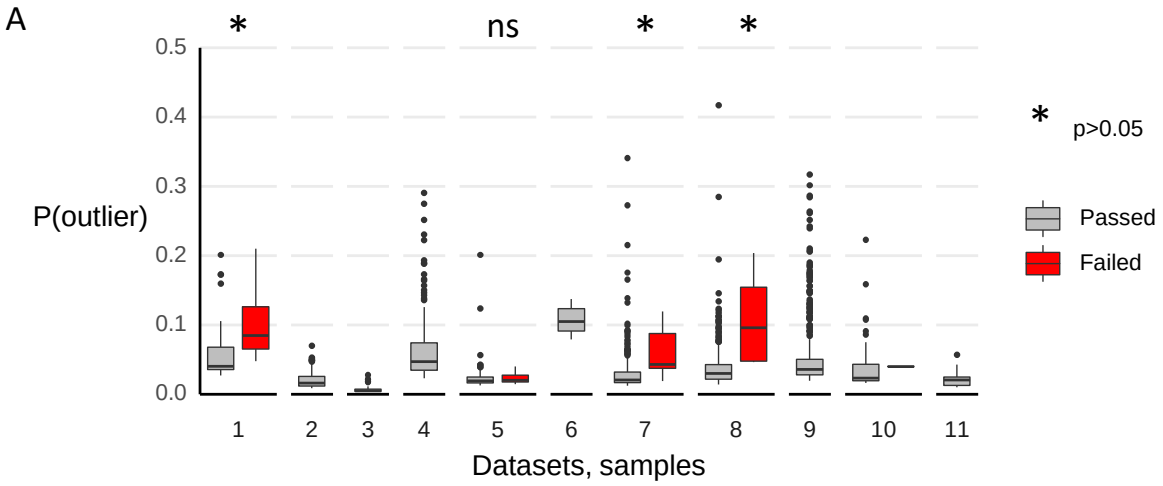

Figure C.1: Samples that failed 1 or more 27 Illumina data quality control checks had higher P(outlier) than samples that passed quality control.

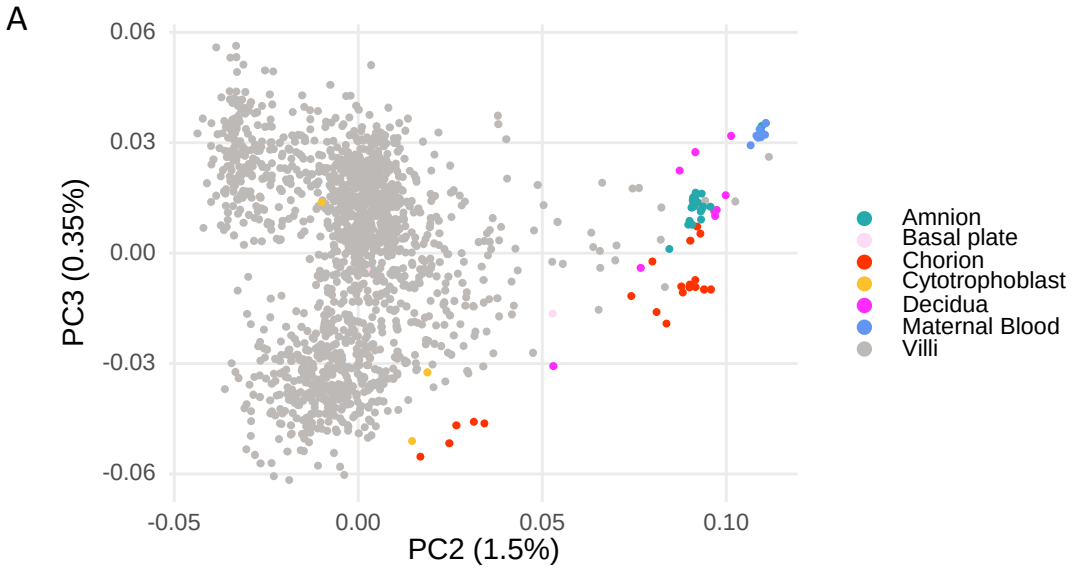

Figure C.2: Principal component analysis (PCA) using placental and maternal tissue samples.

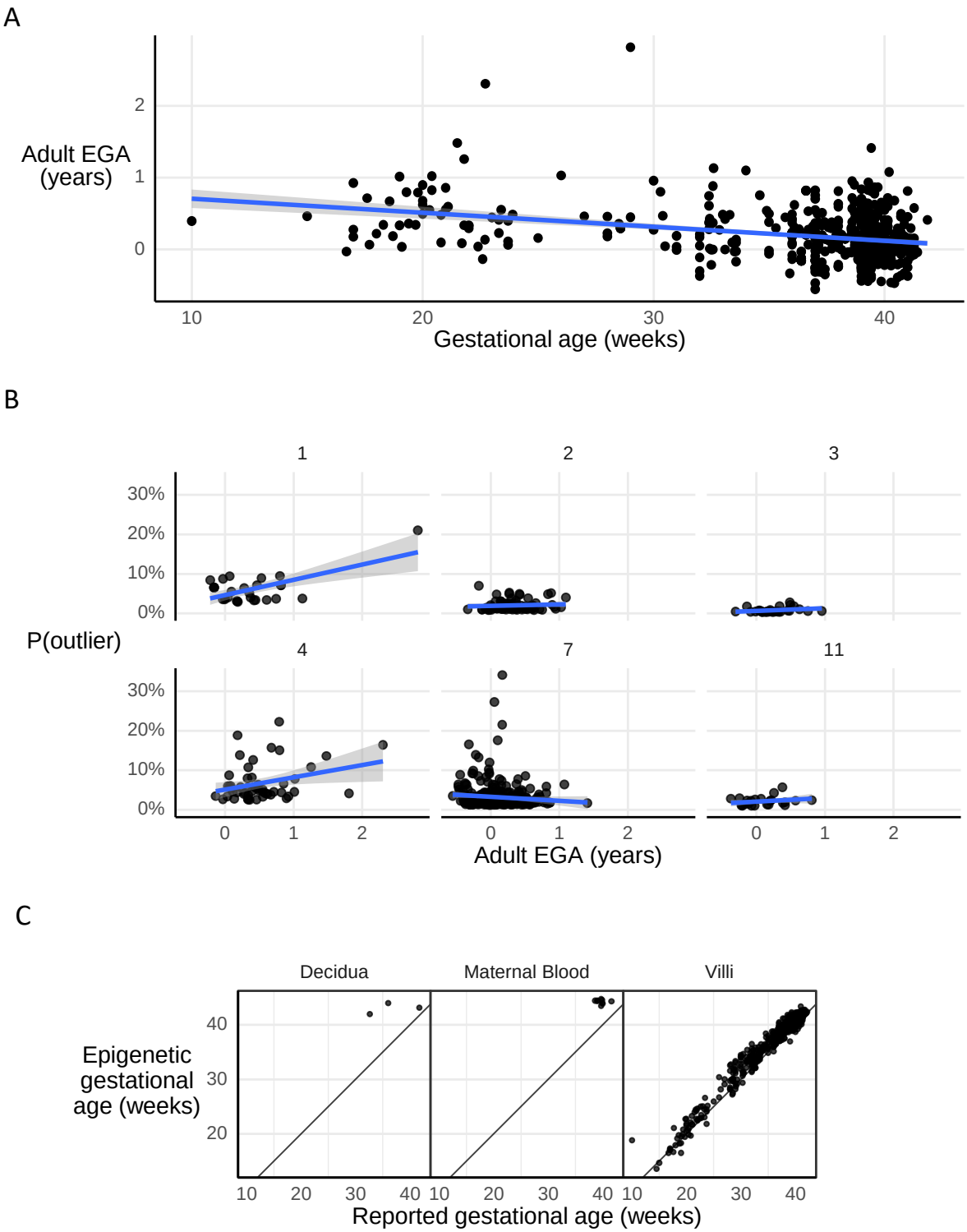

Figure C.3: Epigenetic age analysis. A) Adult epigenetic age (EA) estimated on placental samples was not associated with reported gestational age. B) Adult EA was associated with P(outlier) in 2 placental datasets. C) Placental epigenetic gestational age estimated in decidua, maternal blood, and chorionic villi samples was compared to reported gestational age.

Table C.1: Number of samples that failed quality control checks per dataset.

	Dataset											
Quality control metric	1	2	3	4	5	6	7	8	9	10	11	12
n failed	4	0	0	0	32	0	8	8	0	1	0	6
Restoration	0	0	0	0	0	0	0	0	0	0	0	0
Staining Green	0	0	0	0	0	0	0	0	0	0	0	0
Staining Red	0	0	0	0	4	0	0	1	0	0	0	0
Extension Green	0	0	0	0	0	0	0	0	0	0	0	0
Extension Red	0	0	0	0	0	0	0	0	0	0	0	0
Hybridization High/Medium	0	0	0	0	0	0	0	0	0	0	0	6
Hybridization Medium/Low	0	0	0	0	0	0	0	0	0	0	0	6
Target Removal 1	0	0	0	0	0	0	0	0	0	0	0	0
Target Removal 2	0	0	0	0	0	0	0	0	0	0	0	0
Bisulfite Conversion I Green	2	0	0	0	0	0	1	1	0	0	0	0
Bisulfite Conversion I Red	0	0	0	0	0	0	1	0	0	1	0	0
Bisulfite Conversion II	1	0	0	0	0	0	1	0	0	1	0	0
Specificity I Green	0	0	0	0	0	0	0	0	0	0	0	0
Specificity I Red	0	0	0	0	0	0	0	0	0	0	0	0
Specificity II	0	0	0	0	0	0	0	0	0	0	0	0
Non-polymorphic Green	0	0	0	0	31	0	7	2	0	0	0	0
Non-polymorphic Red	1	0	0	0	0	0	0	4	0	0	0	0