

The role of DNA methylation in gestational age



By Kristine Løkås Haftorn

Dissertation presented for the degree of Philosophical Doctor (Ph.D.)

Institute of Health and Society,

Faculty of Medicine

University of Oslo

Centre for Fertility and Health,

Norwegian Institute of Public Health

2023



© **Kristine Løkås Haftorn, 2023**

*Series of dissertations submitted to the
Faculty of Medicine, University of Oslo*

ISBN 978-82-348-0264-5

All rights reserved. No part of this publication may be reproduced or transmitted, in any form or by any means, without permission.

Cover: UiO.

Print production: Graphic center, University of Oslo.

Table of contents

Acknowledgements	4
Sammendrag	5
Abstract	7
List of papers	9
Abbreviations	10
1. Introduction	12
2. Background	13
2.1 Epigenetics	13
2.1.1 Epigenetic modifications	13
2.1.2 DNA methylation	14
2.1.3 Tissue-specificity of DNA methylation.....	16
2.1.4 Quantification of DNA methylation	16
2.1.5 Epigenetic epidemiology	16
2.2 Development	17
2.2.1 Prenatal development	17
2.2.2 Postnatal development.....	18
2.2.3 The role of DNA methylation in development	19
2.3 Assisted reproductive technology (ART) and DNA methylation	20
2.3.1 Definitions and epidemiology	20
2.3.2 The role of DNA methylation in ART.....	21
2.3.3 ART and perinatal outcomes	22
2.4 Gestational age	22
2.4.1 Definition and epidemiology	22
2.4.2 Gestational age in development and disease	23
2.4.3 Clinical determination of gestational age	24
2.4.4 Associations between DNA methylation and gestational age	24
2.5 Aging	26
2.5.1 What is aging?	26
2.5.2 The role of DNA methylation in aging.....	26
2.6 Epigenetic clocks	27
2.6.1 Using DNA methylation to predict age and gestational age.....	27
2.6.2 A brief history of epigenetic clocks.....	28
2.6.3 Epigenetic age is associated with a range of conditions and diseases	29
2.6.4 Epigenetic clocks specific for the gestational period	29
2.6.5 Epigenetic gestational age as a proxy for developmental maturity	31

3. Aims of the thesis	34
4. Methodological considerations	35
4.1 Sample collection	35
4.1.1 Datasets.....	35
4.1.2 Clinical estimations of gestational age	36
4.2 DNA methylation profiling and data preparation	37
4.2.1 BeadChip arrays for DNA methylation quantification	37
4.2.2 Quality control of microarray-derived DNA methylation data	37
4.3 Epigenome-wide association studies (EWAS)	38
4.3.1 EWAS.....	38
4.3.2 Statistical power and significance thresholds in EWAS.....	38
4.3.3 Confounding factors	39
4.4 Cell types	41
4.4.1 Cell types in cord blood.....	41
4.4.2 Cell-type deconvolution	41
4.4.3 Cell-type proportions as covariates in DNA methylation analyses	42
4.4.4 Cell-type specific analyses of DNA methylation	43
4.5 Prediction	44
4.5.1 Statistical prediction methods.....	44
4.5.2 Penalized linear regression	44
4.5.3 Elastic net and tuning of the penalty term	45
4.5.4 Assessing prediction performance and external validity	45
4.5.5 Drawbacks of penalized regression methods for variable selection	46
4.6 Stability selection	47
4.6.1 The stability selection framework	47
4.6.2 Determining the tuning parameters	48
4.7 Generalized additive models for building epigenetic clocks	48
4.8 Genome annotation & enrichment.....	49
4.8.1 Annotation methods.....	49
4.8.2 Gene-set enrichment analysis	50
5. Result summary	51
5.1 Paper 1.....	51
5.2 Paper 2.....	52
5.3 Paper 3.....	53
6. Discussion	55
6.1 Summary of key findings	55
6.2 Implications of key findings.....	55
6.2.1 Using EPIC-derived DNA methylation data for studying gestational age	55

6.2.2 Cell-type specific DNA methylation signatures of gestational age in cord blood.....	56
6.2.3 nRBCs in cord blood	57
6.2.4 Erythropoiesis.....	57
6.2.5 The switch from fetal to adult hemoglobin.....	60
6.2.6 Glucocorticoids and gestational age	61
6.2.7 The role of nRBCs in gestational age and fetal development.....	62
6.2.8 The role of leukocytes in gestational age	62
6.2.9 ART-children and epigenetic gestational age.....	64
6.2.10 Utility of gestational age clocks	65
6.2.11 Stability of CpGs in gestational age clocks	66
6.2.12 Biological relevance of CpGs that are stably predictive of gestational age	67
6.2.13 Linearity of the relationship between DNA methylation and gestational age	67
6.3 What makes the gestational age clock tick?	68
6.4 Strengths and limitations	70
6.4.1 Sample size and statistical power	70
6.4.2 Array-based DNA methylation data	70
6.4.3 Tissue specificity	71
6.4.4 Using reference data for inferring cell type proportions.....	72
6.4.5 Reliability of cell-type specific DNA methylation analysis methods.....	73
6.4.6 Phenotypic information	74
6.4.7 Range of gestational age.....	74
7. Conclusions and future perspectives.....	75
Appendix	77
References	78
Papers 1-3	

Acknowledgements

The work in this thesis was partly carried out at the Centre for Fertility and Health (CeFH), and partly at the Department of Genetics and Bioinformatics, both at the Norwegian Institute of Public Health. The work was funded by the Research Council of Norway's Centre of Excellence funding scheme, project number 262700, and by the US National Institutes of Health (NIH). I am very thankful to the participants in the Norwegian Mother, Father, and Child Cohort Study (MoBa) and the Finnish Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction study (PREDO). The research presented in this thesis would not be possible to conduct without their contributions.

I would also like to express my sincere gratitude to all my supervisors; Astanand Jugessur (Anil), Jon Bohlin, Siri E. Håberg, Per M. Magnus and Øyvind Næss for their support both scientifically and personally. In particular, I wish to thank my main supervisor, Anil, for letting me develop my own ideas and research skills while at the same time always being available for giving advice and encouragement whenever I needed it. I also want to thank Jon for sharing his knowledge and enthusiasm about the field of epigenetic gestational age, and for challenging me also when I didn't think I needed it.

I feel very privileged to be part of the vibrant and supportive research environment at CeFH and am grateful to all my coworkers and fellow PhD students for contributing to this stimulating and friendly work environment. I would also like to acknowledge my co-authors for their valuable contributions to our work. I want to direct a special thanks to William R. P. Denault for guiding me through the world of statistics, and for always noticing when I needed a reassuring phone call. Further, I wish to thank Yunsung Lee for his kindness and patience while introducing me to R and coding. Also, a big thanks to Christian Page for always having my back and helping me out whenever I needed it. To my fellow PhD student Ellen Ø. Carlsen: thank you for sharing the ups and downs of PhD life with me. I always appreciate our conversations and exchanges, whether they are about horses, kids, or science.

I am incredibly thankful to my amazing family who always supports me. I particularly want to thank Wencke and Jens for being the best parents-in-law I could ever dream of. I cannot express how grateful I am for everything you have done for us during this extremely busy PhD-period. Thanks to my kids, Elida and Lavrans, for all your cuddles and laughter, and for reminding me to be present. And finally, thanks to Trond, for your constant love and support, and for sharing my optimism in conducting projects that others may deem impossible, making the impossible possible yet again.

Kristine Løkås Haftorn,
Algarheim, April 2023

Sammendrag

Bakgrunn: Epigenetiske modifikasjoner, som DNA-metylering, er avgjørende for en rekke prosesser som celledifferensiering og fosterutvikling. Å kombinere epigenetiske analyser med populasjonsbasert epidemiologisk forskning gir betydelige muligheter for å utforske hvilken rolle epigenetisk variasjon spiller for menneskers helse, sykdom og utvikling.

Gestasjonsalder brukes ofte som en indikator på den nyfødtes utvikling og modenhet. Videre er prematur fødsel forbundet med en rekke skadelige utfall hos den nyfødte og senere i livet. Derfor er nøyaktig bestemmelse av gestasjonsalder avgjørende for å kunne sikre riktig perinatal omsorg. Imidlertid har estimering av gestasjonsalder basert på siste menstruasjon eller ultralydmålinger visse begrensninger. For nyfødte som er unnfanget ved hjelp av assistert befruktning (ART) er det nøyaktige tidspunktet når embryoet overføres til livmoren kjent, som dermed gir et mer direkte estimat av gestasjonsalderen. I tillegg sammenfaller tidspunktet for ART med omfattende epigenetiske endringer som skjer i det tidlige embryoet. ART har også vist seg å være forbundet med betydelige DNA-metyleringsendringer hos nyfødte.

DNA-metylering observert ved tusenvis av DNA metyleringsmarkører (CpG-er) over hele genomet har vist seg å variere med gestasjonsalder hos nyfødte, og kronologisk alder hos voksne. Dette har ført til utviklingen av flere nøyaktige DNA-metyleringsbaserte prediksjonsmodeller som kan beregne både alder og gestasjonsalder, kjent som «epigenetiske klokker». Imidlertid er forskningen på epigenetisk gestasjonsalder fortsatt på et tidlig stadium. Det er uklart hvorfor ulike CpG-er plukkes ut i forskjellige epigenetiske klokker. Man vet også lite om de prediktive CpG-ene brukt i klokkene så vel som de biologiske mekanismene som ligger til grunn for forbindelsen mellom DNA-metylering og gestasjonsalder.

Mål: Det overordnede formålet med avhandlingen var å utforske sammenhengen mellom DNA-metyleringsnivåer hos nyfødte og deres gestasjonsalder samt identifisere mekanismer som forklarer denne assosiasjonen. Spesifikt hadde vi som mål å (i) undersøke om det er en sammenheng mellom celletype, gestasjonsalder og DNA-metylering i navlestrengsblod, (ii) utvikle nye epigenetiske klokker for gestasjonsalder og utforske forskjeller mellom nyfødte som ble unnfanget med og uten hjelp av ART, og til slutt, (iii) undersøke, og eventuelt identifisere og karakterisere CpG-er som er stabilt prediktive for gestasjonsalder i den forstand at de konsekvent velges ut. For å nå disse målene studerte vi sammenhengen mellom DNA-metylering i navlestrengsblod og gestasjonsalder ved å bruke innsamlede data fra Den norske mor, far og barn-undersøkelsen (MoBa).

Metoder: Vi brukte BeadChip-arrays for å måle nivået av DNA-metylering for CpGene i genomet til de nyfødte. I tillegg undersøkte vi mulige fordeler ved å bruke det mer omfattende Infinium MethylationEPIC (EPIC) arrayet sammenlignet med et tidligere array, Infinium HumanMethylation

450K (450K). Først gjennomførte vi en epigenom-vid assosiasjonsstudie (EWAS) der vi lette etter sammenhenger mellom gestasjonsalder og DNA metylering. Vi brukte to forskjellige metoder, CellDMC og Tensor Composition Analysis (TCA), for å undersøke om det også fantes celletypespesifikke assosiasjoner. Deretter brukte vi lasso-regresjon for å utvikle en epigenetisk klokke for gestasjonsalder som er spesifikk for EPIC. Informasjon om embryooverføringsdato hos ART-unnfangede nyfødte ble brukt for å evaluere presisjonen til klokken vår, og vi undersøkte om det var forskjeller i epigenetisk gestasjonsalder mellom ART-unnfangede og naturlig unnfangede nyfødte ved å bruke logistisk regresjon. Videre brukte vi en statistisk tilnærming kalt «stability selection» som kombinerer delutvalg med variabelseleksjon for å identifisere CpG-er som er svært prediktive for gestasjonsalder. Vi brukte deretter en ulineær regresjonsmodell (GAM) for å utvikle nye epigenetiske klokker basert på de stabilt prediktive CpG-ene. Til slutt brukte vi ulike algoritmer for annotering og karakterisering av genene som de gestasjonsaldersassosierte CpG-ene var tilknyttet, og deres tilhørende gen-nettverk.

Resultater: Vi oppdaget signifikante sammenhenger mellom DNA-metylering og gestasjonsalder i alle de syv hovedtypene av celler i navlestrengsblod. Imidlertid var de fleste signifikante CpG-ene knyttet til kjerneholdige røde blodceller (nRBCer). Mange av disse CpG-ene er tilknyttet spesifikke prosesser involvert i utvikling av røde blodceller og overgangen fra føtalt til voksent hemoglobin. Vi utviklet også en EPIC-spesifikk epigenetisk klokke for gestasjonsalder som var mer nøyaktig enn tidligere publiserte gestasjonsalderklokker. Å begrense analysen til CpG-er som dekkes både av EPIC og 450K reduserte imidlertid ikke klokkenes presisjon. Bruk av embryooverføringsdato i stedet for ultralydmålinger for å utvikle klokken forbedret heller ikke klokkenes ytelse og klokken fungerte like godt på både ART- og naturlig unnfangede barn. Videre var det ingen signifikante forskjeller i epigenetisk gestasjonsalder mellom de to gruppene. Totalt sett fant vi 24 CpG-er som var stabilt prediktive for gestasjonsalder, og bare opp til 10% av CpG-er i tidligere publiserte epigenetiske gestasjonsalderklokker ble identifisert som stabilt prediktive i vår studie. Flere av de stabilt prediktive CpG-ene var i eller nær gener involvert i immunrespons, metabolisme og diverse utviklingsprosesser. Til slutt brukte vi de stabilt prediktive CpG-ene til å utvikle en ny gestasjonsalderklokke bestående av bare fem CpG-er. Til tross for få CpG-er hadde klokken en tilsvarende presisjon som etablerte gestasjonsalderklokker bestående av dusinvis til hundrevis av CpG-er. GAM metoden avslørte også ulineære sammenhenger mellom DNA metylering og gestasjonsalder hos premature nyfødte.

Konklusjoner: Samlet sett fører våre funn til økt forståelse av sammenhengen mellom DNA-metylering og gestasjonsalder. Resultatene peker på celletypeutvikling, utviklingsprosesser og forberedelse til fødsel og livet utenfor livmoren som plausible mekanismer som ligger til grunn for denne assosiasjonen. Videre har vi utviklet flere nøyaktige epigenetiske klokker som er nyttige verktøy for videre studier av epigenetisk gestasjonsalder.

Abstract

Background: Epigenetic modifications, such as DNA methylation, are essential for a wide array of developmental processes, including cellular differentiation and human development. Combining epigenetic analyses with population-based epidemiological research offers tremendous opportunities for exploring the role of epigenetic variation in human health, disease, and development.

In clinical and research settings, gestational age is often used as an indicator of the newborn's developmental maturity. Furthermore, preterm birth is associated with a range of deleterious outcomes in the neonate and later in life. Thus, accurate determination of gestational age is essential to ensure proper perinatal care. However, estimating gestational age from the last menstrual period or ultrasound measurements has certain limitations. For newborns conceived using assisted reproductive technology (ART), the exact time when the embryo is transferred to the uterus is known and thus provides a more direct estimate of gestational age. Moreover, ART coincides with the extensive epigenetic remodeling that takes place in the early embryo and has also been shown to be associated with DNA methylation alterations in the newborns.

DNA methylation at thousands of CpG sites throughout the epigenome has been shown to be strongly associated with gestational age in newborns and chronological age in adults. This has prompted the development of several accurate DNA methylation-based predictors of age and gestational age, commonly known as 'epigenetic clocks'. Nevertheless, the field of epigenetic gestational aging is still in its nascent stages. The reason for the considerable lack of overlap in predictive CpGs across different epigenetic gestational age clocks remains elusive. Similarly, very little is known about the implications of the predictive CpGs as well as the biological mechanisms underlying the association between DNA methylation and gestational age.

Aims: The overarching aim of this thesis was to explore genome-wide DNA methylation levels in newborns in relation to their gestational age and identify mechanisms that explain this association. Specifically, we aimed to (i) investigate the cell-type specific association between gestational age and DNA methylation in cord blood, (ii) develop new epigenetic clocks for gestational age and explore differences in epigenetic gestational age between ART-conceived newborns and those conceived naturally, and, finally, (iii) identify and characterize CpGs that are stably predictive of gestational age. To achieve these aims, we studied the association between cord blood DNA methylation and gestational age in several subsamples of the Norwegian Mother, Father, and Child Cohort Study (MoBa).

Methods: We used BeadChip arrays to quantify the level of DNA methylation at each CpG site and examined potential advantages of using the more comprehensive Illumina MethylationEPIC (EPIC) array compared to previous arrays. First, we conducted an epigenome-wide association study (EWAS)

of gestational age and applied two different methods, CellDMC and Tensor Composition Analysis (TCA), to elucidate cell-type specific epigenome-wide associations. Second, we used lasso regression to develop a highly performant epigenetic gestational age clock specific for the EPIC array, used information on embryo transfer date in ART-conceived newborns to evaluate the performance of our clock, and used logistic regression to explore differences in epigenetic gestational age between ART-conceived and naturally-conceived newborns. Third, we used a statistical approach called ‘stability selection’ that combines subsampling with variable selection to identify CpGs that are stably predictive of gestational age. We then applied generalized additive model (GAM) regression to develop new and more parsimonious gestational age clocks based on the stably selected CpGs. Finally, we used gene annotation and gene-set enrichment algorithms to examine the genomic location and pathway annotations of the gestational-age associated CpGs.

Results: We discovered significant associations between DNA methylation and gestational age in all the seven main cell types in cord blood. However, most of the significant CpGs were restricted to nucleated red blood cells (nRBCs) and were strongly linked to specific processes involved in red blood cell development (erythropoiesis) and the switch from fetal to adult hemoglobin. We also developed a highly performant epigenetic gestational age clock specific for the EPIC array, which outperformed previously published gestational age clocks. However, restricting the analysis to CpGs shared between EPIC and a previous array (450K) did not reduce the precision of the clock. Using embryo transfer date instead of ultrasound measurements to develop the clock did not improve the prediction performance; our clock performed equally well in ART-conceived and naturally-conceived children. Furthermore, there were no significant differences in epigenetic gestational age or gestational age acceleration between the two groups. Overall, we identified 24 CpGs as being stably predictive of gestational age, and only up to 10% of CpGs in previously published epigenetic gestational age clocks were stably selected in our study. Several of the stably selected CpGs were in or near genes implicated in immune responses, metabolism, and developmental processes. Finally, we used the stably selected CpGs to develop a new gestational age clock consisting of only five CpGs. Strikingly, this clock showed a similar predictive performance to that of established gestational age clocks consisting of dozens to hundreds of CpGs. Furthermore, accounting for nonlinear associations between CpGs and gestational age improved gestational age prediction in preterm newborns.

Conclusions: Overall, our findings contribute to an increased understanding of the association between DNA methylation and gestational age and propose signatures of cell-type development, developmental processes, and preparation for birth and postnatal life as some of the plausible mechanisms underlying this association. Furthermore, we have developed several accurate epigenetic gestational age clocks that will be useful tools for further studies on epigenetic gestational age and developmental maturity.

List of papers

This thesis is based on the following papers:

Paper 1

Haftorn, K.L., Denault, W.R.P., Lee, Y., Page, C.M., Romanowska, J., Lyle, R., Næss Ø.E., Kristjansson, D., Magnus, P.M., Håberg, S.E., Bohlin, J.*, Jugessur, A.* **Nucleated red blood cells explain most of the association between DNA methylation and gestational age.** *Commun Biol* 6, 224 (2023). *Joint senior authors. <https://doi.org/10.1038/s42003-023-04584-w>

Paper 2

Haftorn, K.L., Lee, Y., Denault, W.R.P., Page, C.M., Nustad, H.E., Lyle, R., Gjessing, H.K., Malmberg, A., Magnus, M.C., Næss, Ø., Czamara, D., Räikkönen, K., Lahti, J., Magnus, P., Håberg, S.E., Jugessur, A.*, Bohlin, J.* **An EPIC predictor of gestational age and its application to newborns conceived by assisted reproductive technologies.** *Clin Epigenet* 13, 82 (2021). *Joint senior authors. <https://doi.org/10.1186/s13148-021-01055-z>

Paper 3

Haftorn, K.L., Romanowska, J., Lee, Y., Page, C.M., Magnus, P.M., Håberg, S.E., Bohlin, J., Jugessur, A.*, Denault, W.R.P.* **Stability selection enhances feature selection and enables accurate prediction of gestational age using only five DNA methylation sites.** *Joint senior authors. Submitted to *Genome Biol.* (2023).

Please also see relevant papers not included in the thesis in the Appendix.

Abbreviations

27K: Infinium HumanMethylation27 BeadChip

450K: Infinium HumanMethylation 450K

5hmC: 5-hydroxymethylcytosine

A: adenine

ALSPAC: Avon Longitudinal Study of Parents and Children

ART: assisted reproductive technology

BFU-E: burst-forming unit-erythroid

C: cytosine

C-section: cesarean section

CeFH: Centre for Fertility and Health

CFU-E: colony-forming unit-erythroid

CHARM: comprehensive high-throughput arrays for relative methylation

CpG: 5'-cytosine-phosphate-guanine-3'

DEPICT: Data-driven Expression Prioritized Integration for Complex Traits

DAG: directed acyclic graph

DAVID: Database for Annotation, Visualization and Integrated Discovery

ddNTP: dideoxynucleotide triphosphate

DMR: differentially methylated regions

DNA: deoxyribonucleic acid

DNMT: DNA methylation transferase

DunedinPoAm: Dunedin Pace-of-Aging methylation

EAA: epigenetic age acceleration

EPIC: Infinium MethylationEPIC BeadChip

ETD: embryo transfer date

EWAS: epigenome-wide association study

FACS: fluorescence-activated cell sorting

FDR: false discovery rate

G: guanine

GA: gestational age

GAA: gestational age acceleration

GAM: generalized additive model

GREAT: Genomic Regions Enrichment of Annotation Tool

ICR: imprinting control region

ICSI: intracytoplasmic sperm injection

IUI: intrauterine insemination

IVF: *in vitro* fertilization

LMP: last menstrual period

lncRNA: long noncoding RNA

MAD: median absolute deviation

MBRN: Medical Birth Registry of Norway

MoBa: Norwegian Mother, Father, and Child Cohort Study

NIH: National Institutes of Health

NK cell: natural killer cell

nRBC: nucleated red blood cell

OLS: ordinary least squares

PREDO: Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction study

QC: quality control

qPCR: quantitative real-time polymerase chain reaction

RNA: ribonucleic acid

RRBS: reduced representation bisulfite sequencing

SE: standard error

SNP: single-nucleotide polymorphism

START: study of assisted reproductive technology

T: thymine

TCA: Tensor Composition Analysis

TET: ten-eleven translocation

WebGestalt: Web-based Gene set analysis toolkit

1. Introduction

During approximately nine months of gestation, a single fertilized cell develops into a complex and multicellular human being. Through cell division, differentiation, organ formation, development, and growth, the fetus prepares for life outside the womb until birth marks the end of gestation. The gestational age of the newborn signifies how long he/she was confined within the protective environment of the uterus and can thus provide valuable information regarding the newborn's developmental maturity.

Although environmental factors *in utero* can influence early human development, it is generally under strict genetic control. The 'molecule of life' – DNA – provides the blueprint to create everything that is needed to form a fully functional human body. However, DNA must be decoded and interpreted correctly to convey the genetic information with high fidelity. Consequently, mechanisms that control accessibility and readability of DNA, and those that regulate the timing and rate of gene expression, are essential for normal fetal development. Epigenetics represents one of the many layers of regulatory mechanisms that ensures that the right genes are converted into the right amount of proteins at the right time and in the right cell (1).

In this thesis, I investigate the association between the most widely studied epigenetic modification, namely DNA methylation, and gestational age. I aim to identify mechanisms that underlie this association in order to shed more light on epigenetic processes that characterize fetal growth and development.

2. Background

2.1 Epigenetics

2.1.1 Epigenetic modifications

The DNA molecule is a double-stranded helix held together by weak hydrogen bonds between the following nucleotide base pairs: adenine (A) paired with thymine (T), and guanine (G) paired with cytosine (C). In the eukaryotic cell, proper packaging of DNA into chromatin is essential to fit the entire DNA molecule into the narrow confines of the nucleus. This dense packaging protects DNA from damage and controls DNA accessibility and gene expression (**Figure 1**, see also (2)). The main building block of chromatin is the nucleosome, which consist of DNA wrapped around an octamer of proteins called histones. Modifications to chromatin that do not alter the underlying sequence or backbone of DNA, such as covalent modifications of DNA bases, posttranslational modifications of amino acids on the N-terminal tail of histones, histone variants, and nucleosomal remodeling machines, are able to regulate gene expression by modifying the underlying chromatin structure and access to DNA (2, 3). These epigenetic processes interact with each other to ensure stable states of gene expression.

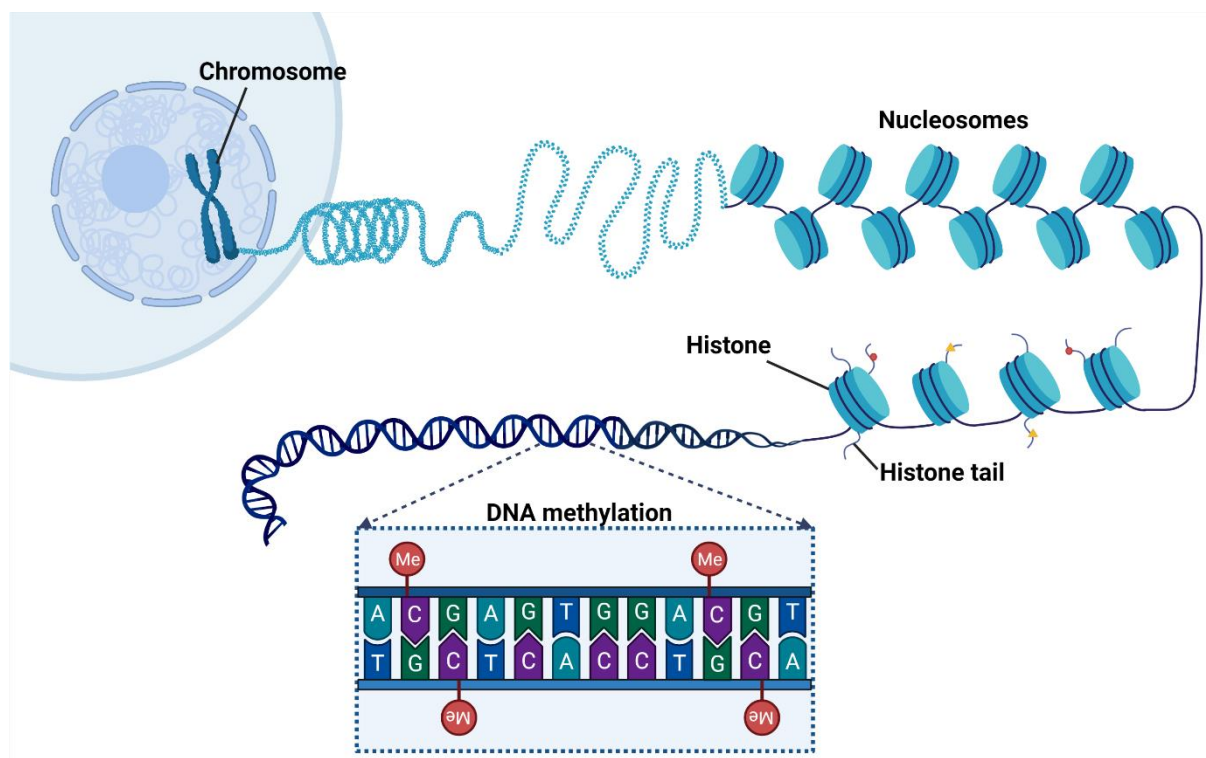


Figure 1. DNA packaging and epigenetic modifications. The DNA double-helix is densely packaged into chromatin in the cell nucleus. DNA wrapped around an octamer of histones constitutes a nucleosome. Amino acids on N-terminal tails protruding from the histones are prone to posttranslational modifications. Underneath the double-helix is a simplified representation of paired nucleotides with methylated CpGs (methyl group shown as red circles). Created with [BioRender.com](https://www.biorender.com).

Epigenetic modifications are mitotically heritable, meaning that they can be transmitted to daughter cells after DNA synthesis and mitosis (4). The epigenome is susceptible to alterations by environmental factors, and to dysregulation during aging and the development of diseases such as cancer (5, 6, 7). Among the many important types of epigenetic modifications, this thesis focuses exclusively on DNA methylation. DNA methylation is by far the most researched type of epigenetic modification. It is also the most stable and accessible epigenetic modification, which makes it ideally suited for epigenome-wide studies (8).

2.1.2 DNA methylation

DNA methylation entails the transfer of a methyl group to a base in the DNA sequence, predominantly at the fifth carbon of a cytosine (C) residue that is attached to a guanine (G) via a phosphate ('p') group. This specific base sequence is hence usually referred to as a CpG site. There are an estimated 29 million CpG sites in the human genome, and except for some tissue-specific differences, 70-80% of these CpGs are estimated to be methylated (9). Groups of unmethylated CpGs are often found near gene promoters, in dense clusters called CpG islands, that play central roles in gene regulation (10). Regions surrounding the CpG islands are called shores (0-2 kb from island edge) (11) and shelves (>2-4 kb from island edge) (12). The remaining regions belong to the 'open sea' (13).

Although an ancient property of eukaryotic genomes (14), DNA cytosine methylation has been lost in several eukaryotic lineages, including common model organisms like *Drosophila melanogaster* (common fruit fly), *Caenorhabditis elegans* (a species of nematode worm), and several yeasts (15, 16). Organisms that exhibit CpG methylation have reduced CpG content because methylated cytosines can deaminate to form uracil, leading to C → T transitions (17).

The process of DNA methylation can be divided into three phases: establishment (*de novo* DNA methylation), maintenance, and demethylation (**Figure 2**). In mammals, there are two major *de novo* DNA methylation enzymes: DNA methylation transferase (DNMT) 3A and DNMT3B (18). DNMT1 maintains symmetrical CpG methylation upon DNA replication. In the absence of a functional DNA methylation maintenance machinery, successive rounds of replication will lead to methylation loss, defined as passive DNA demethylation (19). Active demethylation, on the other hand, is carried out by the Ten-eleven translocation (TET) methylcytosine dioxygenases, which progressively oxidize methylated cytosines to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine, and 5-carboxylcytosine (20, 21), resulting in DNA demethylation during replication or base removal by the base-excision repair pathway (22, 23, 24, 25, 26).

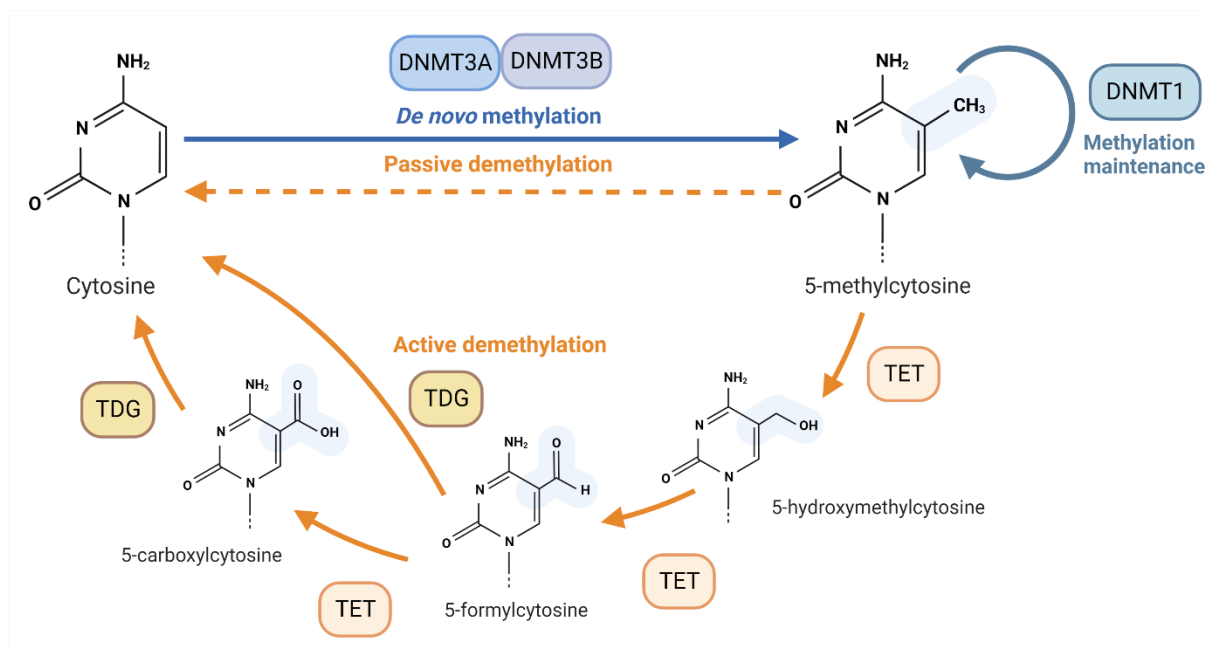


Figure 2. DNA cytosine methylation and demethylation. *De novo* methylation enzymes DNMT3A and DNMT3B are responsible for the establishment of cytosine methylation patterns, whereas DNMT1 maintains symmetrical CpG methylation upon DNA replication. DNA demethylation can either be a passive process due to the absence of DNA methylation maintenance over several rounds of replication, or an active process conducted in several steps by Ten-eleven translocation (TET) methylcytosine dioxygenases and thymine DNA glycosylase (TDG). Methylation processes and enzymes are highlighted in blue, whereas demethylation processes and enzymes are highlighted in orange. Created with [BioRender.com](https://www.biorender.com).

The maintenance of epigenetic status during cell division can lead to coordination of DNA methylation levels at adjacent CpGs (27, 28, 29). Such co-methylation patterns can also exist between distal CpGs because they may be brought into spatial proximity through chromatin folding (30).

Depending on the location of the CpG, DNA methylation is involved in transcriptional regulation and alternative splicing (31) and plays a major role in repressing transposons (32). Although highly context specific, promoter methylation has generally been associated with repression of transcription (33). DNA methylation is also highly enriched in pericentromeric satellite repeats and in the bodies of actively transcribed genes, in contrast to CpG islands in promoters of actively transcribed genes, which are usually unmethylated (14). DNA methylation is essential for maintaining long-term repression of genes, including germline-specific genes (34), imprinted genes (i.e., genes of which the maternally derived or paternally derived allele is suppressed in the embryo) (35, 36), and genes located on the inactive X-chromosome (37).

2.1.3 Tissue-specificity of DNA methylation

DNA methylation plays a critical role in cellular differentiation, particularly in establishing and maintaining cellular identity (38). For example, through silencing pluripotent factors, DNA methylation is directly involved in the initiation of cellular differentiation of pluripotent cells. Although there is a global increase in DNA methylation during cellular differentiation, specific loci show a cell-type dependent *decrease* in DNA methylation (39, 40). These demethylated lineage-specific loci are not restricted to promoter regions but are also found in distal gene sequences and intronic regions. Moreover, CpG island shores are susceptible to methylation in a tissue-specific manner that correlates with gene expression (11), and the DNA methylation status of enhancer regions contributes to forming the epigenetic memory of specific cell-types (41). Although DNA methylation is regarded as a relatively stable epigenetic mark, it has been identified as the most discriminant epigenetic feature across different tissues (42).

2.1.4 Quantification of DNA methylation

DNA methylation level can be quantified in different ways. For example, bisulfite-converted DNA or immunoprecipitation of methylated fragments can be used either together with quantitative real-time PCR (qPCR), to analyze specific loci of interest, or with sequencing methods for whole-genome analysis (43). These methods are relatively expensive, time-consuming, and laborious, especially when many samples need to be analyzed. An alternative method for quantifying DNA methylation at the genome-wide level is to use BeadChip arrays targeting specific CpGs distributed across the genome. ‘Illumina’, a biotechnological company headquartered in San Diego, CA, USA, has developed several of these arrays during the last decade. The first platform, GoldenGate, was launched in 2006 and included 1,536 CpG sites related to cancer (44). Three years later, the Infinium HumanMethylation27 BeadChip (27K) harboring ~27,000 CpGs was released (45). Next came the Infinium Human Methylation 450K BeadChip (450K) in 2011, covering approximately 17-fold more CpGs (~450,000) than the 27K array and targeting 99% of all RefSeq genes with an average of 17 probes per gene ((12); <https://www.ncbi.nlm.nih.gov/refseq/>). Finally, the Infinium MethylationEPIC BeadChip (EPIC) was released in December 2015 and covered ~850,000 CpGs, including several in regulatory regions (46). At the time of writing, Illumina has launched an updated version of its EPIC array, Infinium MethylationEPIC v2.0, in which poorly performing probes were removed and an additional 186,000 CpGs were included (47). Recently, other platforms have been launched for the mouse methylome (48), as well as custom-arrays housing anywhere between 3,000-100,000 markers.

2.1.5 Epigenetic epidemiology

There are many different approaches to studying epigenetics, and, more specifically, DNA methylation. Combined knowledge from various disciplines, such as biochemistry, molecular biology,

and physiology, is essential for gaining a deeper understanding of the enzymatic reactions required to establish and maintain methylation levels. Such knowledge is also key to elucidating the biological functions of DNA methylation and for mapping out complex interactions between various epigenetic modifications and other key cellular processes.

Rather than focusing on processes that operate consistently in every person, the field of epidemiology seeks to understand the reasons for variability in different traits within a population (49). The emergence of high-throughput technologies for more comprehensive epigenetic analyses coupled with an increasing recognition of the role of epigenetic variation in human health and disease have introduced the field of epigenetic epidemiology (50). This usually entails studying large cohorts of individuals to determine the extent to which epigenetic marks vary among individuals and throughout the life course due to a combination of genetics, environmental exposures, and life experiences. Specifically, epigenome-wide association studies (EWASes) have become an integral part of exploring the links between DNA methylation and a range of exposures and phenotypes. The timely development of DNA methylation BeadChip arrays has been especially relevant, as they provide an affordable option for quantifying DNA methylation in large studies while still retaining an adequate precision and coverage of DNA methylation sites. The integration of epigenetic analyses into population-based epidemiological research has thus created a suitable framework for exploring the role of epigenetic variation in human development, including descriptive studies as well as studies that specifically target the causes and consequences of epigenetic variation (50).

2.2 Development

2.2.1 Prenatal development

The development of a single-celled zygote to a multicellular adult organism is complex and requires a wide variety of mechanisms and processes. I will give a brief overview of some of these processes in this chapter, which is primarily based on the seminal book “Larsen’s Human Embryology” (51).

Human prenatal development is usually divided into three trimesters. The first starts with the fertilization of an ovum by a sperm in one of the ovarian ducts. The resulting zygote undergoes a series of cell divisions by mitosis which give rise to a cluster of multiple cells. These cells then go on to form the blastocyst, which consists of a large fluid-filled central cavity. During the first week of embryonic development, the embryo travels from the ovarian duct to the uterine cavity and initiates implantation into the uterine wall, becoming fully implanted during the first 6-9 days after conception. Implantation is followed by the formation of the ‘yolk sac’, a structure associated with the developing embryo through the fourth week of development. The yolk sac has several important functions, including the formation of blood cells (hematopoiesis) and holding primordial germ cells that give rise to male and female gametes. The embryo continues to grow rapidly, with most of the major organ

systems formed in the following few weeks. The embryonic period ends after the eighth week of development. The fetal period, which is the remaining period of gestation, is devoted mainly to maturation and growth. The weight of the fetus increases by approximately 250-fold during the second and third trimesters. Whereas most of the growth in length occurs during the second trimester, most of the weight is added during the third trimester.

The placenta is an organ consisting of both maternal and fetal components. It provides nutrients to the fetus and eliminates metabolic waste. The implanting blastocyst induces the development of the placenta. The uteroplacental circulation system begins to develop during the second week of development. This system allows the exchange of gas and metabolites between maternal and fetal blood by diffusion. The placenta is connected to the fetus via the umbilical cord, and umbilical arteries and veins develop within the umbilical cord to allow blood to circulate between the fetus and the placenta. In addition to nutrient and gas exchange, the placenta secretes hormones such as sex steroids that help maintain pregnancy. Moreover, maternal antibodies can cross the placenta into the fetus, where they provide protection against infections.

2.2.2 Postnatal development

Unlike prenatal development, which takes place in the relatively predictable and stable environment of the uterus, postnatal development can be affected by a range of highly variable environmental factors. In general, postnatal development can be divided into four main phases: infancy, childhood, puberty, and adulthood.

Infant and childhood development is, to a large extent, a continuation of *in utero* growth and maturation. These phases are also characterized by extensive developmental changes in the neurobehavioral, gastrointestinal, and immune systems due to the substantial differences in sensory input, microbiota and other environmental factors in postnatal life compared to life *in utero* (52, 53, 54, 55). Puberty marks the transition from childhood to adulthood and is characterized by rapid growth, development of secondary sexual characteristics, gonadal maturation, and attainment of reproductive capacity, as well as changes in brain function and cognitive development (56, 57). Pubertal development is mainly controlled by hormonal activity through the hypothalamic-pituitary-gonadal axis (56).

An adult can be defined as a person who is fully grown and mature and thus no longer developing. However, there is a lack of agreement on the age at which individuals should be considered adults (58). Moreover, processes such as brain and cognitive development continue after early adulthood (58, 59). It has been proposed that aging represents a continuation of developmental processes, implying that development continues throughout the human lifespan (60, 61, 62).

Although I have provided a backdrop for the developmental events spanning the prenatal period to adult life, this thesis focuses primarily on the pre- and perinatal period (i.e., the period before and around the time of birth) because gestational age was the main phenotype of interest.

2.2.3 The role of DNA methylation in development

DNA methylation is integral to mammalian development. As mentioned in chapter 2.1.2, DNA methylation is, among its many other functions, implicated in genomic imprinting, X-chromosome inactivation, and the repression of germline-specific genes, all of which are key processes during embryonic development (63). Notably, DNMT-deficient mice exhibit severe developmental abnormalities leading to early embryonic lethality, further emphasizing the importance of a functional DNA methylation machinery during development (18, 64).

The regulation of DNA methylation erasure and (re-)establishment varies considerably between different developmental stages (63, 65). The mammalian genome undergoes two extensive waves of DNA methylation reprogramming during embryogenesis. The first happens shortly after fertilization and the other after germline cell specification (66, 67). During post-fertilization reprogramming, the embryo loses gamete-specific DNA methylation patterns inherited from the oocyte and the sperm. Although embryonic and germline demethylation is genome-wide, a substantial amount of DNA methylation persists at the end of both processes. In both humans and mice, approximately 20% of CpGs in the pluripotent cells of preimplantation embryos retain gamete-inherited methylation (68, 69). These CpGs are found in imprinting control regions (ICRs), in transposable elements (70) and in transiently imprinted regions with mainly maternal-specific DNA methylation patterns that are maintained until the blastocyst stage but are lost after implantation (63, 71). These transient imprints may have specific roles in development, as exemplified by the *Zdbf2* locus in mice, where transient hypomethylation of the paternal allele can cause long-lasting imprinting through a cascade of downstream epigenetic changes that affect postnatal growth and feeding behavior (72, 73). Remethylation after implantation is very rapid, and somatic levels of DNA methylation are retained when the epiblast stem cells are still pluripotent (74). This means that DNA methylation patterns established at this very early stage have the potential to be propagated through life in all tissues, thus maintaining epigenetic memory of early embryogenesis (74). However, as described in chapter 2.1.3, tissue differentiation also induces changes in DNA methylation patterns (75, 76).

Whereas the reprogramming of the human methylome during embryonic development has been extensively studied, less is known about the DNA methylation dynamics in the fetal period of development. A study profiling 12 mouse tissues and organs at nine developmental stages from embryogenesis to adulthood observed continuous loss of CpG methylation throughout fetal development and CpG remethylation postnatally, primarily at distal regulatory elements (77). The

authors also discovered the accumulation of non-CpG methylation within the bodies of key developmental transcription factor genes during late stages of fetal development, coinciding with transcriptional repression of these genes (77).

Another study examining DNA methylation patterns in four human fetal tissues during the first and second trimesters observed large-scale remodelling of DNA methylation from gestational week 9 to 22, with specific hypomethylation near tissue-specific genes and hypermethylation near developmental genes (78). Dynamic DNA methylation was associated with the progressive repression of developmental programs and the activation of genes involved in tissue-specific processes. These studies indicate that DNA methylation is not only an integral part of early embryo development but continues to be a key feature of epigenetic remodeling throughout fetal development.

DNA methylation may also be implicated in placental growth and differentiation, although its role in these processes is poorly understood (79). The placental genome is generally hypomethylated compared to that of other healthy tissues (79, 80), but several studies have showed a progressive increase in placental DNA methylation with gestational age (81, 82, 83). Associations between placental DNA methylation and fetal growth has also been observed (84, 85, 86). Finally, imprinting also appears to be particularly important for placental development (87, 88).

2.3 Assisted reproductive technology (ART) and DNA methylation

2.3.1 Definitions and epidemiology

ART is a collective term used to define all interventions that include the *in vitro* handling of both human oocytes and sperm or embryos for the purpose of reproduction (89). By this definition, intrauterine insemination (IUI) is not considered ART, but *in vitro* fertilization (IVF) is. In brief, IVF entails the harvesting of oocytes from a woman's ovary to be fertilized by sperm outside of the female body (**Figure 3**). The resulting embryos are cultured in the laboratory for some days before they are either transferred to the uterus ('fresh' transfer) or cryopreserved. Cryopreserved embryos may later be thawed and transferred to the uterus ('frozen' transfer). IVF may be combined with intracytoplasmic sperm injection (ICSI), wherein a spermatozoon is injected directly into an oocyte's cytoplasm.

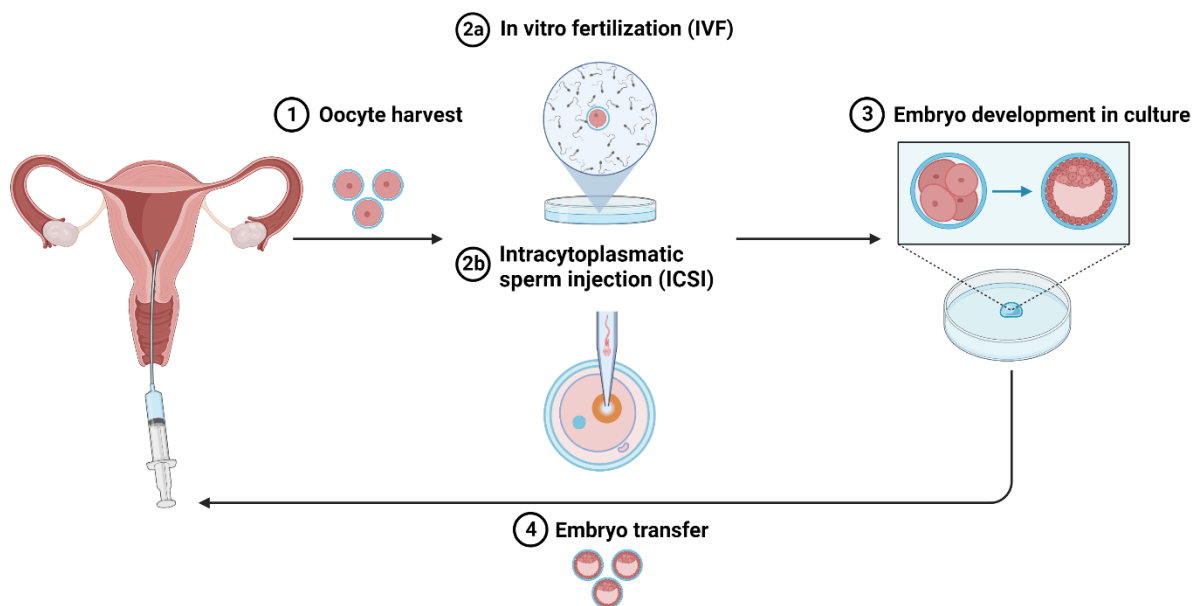


Figure 3. The process of *in vitro* fertilization. (1) Oocytes are harvested from a woman's ovary. (2) the oocyte is fertilized either (a) by being mixed with many spermatozoa in a laboratory dish, or (b) via direct injection of a spermatozoon into its cytoplasm. (3) The resulting embryo is cultured in the laboratory for some days. (4) Embryo(s) are transferred to the uterus. Created with [BioRender.com](https://www.biorender.com).

The use of ART has been on the rise since the first live birth of an IVF-baby in England in 1978 (90). More than eight million children have since then been born with the aid of ART worldwide (91). In Norway, the percentage of children born after ART each year was 5.4% in 2020 (data derived from the Medical Birth Registry of Norway), and the prevalence of ART use is higher amongst parents of older age and with higher socioeconomic status (89).

2.3.2 The role of DNA methylation in ART

ART involves several manipulations of the gametes and early embryo at the same time as the embryo undergoes the extensive epigenetic reprogramming described in chapter 2.2.3. These manipulations include hormonal stimulation of the ovaries, surgical retrieval of oocytes, IVF with or without ICSI, culturing, storing, and transferring embryos. During these procedures, the cells involved are exposed to nonphysiological changes in temperature, light, oxygen levels, pH, and different culture media. Therefore, it is highly likely that ART affects the establishment and/or maintenance of epigenetic marks (92). Both animal and human studies have found epigenetic changes in placental tissues and cord blood from ART pregnancies, although the effects of these changes are not well established (92, 93, 94). ART has also been linked to a few rare imprinting disorders (95, 96).

2.3.3 ART and perinatal outcomes

Conceiving with the help of ART is associated with several adverse perinatal outcomes, such as fetal growth restriction, preeclampsia, and birth defects (97). There is a higher prevalence of multiple births amongst mothers who conceived using ART, but this is largely due to the common use of multiple embryo transfer (97). The twinning rate among ART conceptions has declined over time due to a gradual shift to single-embryo transfer, leading to considerably reduced perinatal risks for ART children in recent years (98). Nevertheless, singletons born after ART have a 2- to 3-fold increased risk of several adverse perinatal outcomes (97). The risks seem to differ between different ART procedures. For instance, children born after fresh embryo transfer have a higher risk of low birthweight, whereas being born after frozen embryo transfer is associated with higher birthweight and maternal preeclampsia (97). In general, pregnancies due to ART use are associated with a shorter gestational age at birth and a higher rate of preterm birth (99, 100).

2.4 Gestational age

2.4.1 Definition and epidemiology

Gestation is the period between conception and birth. Gestational age is defined as the duration of pregnancy, measured in weeks or days, from the first day of a woman's last menstrual period (LMP) to the current date. It is important to note that by using this definition the estimated gestational age is approximately two weeks longer than the actual time that has elapsed since conception. This is because of the interval between the onset of the last menstruation to the actual ovulation and conception. The average gestational age at birth in singleton pregnancies is approximately 282 days (40 weeks) but ranges between 259 to 293 days for what is considered 'term birth' (101, 102). Conversely, birth before 259 days (37 weeks) is considered preterm, while birth after 294 days or more is considered post term. However, as fetal development is continuous across these cutoffs, different subcategories of preterm and term birth definitions have been suggested to describe deliveries and their outcomes more accurately and to better understand the impact of gestational age on perinatal outcomes (**Figure 4**; see also (101, 103))

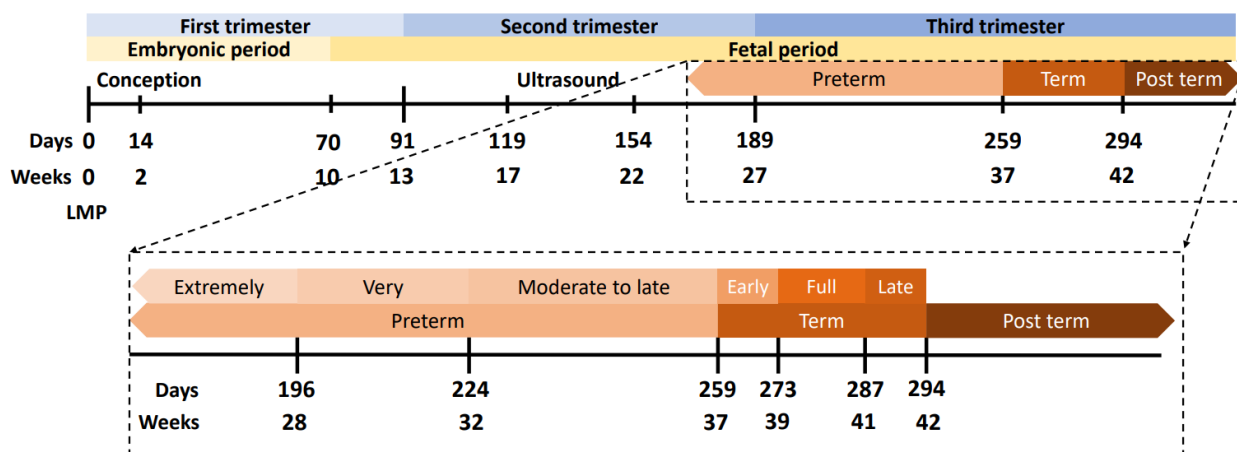


Figure 4. The timeline of gestation. The start of gestation is marked by the first day of the last menstrual period (LMP) before pregnancy, approximately 14 days before conception. The three trimesters are highlighted in different shades of blue, whereas the embryonic and fetal periods are separated by different shades of yellow. Ultrasound measurements in MoBa were performed between gestational weeks 17 and 22. The excerpt shows the different subcategories of preterm, term and post term birth definitions, highlighted in different shades of salmon, orange and brown, respectively.

As mentioned above, there is considerable variation in the duration of term pregnancies, even when excluding pathological pregnancies and accounting for measurement error in the method used to estimate gestational age (49). Such variation may reflect differences in the pace of fetal maturation, or differences in the mother's capacity to carry the fetus to term. Importantly, clinical interventions on gestational age, such as cesarean section (c-section) and induction of labor before it occurs naturally, influence the gestational age distribution (49). Reasons for applying such interventions include obstetric indications such as preeclampsia, fetal distress, placental abruption, and prolonged or difficult labor (104). However, some are due to elective c-section in the absence of medical indications.

2.4.2 Gestational age in development and disease

In clinical and research settings, gestational age is often used as an indication of developmental maturity (105). Preterm birth has been reported to be associated with a range of deleterious outcomes in the neonate and later in life (106, 107, 108). Several studies have showed that early term and post term newborns have worse health outcomes compared to full-term babies (109, 110). Births at 40-41 weeks have been suggested as the ideal window for optimal neurodevelopmental outcomes (111).

2.4.3 Clinical determination of gestational age

The ideal measure of gestational age would be from the day of conception to the date of delivery. However, as there are currently no methods that can measure or detect conception with such certainty, gestational age can only be estimated approximately. Gestational age can be estimated from LMP, but there are many possible sources of error in this estimate. For instance, the length and regularity of menstrual cycles can vary both between and within individuals (49), the use of contraceptives may influence the cycle, and recall bias may influence the estimates (112, 113). Ultrasonographic biometry is another widely used method for gestational age determination during pregnancy and is generally considered to be more accurate than LMP (113). However, the main limitation of ultrasound dating is the assumption that, at the time of ultrasound measurement, all fetuses should have the same dimensions. Hence, the certainty of ultrasound estimates may be influenced by differences in early growth of the fetus and the timing of ultrasound measurements (114). Measurement error is also a limitation; for example, maternal obesity may affect the ultrasound measurements (115). Moreover, there are marked discrepancies between gestational age at birth estimated by LMP and ultrasound data, especially in the earlier and later gestational ages (116).

For ART pregnancies, there is a third option to estimate gestational age before delivery in addition to the LMP date and ultrasound measurements. Because the timing of fertilization and embryo transfer is known, this information can in theory be used to calculate the true gestational age of the fetus. However, using the date of fertilization or embryo transfer for gestational age estimation is not entirely unproblematic. The time span from ovulation to fertilization and implantation in pregnancies conceived in natural cycles might differ from that in ART pregnancies. Furthermore, the timing between oocyte retrieval and fertilization, and the culture time of the embryo before transfer differ between individuals and between ART clinics. This discrepancy is not always accounted for in the registries. Despite these limitations, however, determining gestational age based on the date of fertilization or implantation has been proposed as a valuable approach to validate gestational age estimated by LMP or ultrasound (116). Several studies have shown a high correlation between estimates of gestational age based on the time of IVF and those based on ultrasound measurements, although ultrasound-based estimates were 1-2-days shorter on average, with a difference of up to 14 days between the two methods in individual fetuses (117, 118).

2.4.4 Associations between DNA methylation and gestational age

Several EWASes conducted in the last decade have clearly shown that gestational age is strongly associated with DNA methylation at thousands of CpG sites throughout the genome (**Table 1**). In 2011, Schroeder *et al.* conducted the first study investigating the relationship between DNA methylation and gestational age in healthy neonates (119). They identified numerous candidate genes associated with gestational age, including several genes implicated in the timing of delivery or

postnatal outcomes (119). Several other studies have found associations in CpGs located in genes or regions that are implicated in a wide array of developmental processes and the regulation of epigenetic patterns (120, 121). In 2016, Bohlin *et al.* identified 5,475 CpGs associated with gestational age using newborn samples from the Norwegian Mother, Father, and Child Cohort study (MoBa) (122). The largest EWAS on gestational age to date is a large meta-analysis of 450K methylation data in 3,648 newborns from 17 cohorts, in which 8,899 CpGs were found to be associated with gestational age (123). Most studies investigating the relationship between gestational age and DNA methylation have used cord blood samples or dried blood spots, but there are also a few studies that have found substantial DNA methylation changes with gestational age in placental samples (81, 82, 124).

Table 1. EWASes of gestational age based on DNA methylation data generated from cord blood or dried blood spots.

Study	Year	Sample size	Platform	Significance threshold	Significant associations	Reference
Schroeder <i>et al.</i>	2011	259	27K	FDR < 0.05	41 CpGs	(119)
Lee <i>et al.</i>	2012	141	CHARM 2.0**	FDR < 0.05	3 regions	(120)
Parets <i>et al.</i>	2013	50	450K	FDR < 0.05	9,637 CpGs	(121)
Simpkin <i>et al.</i>	2015	914	450K	$p < 1.03 \times 10^{-7}$	224 CpGs	(125)
Bohlin <i>et al.</i>	2016	1,068	450K	Bonferroni	5,475 CpGs	(122)
Knight <i>et al.</i>	2016	1,434	27K	FDR < 0.05	3,155 CpGs	(126)
Hannon <i>et al.</i>	2019	1,316	450K	$p < 1 \times 10^{-7}$	4,299 CpGs	(127)
York <i>et al.</i>	2020	124 + 378*	450K	FDR < 0.05	2,372 CpGs	(128)
Merid <i>et al.</i>	2020	3,648	450K	$p < 1.06 \times 10^{-7}$	8,899 CpGs	(123)

Abbreviations: 27K, Infinium HumanMethylation27 BeadChip; CHARM, Comprehensive high-throughput arrays for relative methylation; 450K, Infinium Human Methylation 450K BeadChip; FDR, false discovery rate.

* Two different cohorts were analyzed individually in this study.

** CHARM 2.0 is a customized microarray method covering 5.2 million CpGs arranged into probe groups (120).

Different EWASes of gestational age have shown that there is a large overlap in CpG sites and associated genes. Several of these studies reported a higher proportion of *hypomethylated* CpGs (i.e., CpGs with decreasing methylation levels) amongst those associated with gestational age (119, 122, 123, 125). In addition, the gestational-age associated CpGs in cord blood seem to be relatively depleted in CpG islands and promoter regions (121, 123, 129), although the opposite seems to be the case for placental samples (124).

Most of the associations between DNA methylation and gestational age seem to be restricted to the perinatal period, although a smaller proportion of gestational-age related CpGs also appears to be

associated with postnatal aging (123). A longitudinal study of approximately 950 individuals from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort from the UK (130) indicated that the association between DNA methylation (450K-derived data) and gestational age fades away by early childhood (125). Another study comparing DNA methylation (also based on the 450K platform) between extreme preterm and term newborns also found widespread methylation differences between the groups that were largely resolved by 18 years of age (129).

2.5 Aging

2.5.1 What is aging?

A person's age is usually reported as the time that has passed since birth, which is also commonly referred to as the person's chronological age. However, aging is not merely the passing of time but is inherently linked to biological changes and the functional capability of the organism. Thus, aging can be loosely defined as the time-dependent decline in functional capacity across the lifespan. Although these types of changes are highly correlated with chronological age, they do not always happen at the same rate as chronological aging and are often referred to as biological aging (131). These biological aging processes may include a continuation of development, damage accumulation, cumulative mutational load, decreased fitness, and increased functional and cognitive decline. A recent review proposed the following twelve molecular, cellular, and systemic hallmarks of aging: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, disabled macroautophagy, deregulated nutrient-sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, altered intercellular communication, chronic inflammation, and dysbiosis (disruption of the microbiome) (132). These hallmarks are further grouped into three categories: (i) the primary hallmarks that cause damage, (ii) antagonistic hallmarks that respond to and compensate for the damage induced by primary hallmarks, and (iii) integrative hallmarks that are responsible for aging phenotypes when the damage accumulation caused by the primary and antagonistic hallmarks cannot be compensated for anymore. These three interconnected groups of hallmarks point to aging as being the result of the accumulation of multiple types of molecular damage and dysfunction due to a diminished damage-repair capacity. Epigenetic alterations thus represent a primary hallmark of aging (132).

2.5.2 The role of DNA methylation in aging

Aging is associated with a range of epigenetic changes, including aberrant chromatin remodeling, abnormal modification of histones, and alterations in DNA methylation patterns (133, 134). Early studies reported a global loss of methylation with increasing age (135), but conflicting evidence from later studies suggests that the overall effect of global DNA methylation changes might be dependent

on the methods used or the specific tissue under study (136). Several EWASes have revealed predictable and consistent shifts in average DNA methylation level at specific CpGs with age (137, 138, 139). However, the functional consequences of DNA methylation alterations are not clear. Altered DNA methylation patterns have been found in many age-related diseases such as cardiovascular disease (140), type 2 diabetes (141), Alzheimer disease (142), and cancer (143). Some of these altered DNA methylation patterns overlap with age-associated differentially methylated CpGs (138, 141, 143, 144). Although tissues and cells have unique DNA methylation aging signatures, there are conserved DNA methylation changes across cell types during aging (145, 146). CpG-rich regions tend to become *hypermethylated* (i.e., more methylated) with age, especially in the promoters of key developmental genes harboring both active and inactive histone marks (138, 147). *Hypomethylation*, on the other hand, occurs mostly in regions of low CpG density, often at introns or intergenic regions harboring active histone marks associated with enhancers (148). Compared to hypermethylation, patterns of hypomethylation are less conserved across tissues, possibly owing to the role of enhancers in tissue-specific gene expression (63, 148).

DNA methylation is central to the field of epigenetic reprogramming and rejuvenation, which aims to reset epigenetic patterns to youthful states by reversing cellular age. One such strategy is the overexpression of four genes coding for the transcription factors OCT4, SOX2, KLF4 and MYC that are collectively known as Yamanaka factors (149). Induction of these transcription factors can make the cell regain its pluripotency, reverse age-related DNA methylation signatures, yield a younger transcriptome, and even improve tissue-function in older individuals (150, 151). TET demethylating enzymes were shown to be necessary for this reprogramming to occur, suggesting that DNA methylation changes play a fundamental role in the aging process and its reversal (151).

2.6 Epigenetic clocks

2.6.1 Using DNA methylation to predict age and gestational age

‘Epigenetic clock’ refers to an innate biological process that gives rise to age-related DNA methylation changes and plays a purposeful role in development and aging (152). These age-related changes make DNA methylation a good candidate for a biomarker of aging. Some CpGs change almost linearly with age or gestational age and can therefore be used for prediction (153). Such a prediction model can be constructed by measuring the DNA methylation levels in CpGs throughout the genome and using statistical methods such as supervised machine learning to choose the set of CpGs providing the best predictive ability. Because of their great clock-like precision in predicting age, these models are themselves often called ‘epigenetic clocks’, and the predicted age is called ‘epigenetic age’ or ‘DNAm age’ (152).

2.6.2 A brief history of epigenetic clocks

In 2011, Bocklandt and colleagues published the first epigenetic predictor for age from 27K-derived DNA methylation data from saliva samples (154). They showed that it was possible to use DNA methylation levels from only two CpGs to construct a prediction model that was able to explain 73% of the variance in age and predict the age of an individual with an average accuracy of 5.2 years (154). Two years later, the ‘Hannum clock’, which included 71 CpGs, was developed on DNA methylation levels quantified by 450K in whole blood samples (155). This clock showed an even higher correlation with age ($r = 0.91$) and an average accuracy of 4.9 years. The Hannum clock also demonstrated a strong predictive power for chronological age in several other tissue types (breast, kidney, lung, and skin), although each tissue showed a different intercept and slope than expected from the blood-based model (155). In the same year, Horvath developed a pan-tissue clock using 8000 samples from 82 different datasets comprising 51 different tissues and cell types (145). This clock consisted of 353 CpGs and performed very well in heterogeneous tissues as well as in individual cell types. Later, in 2018, Horvath and colleagues developed another clock that was optimized for skin and blood samples, as the previous pan-tissue clock was suboptimal for some of these tissue types (156). In addition to the Bocklandt clock, which consisted of two CpGs, a couple of models developed subsequently comprised only a few CpGs (153), including a striking example from Garagnani and colleagues (157) showing a correlation of 0.92 between chronological age and hypermethylation of the ‘ELOVL fatty acid elongase 2’ (*ELOVL2*) gene.

Because clocks that were mainly trained on adult samples showed a lower accuracy in pediatric samples (158), and the rate of DNA methylation change is greater in children and adolescents compared to adults (159), epigenetic clocks that are more specific for pediatric populations were subsequently developed (160, 161).

The clocks mentioned thus far were developed for predicting chronological age and are labeled ‘first generation’ epigenetic clocks. A second generation of epigenetic clocks was developed to target phenotypic age more specifically and capture more biologically relevant measures of physiological dysregulation. This was done by replacing chronological age with a surrogate measure of biological age that differentiates mortality and morbidity risk among individuals of the same age, as demonstrated by Levine and colleagues’ PhenoAge clock from 2018 (162). Another second-generation clock, GrimAge, was developed in a similar manner for predicting lifespan (163).

More recently (in 2020), Belsky and colleagues developed an algorithm called ‘Dunedin pace-of-aging methylation’ (DunedinPoAm) for quantifying the pace of biological aging, which was updated in 2022 (164, 165). The updated clock, DunedinPACE, is based on a longitudinal study where a wide array of biomarkers of health and disease as well as various blood biomarkers were measured in 954 individuals born the same year. The authors modelled change-over-time across 12 years of follow-up.

The rates of change were composited to form a score for aging-related decline which was subsequently used to train the epigenetic clock (164).

The clocks developed thus far have been based on an average DNA methylation level in a heterogeneous cell population, but Trapp and colleagues (166) published a clock in 2021 that could predict epigenetic age at single-cell resolution, providing novel insights into the heterogeneity in aging of individual cells. While the vast majority of epigenetic clocks have been based on regularized linear regression, de Lima Camillo and colleagues (167) used deep-learning approaches to develop a multi-tissue clock that could take non-linear interactions into account. Epigenetic clocks for a range of different species have recently been developed (168, 169, 170, 171, 172, 173, 174), and in a study currently deposited in the open-access preprint repository, bioRxiv, the authors propose a universal clock for mammals, suggesting that the underlying principles of epigenetic clocks are evolutionarily conserved (175).

In summary, a large body of work conducted over the last decade has provided new insights into how the methylome can predict age and how this feature can be exploited to gain a deeper understanding of the mechanistic underpinnings of aging and healthspan, lifespan and mortality, as well as development and evolution.

2.6.3 Epigenetic age is associated with a range of conditions and diseases

When age is predicted using an epigenetic clock, some individuals will exhibit an epigenetic age that is higher or lower than their chronological age. This discrepancy is often called epigenetic age acceleration (EAA) which is usually defined as the residuals from a regression of epigenetic age on chronological age. In particular, EAA calculated from several of the published clocks has been associated with a range of age-related conditions and diseases, including diabetic complications (176), Down syndrome (177, 178), Parkinson (179), Hutchinson Gilford Progeria Syndrome (156, 180), Alzheimer's disease (181), cancer (182), coronary heart disease (162), centenarian status (183), physical and cognitive fitness (184), and life expectancy (185).

2.6.4 Epigenetic clocks specific for the gestational period

Although DNA methylation is strongly associated with both gestational age and chronological age, epigenetic clocks trained on samples from children or adults do not perform well in predicting gestational age in newborns. In most of these clocks, newborn samples were either not included at all or were assigned an age of zero, and thus did not attempt to differentiate between different gestational ages (145, 155, 160). The Skin & blood clock takes gestational age into account by assigning negative ages to those samples, but it is still not able to predict gestational age with a similar precision as adult

age (156). Therefore, there was a great need for separate epigenetic clocks that are specific for the gestational period.

In 2016, two separate cord-blood-based epigenetic clocks specific for gestational age were published in the same issue of the journal *Genome Biology*: the Bohlin (122) and Knight (126) clocks. The Bohlin clock was trained on 1,068 450K-derived DNA methylation samples from a subset of MoBa (186). When trained on ultrasound-estimated gestational age, 96 CpGs were selected for being predictive of gestational age. The clock was subsequently tested in another subset of 685 newborns from MoBa and showed a precision of $R^2 = 0.66$ and a standard error (SE) of ± 12.5 days (95% prediction interval). The authors also developed a clock trained on LMP-estimated gestational age, which selected fewer CpGs for prediction (56 CpGs) and did not perform as well as the ultrasound-based clock ($R^2 = 0.5$, SE ± 14.9 days).

The clock developed by Knight *et al.* was trained on 207 DNA methylation samples from six independent cohorts encompassing 16,676 CpG sites from the 27K and 450K arrays. Both cord-blood and dried-blood spot samples were used. In this clock, 148 CpGs were selected as being predictive for gestational age. The clock was subsequently tested in 1,134 samples from six independent datasets, showing an overall correlation of 0.91 and a median absolute deviation of 1.24 weeks.

Although both clocks showed a good predictive performance in their test sets, there are several marked differences between the Bohlin and Knight clocks, as highlighted in a commentary by Simpkin *et al.* (187). Knight and colleagues used data from both 27K and 450K arrays, ending up with substantially fewer CpGs to select from (16,838 compared to 473,731 in the Bohlin *et al.* study), which may partly explain the lack of overlap in CpGs selected for the two clocks (only two CpGs were in common between the two clocks). Knight *et al.* also included a wider range of gestational age and ancestries in their training data compared to Bohlin *et al.*; specifically, they included more preterm newborns. The latter may have a large impact on prediction performance in the sample set one would like to study, depending on the gestational age range in the samples (187). Finally, Knight *et al.* included substantially fewer samples in their training set ($n = 207$), saving most of their data for testing, whereas the Bohlin clock was trained on over five times more samples than the Knight clock ($n = 1,068$). In general, using more samples for training creates a more robust prediction model that is less prone to overfitting, especially when the number of CpGs in the clock is nearly as large as the number of training samples (187).

Aside from cord-blood and dried-blood spot samples, epigenetic clocks based on DNA methylation in placental tissue have also been developed. In 2017, Mayne and colleagues (188) built a clock consisting of 62 CpG sites using 27K DNA methylation data from 170 placental samples. That clock predicted gestational age in the test samples with a correlation of 0.95. Later, Lee *et al.* (124) published three different placental clocks based on a larger training set of 1,102 samples. They also

showed that the placental clocks were fundamentally different from both adult-age clocks and cord blood-based clocks for gestational age (124). Falick Michaeli *et al.* (189) performed reduced representation bisulfite sequencing (RRBS) to develop gestational age clocks from both cord blood and placental samples. RRBS is a very different method from the Illumina arrays, covering different loci in the genome. Thus, although they were able to develop a relatively precise clock using RRBS ($r = 0.77$), it is difficult to compare their clock with the clocks developed using microarray samples.

Other relevant clocks include one developed by Steg *et al.* (190), which was trained on a dataset of 193 fetal brain samples. That clock predicted gestational age in fetal brain samples as well as cellular stem cell models and derived neurons (190). Graw *et al.* (191) developed clocks for estimating post-menstrual and postnatal age in preterm infants. This was based on buccal cell tissue and was only compared to pediatric and adult clocks based on skin/buccal cells, not gestational age clocks.

2.6.5 Epigenetic gestational age as a proxy for developmental maturity

Although DNA methylation patterns can predict a newborn's gestational age very accurately, there is still a lack of understanding regarding why these patterns change so predictably with gestational age and whether variation in these patterns have any biological cause or consequence. It has long been proposed that the gestational age clocks track the development of the fetus and mirror its developmental maturity (126). Just as EAA represents the discrepancy between a person's epigenetic age and his/her chronological age, it is possible to calculate gestational age acceleration (GAA) for a newborn, defined as the discrepancy between epigenetic gestational age and clinically-estimated gestational age (126). It has been hypothesized that if the epigenetic gestational age is indeed a proxy for the newborn's developmental maturity, GAA could function as an instrument to assess the relationship between epigenetic developmental maturity and diverse exposures and outcomes (126).

Exploration of GAA in newborns was first demonstrated by Knight *et al.* in 2016, where the authors found an association between positive GAA (i.e., the epigenetic gestational age is *higher* than the clinically-estimated gestational age) and both birthweight percentile and birthweight itself (126). Curiously, they also found an association between GAA and maternal insurance status. Later studies have found associations between positive GAA and a range of maternal conditions and exposures, such as maternal age (192), pre-eclampsia (192), fetal demise in a previous pregnancy (192), treatment with the inflammation-reducing corticosteroid betamethasone in pregnancy (192), pregnancy fatty acid status (193), higher maternal plasma homocysteine concentrations (101), pre-pregnancy maternal overweight and obesity (194), and maternal smoking (195, 196). In addition, birth length (195, 197), lower 1-min Apgar score (192), reduced risk of needing respiratory interventions, and lower bronchopulmonary dysplasia rate (198) have all been associated with a positive GAA.

Negative GAA (i.e., the epigenetic gestational age is *lower* than the clinically-estimated gestational age) has also been associated with maternal factors and exposures, including maternal serum triglyceride levels (199), high maternal serum lipid levels (199), D3 supplementation (200), gestational diabetes in a previous pregnancy (192), air pollution exposure (201), neighborhood adversity (105), and higher parity (199). Cord serum B12 concentrations have also been associated with a negative GAA (202). Two separate studies found an association between maternal depression and negative GAA (203, 204).

Thus far, only few studies have assessed the relationship between GAA and outcomes later in life. Suarez *et al.* (203) found that a negative GAA was associated with internalizing problems in early childhood in boys, and that GAA partly mediated the effect of maternal depression on this outcome. Monasso *et al.* (205) investigated associations between GAA and cardiovascular outcome in 10-year-old children and did not find any associations with blood pressure, carotid intima-media thickness or carotid distensibility. Bright *et al.* (197) found that the association between positive GAA and birth size (weight and length) persisted until nine months of age, but that the association of GAA and weight reverses from age five years onwards, such that by age 10 years, positive GAA is associated with lower childhood weight.

The evidence for GAA-associations with newborn sex and birthweight is somewhat conflicting. The initial result from Knight *et al.* showing increased birthweight with positive GAA was later replicated in several studies (194, 197, 200), whereas Girchenko *et al.* (126) found significant associations between negative GAA and all measures of birth size. Further, Girchenko *et al.* also reported an association between female sex and positive GAA, while Khouja *et al.* (194) reported an association between male sex and positive GAA. A third study found an association between male sex and negative GAA (199).

The results from studies assessing the significance of GAA for diverse exposures and outcomes are thus far more inconclusive than for EAA. This could be due to the relative accuracy of clinical gestational age estimates used in the different studies (193). As mentioned previously, there can be a large gap between ultrasound-based and LMP-based gestational age estimation for the same individual, which means that the choice and accuracy of gestational age estimate may have a large impact on the calculated gestational age acceleration. In addition, the choice of clock may influence the analyses (187, 193). Different gestational age clocks have few overlapping CpGs and may track different biological processes. This particularly applies to clocks developed from different tissues. Dieckmann *et al.* (195) compared epigenetic gestational age predictions in cord blood and placental tissue from the same individuals and found that GAA was not correlated across tissues. The precision of gestational age prediction may also have an impact the associations with phenotypes of interest, since a very precise clock would display a smaller deviation between epigenetic gestational age and

clinically-estimated gestational age (206). Finally, the way GAA is calculated often differs between studies and may affect the results (194). In most of the previously mentioned studies, GAA was calculated using the residuals from a regression of epigenetic gestational age on clinically-estimated gestational age, while in some studies GAA was defined as the raw difference between epigenetic gestational age and clinically-estimated gestational age (126, 192).

3. Aims of the thesis

The overarching aim of this thesis was to explore genome-wide DNA methylation levels in newborns in relation to their gestational age and identify mechanisms that may explain this association. The individual aims of each of the papers included in this thesis were as follows:

Paper 1: Investigate cell-type specific association between gestational age and DNA methylation in cord blood.

Paper 2: Develop cord blood-based epigenetic gestational age clock specifically for the Illumina EPIC array and investigate whether the additional probes on EPIC improve the prediction performance compared to clocks developed on 450K array data. Furthermore, evaluate the precision and accuracy of the new clock using the embryo transfer date of newborns conceived with the use of ART, and explore differences in epigenetic gestational age between ART-conceived newborns and newborns conceived naturally.

Paper 3: Identify CpGs that are stably predictive of gestational age in cord blood and determine whether stably selected CpGs can be used to build an even more parsimonious gestational age clock.

4. Methodological considerations

4.1 Sample collection

4.1.1 Datasets

We used three different subsamples of cord-blood DNA methylation data from the Norwegian Mother, Father, and Child Cohort Study (MoBa) (186), and one cord-blood DNA methylation dataset from the Finnish Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction (PREDO) study (207). **Table 2** summarizes these datasets with respect to the three papers. All newborns included in this study were singletons. Further details regarding each of these datasets can be found in the Methods section of the respective papers.

Table 2. Datasets used in the current study.

Dataset	Sample size*	Source	DNAm array	Paper
START	1,794	MoBa	EPIC	1,2,3
MoBa1	1,062	MoBa	450K	1
met008	1,182	MoBa	EPIC	3
PREDO	148	PREDO	EPIC	2

Abbreviations: DNAm, DNA methylation; START, SStudy of Assisted Reproductive Technology; MoBa, Mother, Father, and Child Cohort Study; PREDO, Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction Study

* Total sample size used in the work of this thesis.

MoBa is a prospective pregnancy cohort study in which pregnant women attending a routine ultrasound examination around gestational week 18 between 1999 and 2008 were recruited (186). The cohort includes more than 140,000 children, 95,000 mothers and 75,000 fathers. Several questionnaires were sent out to the participants during and after pregnancy. Furthermore, several different biological samples were obtained from the participants, including umbilical cord blood that was taken immediately after birth and frozen at -80°C (208).

The Study of Assisted Reproductive Technology (**START**) is a subsample from MoBa consisting of a random selection of 992 mother-father-child trios with naturally conceived children and 978 trios with children conceived with the use of ART (93). DNA methylation was quantified using the EPIC array. This subsample was used in all three papers, with slightly different sample sizes in each paper according to the availability of the covariates that were needed for each specific study. In **Paper 1**, 953 samples from naturally conceived newborns were included. In **Paper 2**, 955 samples from naturally conceived newborns, and 838 samples from ART conceived newborns were included. Finally, in **Paper 3**, 956 samples from naturally conceived newborns were included.

The **MoBa1** samples were selected from a substudy of MoBa that evaluated the association between maternal plasma folate during pregnancy and childhood asthma status at three years of age (209). DNA methylation levels from 1,062 newborns in this study were quantified using the 450K array (210). This subsample was used in **Paper 1**.

met008 is a random subsample of 1,186 newborns from MoBa that was selected based on the same criteria as those for START except that the mode of conception was not included as a criterion (93). Like START, DNA methylation was also quantified using the EPIC array. Four samples overlapped with START and were therefore excluded, resulting in 1,182 samples of this subset being included in **Paper 3**.

PREDO is a prospective pregnancy cohort of 1,079 Finnish women who gave birth to a singleton live child between 2006 and 2010 (207). 148 cord blood samples with DNA methylation levels quantified using EPIC were available for analysis and were included in **Paper 2**.

4.1.2 Clinical estimations of gestational age

As explained in chapter 2.4.3, although the true gestational age of a newborn cannot be known with absolute certainty, there are different approaches to estimating gestational age, each with its own set of advantages and limitations. For all the papers in this study, we used estimations of gestational age extracted from the Medical Birth Registry of Norway (MBRN). Most newborns in MBRN have gestational age estimated both from the mother's LMP and from ultrasound measurements, but a few individuals lacked one or the other. To avoid any systematic error due to differences in estimation method, we only analyzed data on newborns for whom an ultrasound-based gestational age estimate was available. Another reason for our preference for gestational age estimated from ultrasound measurements is that they are generally considered to be more accurate than LMP-based estimates. Notably, Bohlin and colleagues demonstrated that ultrasound measurements were more strongly associated with DNA methylation than LMP-based estimates (122). Estimates from the Knight clock also correlated more strongly with ultrasound-based estimates than those based on LMP (126).

For ART pregnancies in **Paper 2**, we also used a second estimate of gestational age recorded in MBRN that was based on the date of egg retrieval and/or embryo insertion. For most of the ART pregnancies, this estimate was based on the date of egg retrieval plus 14 days (as a proxy for the time interval between LMP and conception). When the date of egg retrieval was not known, the date of embryo insertion minus two days was used instead. For frozen embryos, we used the date of embryo insertion plus 14 days, and the number of days between egg retrieval and freezing. These three estimations of gestational age were combined into a variable called embryo transfer date (ETD).

4.2 DNA methylation profiling and data preparation

4.2.1 BeadChip arrays for DNA methylation quantification

DNA methylation was measured using Illumina BeadChip arrays. These arrays use bead technology to measure DNA methylation status for a given number of probes at specific positions in the genome as an average over all cells included in the input. In this study, we used two different methylation arrays: EPIC, covering 863,904 CpGs (46), and 450K, covering 482,421 CpG sites (45).

Briefly, the laboratory workflow of the Illumina Infinium BeadChip includes the following steps (45): First, bisulfite conversion is applied to DNA isolated from a given individual. Sodium bisulfite converts unmethylated cytosines to uracils but leaves methylated cytosines unchanged. The DNA is then amplified, fragmented, and precipitated before it is hybridized to probes on the BeadChip. Each bead on the chip contains a 50-base probe and a 23-base address to identify its physical location on the BeadChip. The probe sequences are designed to be complementary to specific DNA regions containing a CpG site at the 3' end of the probe. Hybridization of the bisulfite-converted DNA incorporates a fluorescently labeled dideoxynucleotide triphosphate (ddNTP) that can differentiate between a methylated and unmethylated signal. The fluorescence signal is then quantified. The resulting methylation β -value is defined by the following formula:

$$\beta = M/(M+U+100) \quad (1)$$

Where M is the intensity of the methylated signal and U is the intensity of the unmethylated signal. The number 100 is a constant added to avoid having zero in the denominator. A β -value of 0 represents an entirely unmethylated CpG site and a β -value approaching 1 represents a completely methylated CpG site.

The Illumina Infinium platforms include two different types of probe design (45). Type I probes have two separate probe sequences per CpG site, whereas Type II probes have only one probe sequence per CpG site. While Type I probes can measure methylation at more CpG dense regions, Type II probes use half the physical space on the BeadChip compared to Type I. These probe type differences need to be accounted for during the quality control (QC) process (described in the next chapter).

4.2.2 Quality control of microarray-derived DNA methylation data

After quantifying DNA methylation levels using BeadChip arrays as described above, it is critical to assess the reliability of each data point by performing a series of QC steps on the data before conducting any downstream analyses. The specific QC pipeline may vary between different studies, but a few of the main points are elaborated below.

First, the total fluorescence intensity for each probe in each sample is evaluated by computing the detection p-values and applying a cutoff to remove those data points with too large detection p-values

(otherwise, these cannot be distinguished from noise signals). Samples or probes with a high proportion of large detection p-values are excluded from the data during the QC process. Further, probes that are located near a single-nucleotide polymorphism (SNP) or are cross-hybridized are also removed. Samples displaying a sex mismatch, determined by using methylation signals from the sex chromosomes, are either flagged out for further verification or excluded altogether. After the exclusion of probes and samples, the next step is background correction and normalization of probe types. Background correction is performed to minimize the background noise in intensities. Normalization is applied to reshape the distributions of intensities across probe types and samples so that they are comparable to each other. Finally, some pipelines address systematic technical variation in the form of batch effects, but in our analyses, we included batch (plate) instead as a variable to be adjusted for in the statistical model. Additionally, the positioning of the case (naturally conceived newborns) and control (ART-conceived newborns) samples on the 96 well plates was randomized to minimize systematic technical variation between the samples. More details on the specific QC pipelines for the datasets used in this study are provided in the Methods sections of the respective papers.

The choice of preprocessing method and QC pipeline may have consequences for downstream analyses. A study investigating the effect of different preprocessing methods in epigenetic age estimation in adults found that epigenetic age was highly correlated across preprocessing methods, but that different methods could lead to a systematic offset in the age estimate (211). In association studies, the choice of normalization approach may influence reproducibility and variability of the data, which are largely dependent on the strength of the signal (i.e., its p-value) (212).

4.3 Epigenome-wide association studies (EWAS)

4.3.1 EWAS

EWASes are commonly used to investigate the association between a phenotype and epigenetic variants, most commonly, DNA methylation level differences. Different study designs can be employed in an EWAS, such as case-control, longitudinal, quantitative trait or family-based study designs, and the most common platform for quantifying DNA methylation levels in an EWAS involves the use of microarrays such as those described above (see also (213)). The analytic workflow in an EWAS involves running one regression at a time for each of the CpGs interrogated in the study to search for associations with the phenotype of interest. In **Paper 2**, we used EWAS as a tool to search for epigenome-wide associations with gestational age.

4.3.2 Statistical power and significance thresholds in EWAS

An EWAS must have sufficient statistical power to identify true associations between DNA methylation level at a given CpG site and a phenotype of interest. Statistical power is the probability

that a statistical test rejects the null hypothesis when the alternative hypothesis is correct (true positive), and in an EWAS setting, statistical power is affected by the sample size of the cohort, the chosen significance threshold, and the effect size of the CpG. The most straightforward way to improve power is to increase sample size. A sample size of 1,000 (e.g., 500 cases and 500 controls) has been recommended for case-control or family-based EWASes (214, 215), but there does not seem to be *a priori* consensus regarding the minimum sample size required for an EWAS investigating a quantitative phenotype like gestational age. Several power calculators have also become available, but these are mostly restricted to case-control studies (214, 216).

As a vast number of CpGs are routinely analyzed in an EWAS, significance thresholds for controlling the number of false positives are impacted by multiple testing. A Bonferroni correction can be applied to the total number of CpGs analyzed to adjust for multiple testing, but this is widely considered to be too conservative for an EWAS due to the high correlation in DNA methylation levels at CpG sites across the genome, which reduces the actual number of independent tests (214). An alternative approach is to apply a false discovery rate (FDR) threshold to the p-value distribution, which provides a more balanced compromise between false positives and false negatives by identifying the top associated sites relative to the threshold. FDR is commonly applied using the approach described by Benjamini and Hochberg (217), which makes the assumption that the p-values are independent and uniformly distributed under the null hypothesis (214). However, this uniform distribution does not necessarily apply to p-values across CpGs analyzed in an EWAS and may limit reproducibility of results across studies and lead to severely inflated test statistics (213). Thus, several different methods have been developed to control for such inflation (218, 219). As previous studies have shown that the DNA methylation levels of many CpGs are significantly associated with gestational age, we applied Bonferroni correction to restrict the number false positive associations. Although this may come at a cost of having a higher number of false negatives, it may generate more robust results with a higher level of reproducibility.

4.3.3 Confounding factors

Several environmental and phenotypic factors can directly confound EWASes by affecting both the methylome and the phenotype of interest (8). One way of controlling for such confounders is by stratification or adjustment in the statistical analyses. A useful tool for assessing the relationship between the exposure, outcome and potential confounders, and thus decide which variables should be adjusted for, is the directed acyclic graph (DAG) (220). However, a DAG is only a simplified illustration of the relationship between the variables and must therefore be interpreted with caution. This especially applies to epigenome-wide studies where the relationship between the variables may vary between different CpGs. **Figure 5** shows a simplified DAG illustrating the relationship between DNA methylation, gestational age, and other relevant variables.

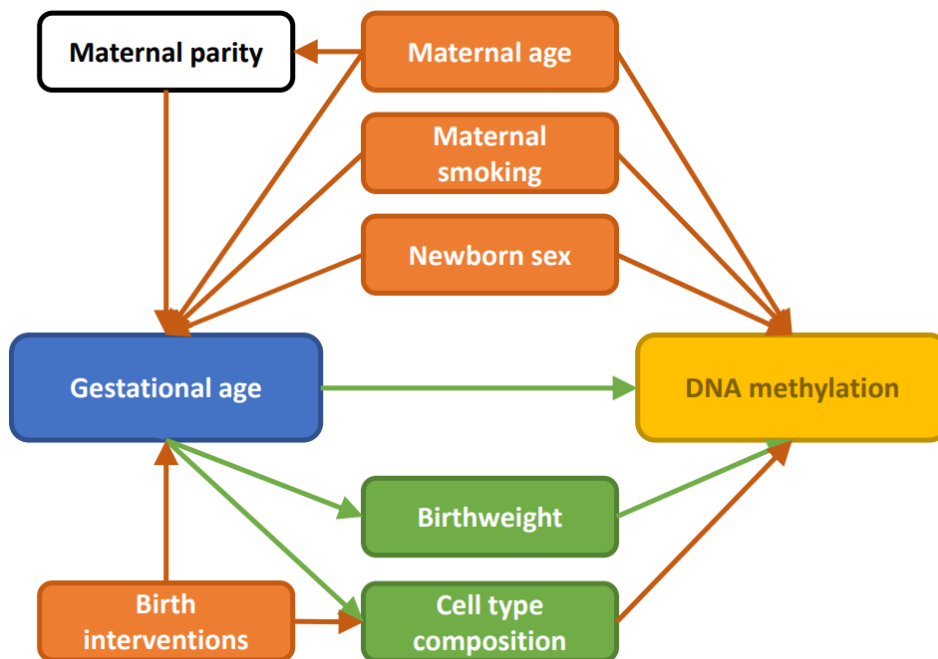


Figure 5. A Directed Acyclic Graph (DAG) to illustrate the relationship between gestational age (exposure, in blue) and DNA methylation (outcome, in yellow), and other relevant variables. In the DAG, green arrows mark “causal” paths, whereas orange arrows mark confounding paths. Variables in orange boxes are considered confounding factors, whereas variables in green boxes are considered mediating factors. Maternal parity (in white and black) is on a confounding path, but due to its assumed relationship with maternal age, it does not need to be adjusted for in the statistical model. Birth interventions (e.g., birth induction and cesarean section) are considered cofounders in this DAG, but due to their assumed relationship with cell type composition, they do not need to be adjusted for as long as we adjust for cell-type composition.

The variables we chose to adjust for in our study (in addition to batch (plate) and cell-type composition) were newborn sex, maternal smoking, and maternal age. Sex is known to be strongly associated with cord blood DNA methylation on autosomal chromosomes (221). There are also differences in growth patterns and risk for preterm birth between the sexes (222, 223, 224). Furthermore, maternal smoking is robustly associated with DNA methylation alterations in the newborn (225) and risk of preterm birth (226). Maternal age is also associated both with newborn DNA methylation (227) and risk of preterm birth (228).

Birthweight is another trait that is associated both with DNA methylation (229) and gestational age (230). Although previous studies of gestational age adjusted for birthweight (123, 125), we did not do so in our main analyses. As birthweight may be viewed as a composite of fetal growth and the length of gestation, it can act as a mediator of the association between gestational age and DNA methylation (**Figure 5**). Thus, adjustment for birthweight would remove part of the gestational age signal. We did, however, adjust for birthweight in a sensitivity analysis, but this did not change our conclusions. Because gestational age and birthweight are so closely correlated, it is difficult to disentangle their

separate effects on DNA methylation. Therefore, we did not consider this to fall within the scope of our study.

Parity is another variable that has been included in a previous study of gestational age (123), and that has been associated with maternal age and risk of preterm birth (231). Although parity is associated with maternal DNA methylation changes (232), such associations have not been shown for newborn DNA methylation.

Medical birth interventions, such as induction of labor and c-section, may have an impact on gestational age at birth. They have thus either been adjusted for (122) or treated as an exclusion criterion (123) in previous studies of gestational age. C-section is associated with DNA methylation in the newborn, but this association may be largely explained by differences in cell-type proportions (233).

Furthermore, cell-type heterogeneity is a well-known confounder in epigenetic analyses, which I will describe in more detail in the next chapter.

4.4 Cell types

4.4.1 Cell types in cord blood

The work presented in this thesis was based on DNA methylation data from cord blood samples. Cord blood consists of several different cell types, usually divided into eight main categories: B-cells, CD4+ T-cells, CD8+ T-cells, granulocytes, monocytes, natural killer (NK) cells, nucleated red blood cells (nRBCs), and enucleated red blood cells (erythrocytes). B-cells, T-cells, granulocytes, monocytes and NK cells are all leukocytes, or white blood cells, whereas nRBCs and erythrocytes are red blood cells. Each of these cell-types can be further divided into more specific subtypes. As erythrocytes do not contain a nucleus, the remaining seven cell-types are those that are relevant for DNA methylation analyses of cord blood. Cord blood differs from adult peripheral blood in many important aspects, including cell-type composition (234) and the level of immune cell maturity (235). Moreover, whereas cord blood has a considerable proportion of nRBCs, these cells are normally not present in adult peripheral blood (236).

4.4.2 Cell-type deconvolution

The methylation β -value derived from array-based quantification represents an average of DNA methylation levels for each specific CpG in all cells of a given tissue, which, in our case, is cord blood. It is important to bear in mind that different cell types have distinct DNA methylation patterns when DNA methylation level is quantified in tissues consisting of a mixture of several cell types. Some of the cell types may have a larger impact on the average DNA methylation measurement than

others, and intraindividual differences in cell-type composition may affect the results considerably if they are not accounted for in the analysis (237). To adjust for differential DNA methylation in mixed cell samples, it is common to include the proportions of the different cell types as variables in the statistical model.

Actual cell counts are not always available for each sample when analyzing DNA methylation data, and, consequently, several algorithms for *in silico* estimation of cell-type composition have been developed (238). These algorithms are either reference-based or reference-free. The reference-based methods have been developed by identifying differentially methylated regions (DMRs) that are specific for a given cell type in a reference dataset of purified cell populations. These references are based on cell-sorting methods such as fluorescence-activated cell sorting (FACS) and rely on carefully selected cell-type DMRs using sorted, purified cell populations from multiple subjects. The original reference-based algorithm using such DMRs was developed by Houseman *et al.* (239) in 2012. This algorithm models DNA methylation of a given sample as a weighted combination of the individual DNA methylation patterns of cell types in that sample. Reference-free algorithms, on the other hand, are based on unsupervised methods. In general, when a reference database of DNA methylation is available for the tissue type of interest, a reference-based approach is recommended over reference-free methods, but the latter can be useful when cell-type specific DMRs and reference datasets are unavailable (240, 241).

In our analyses, we used a reference-based method by Gervin *et al.* (242) based on the framework proposed by Houseman *et al.* (239). Currently, four cord blood references consisting of cell-type-specific DNA methylation array data have been published (243, 244, 245, 246). Gervin *et al.* (242) systematically evaluated and compared these datasets in 2019. By filtering and combining samples, they obtained a joint reference for cord blood – which is the one we have used in this study.

4.4.3 Cell-type proportions as covariates in DNA methylation analyses

The importance of adjusting for cellular heterogeneity in epigenetic studies is widely acknowledged (237, 247). As mentioned in the previous chapter, it is common to adjust for differential DNA methylation in mixed cell samples by including the proportions of the different cell types as variables in the statistical model. We show, in **Paper 1**, that there is considerable interindividual variation in cord blood cell-type composition, and that the proportions of several cell types are significantly associated with gestational age. Due to these observations, and the strong cell-type specificity of DNA methylation patterns, we included cell-type proportions as covariates in our EWAS of gestational age (**Figure 5**).

4.4.4 Cell-type specific analyses of DNA methylation

For some research questions, it may be sufficient to adjust for cell-type composition. However, merely including the cell-type proportions as linear covariates in the model does not tell us whether the associations are similar across all cell types, or if any of them are specific for a given cell type(s).

Determining if there are any cell-type specific associations between DNA methylation and gestational age is therefore critical in unraveling the mechanistic underpinnings behind such associations. Several methods for identifying cell-type specific DNA methylation differences in EWAS studies have recently been developed. These methods can be used on DNA methylation data quantified from heterogeneous tissues (like cord blood) but require information on the cell-type proportions for each sample. In our study, we have used two different methods for cell-type specific association analysis: CellDMC, developed by Zheng *et al.* (248), and Tensor Composition Analysis (TCA), developed by Rahmani *et al.* (249).

The key idea behind CellDMC is that DNA methylation alterations occurring in a specific cell type will exhibit a significant interaction between the corresponding cell-type proportion and the phenotype of interest (248). CellDMC incorporates multiplicative terms between the phenotype of interest and the estimated cell-type proportions in a linear modeling framework. By contrast, TCA exploits the available information on DNA methylation levels for each sample and the information on cell-type composition for each sample to infer a so-called ‘tensor’ of samples by CpGs by cell-types (249). In principle, TCA enables the inference of DNA methylation levels in each individual for each cell type, and subsequent analyses (such as an EWAS of GA) can then be performed on each cell type separately (an approach referred to as ‘two-stage TCA’ in **Paper 1**). However, for association testing, Rahmani *et al.* recommend using a marginal conditional approach instead, in which a model is first fitted to all cell types jointly and then the effect of each cell type is tested separately for its statistical significance (249). This approach, referred to as ‘one-stage TCA’ in **Paper 1**, is very similar to the one implemented in CellDMC, which is also based on marginal conditional tests (248). We, therefore, decided to include both one-stage and two-stage TCA in our analysis pipeline.

Furthermore, the authors of the TCA method stress the importance of making appropriate assumptions regarding the directionality of the model in cell-type specific DNA methylation analyses, because these assumptions can affect statistical performance, and, if incorrect, may result in spurious findings (250). In epigenetic association studies, one may hypothesize that DNA methylation (X) *affects* a condition of interest (Y), in which the directionality assumption can be designated as $X \rightarrow Y$, or, conversely, that it may *be affected* by the condition of interest (designated $Y \rightarrow X$). In our studies, we assumed that the latter directionality is the more plausible scenario describing the relationship between DNA methylation and gestational age, given that chronological gestational age cannot be altered by methylation changes. Thus, for one-stage TCA, we implemented the statistical model recommended by the original authors for the assumption $Y \rightarrow X$ (the function `tca` in their R package TCA). The

CellDMC method, on the other hand, does not accommodate the assumption $X \rightarrow Y$ and is restricted to $Y \rightarrow X$.

4.5 Prediction

4.5.1 Statistical prediction methods

A prediction is an inference made about the future, often based on current or past evidence. In statistics, this involves exploring systematic patterns in data to forecast outcomes. Although methods such as EWAS and CellDMC focus on understanding the relationship between a specific factor and an outcome of interest, and may thus provide novel insights into the DNA methylation-gestational age association, the associations detected in such studies do not necessarily have significant predictive power.

The main goal of a predictive method is to accurately predict what the response is going to be in relation to future input variables (251), and thus requires fewer assumptions of the relationship between the input and the response variables compared to methods used for studying association patterns. Predictive modeling can be used to identify a set of CpGs that can predict gestational age. Examples include penalized regression methods, such as lasso or elastic net, but also decision trees and neural networks. The focus of this thesis is on penalized regression methods, which are the type of methods most commonly used to develop epigenetic clocks.

4.5.2 Penalized linear regression

Predictive modeling using high-dimensional datasets, such as array-based DNA methylation datasets, can be challenging due to the markedly lower number of samples compared to the number of variables (CpGs) (252). When the number of variables in a model substantially exceeds the number of samples, the model might include random variation present in the data that does not represent true variation associated with the phenotype of interest (253). This is known as ‘overfitting’. One solution to this problem is to shrink the regression coefficients or set them to zero by imposing a penalty on their size and thereby decrease the variance (254).

Two commonly used penalized regression methods are lasso and ridge regression. The main difference between these methods is that lasso performs both parameter shrinking and subset selection by setting some of the coefficients to zero, whereas ridge regression only shrinks the coefficients and thereby keeps all the variables in the model (252). Thus, if a large fraction of the total set of variables is expected to be associated with the outcome, ridge regression should probably be favored, but if only a few variables are expected to be predictive of the outcome, lasso may be preferable. The two methods also differ in how they handle correlated variables (252). Whereas ridge regression shrinks correlated variables toward each other, lasso typically selects just one of them. Thus, ridge regression may

perform better in a set of highly correlated variables (252), which is typical of CpGs showing varying degrees of co-methylation patterns across the epigenome.

4.5.3 Elastic net and tuning of the penalty term

Elastic net is a method that contains a combination of lasso and ridge regression (255). It allows tuning of the penalty term through adjustment of the parameters α and λ , where α controls the *type* of shrinkage and λ the *amount* of shrinkage. Put simply, an α value of 0 corresponds to ridge regression, while an α value of 1 corresponds to lasso regression. Setting α between the values of 0 and 1 gives a penalty term that is dominated by the end point to which the α is closest (252). The tuning parameters α and λ can be chosen by k-fold cross validation. In such cross-validation, the training set is split into k equally sized subsets ('folds'), where all subsets except one are used for fitting the model which is subsequently used to estimate the prediction error in the left-out subset. The procedure is then repeated for a total of k times, each time leaving out a different subset. Finally, the prediction errors from all the subsets are merged and the optimal parameter value is identified. Typically, the parameters that give the smallest prediction error are chosen, but to obtain a more parsimonious model with fewer variables, the 'one standard error rule' can be applied, where the largest λ within one standard error of the minimum prediction error is chosen (252). By using the one standard error rule, as was done in **Paper 2** and **Paper 3**, one can produce a model with fewer variables and thus ease interpretation. However, the downside of using this rule is that it may also lead to increased prediction error and more biased parameter estimates (252).

4.5.4 Assessing prediction performance and external validity

The best way to properly validate a prediction model is to assess its performance in a dataset that was not included in the training of the model (253). For large enough studies, it is thus recommended to split the original dataset into three sets: a training set (consisting of 50% of the dataset), a validation set (25%), and a test set (25%) (254). Alternatively, one can split the original dataset into just two sets, a training set (70-80%) and a test set (20-30%), before running cross-validation and model selection on the training set. There are no clear-cut recommendations on how large the training and test set should be, but, in general, the precision of the model will increase with the number of samples in the training set (249). To create an appropriate prediction model, it is important to ensure that the data used for training and testing the model are representative of the population on which predictions are to be made.

Another important aspect to consider is the effect of confounders, which can make it difficult to interpret the results. When training and test datasets are generated together as part of the same study, an unknown confounder could affect the predictive performance in both the training and test set (253).

Thus, it is preferable to have a test dataset generated in an independent laboratory/core facility whenever feasible.

To assess the performance of the prediction model, i.e., the gestational age clock, clinically-estimated gestational age is regressed on gestational age predicted from DNA methylation data in the test set. We used MM-type robust linear regression in all three papers because it is less influenced by outliers than ordinary least squares (OLS) regression (256). Different metrics can be applied to assess the prediction performance. The R^2 statistic provides a measure of the proportion of variance explained, while the median absolute deviation (MAD) between clinically-estimated gestational age and gestational age predicted using DNA methylation data provides a measure of the accuracy of the predictions in terms of the difference in number of days.

When developing a new prediction model, such as an epigenetic clock, it may be of interest to compare its performance to that of previously developed epigenetic clocks. One way of assessing the size and significance of differences in performance between two different prediction models is to compute bootstrap confidence intervals for differences in R^2 , MAD, and standard error (SE) between the two models, as we did in **Paper 2**.

4.5.5 Drawbacks of penalized regression methods for variable selection

Although penalized regression methods are very useful for building accurate prediction models for gestational age, they have some limitations when it comes to variable selection. Ridge regression and elastic net with a penalty term close to ridge ($\alpha < 0.5$) will keep all (or most of) the variables and is thus not useful for variable selection. In contrast, lasso and elastic net with a penalty term close to lasso ($\alpha > 0.5$) will set some coefficients to zero and thus keep fewer variables in the model, which means that they also perform variable selection in addition to prediction. However, there is some inconsistency in terms of which variables are selected by lasso or elastic net when the covariates are measured with error (257, 258). Because some measurement error is introduced when using arrays to quantify DNA methylation levels (8, 259), this drawback of penalized regression methods may explain some of the lack of overlap in selected CpGs between different epigenetic clocks for gestational age. Moreover, the handling of correlated variables may also lead to inconsistency in variable selection, because elastic net may select variables for prediction that are not actually related to the outcome but are only correlated with variables that are associated with the outcome (252). As mentioned in chapter 2.1.2, neighboring CpGs often exhibit local correlation that may influence which CpGs are selected for different gestational age clocks.

Another potential disadvantage of penalized regression methods is that they assume a linear relationship between the variables and the outcome. Different biological processes occurring in different phases of gestation, combined with a non-linear relationship between growth rate and

gestational age (51), may give rise to a non-linear relationship between DNA methylation levels and gestational age.

Finally, interpretation of the predictive model generated by penalized regression methods is not straightforward. Due to the penalty term of the elastic net equation, the selected variables are influenced by every other variable selected (252). Thus, it is not clear whether the selected CpGs are the ones most strongly associated with gestational age (255) and/or whether they are biologically relevant.

4.6 Stability selection

4.6.1 The stability selection framework

One way of dealing with some of the drawbacks of penalized regression methods is to combine them with subsampling. As Meinshausen and Bühlmann demonstrated in 2010 (260), stability selection is a useful framework for combining high-dimensional selection algorithms with subsampling to control the number of false discoveries in the set of selected variables. This is achieved by resampling the dataset multiple times and fitting a variable selection model, in our case, lasso, to each subsample. The relative selection probabilities are then calculated for each variable (here, CpG), and the variables that have a selection probability higher than π_{thresh} – a pre-specified threshold value – are considered stably selected, i.e., stably predictive of the outcome (here, gestational age). This selection procedure controls the per-family error rate $E(V)$, which is the expected number of false positive variables. An upper bound is given by the following formula:

$$E(V) \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p} \quad (2)$$

Where q is the average number of variables (CpGs) selected by the variable selection method and p is the total number of variables included in the analyses. π_{thresh} is the selection probability threshold. The theory requires two assumptions to ensure that the error bound holds:

- (i) All noise variables should have the same selection probability.
- (ii) The variable selection method used must not be worse than random guessing.

The first assumption is difficult to test explicitly, but the distribution of the selection probabilities calculated in our main analysis in **Paper 3** indicated that most of the CpGs included in the analysis have a selection probability around zero.

Regarding the second assumption, our analyses in **Paper 2** as well as previous studies on gestational age clocks have shown that lasso regression is better than random guessing at selecting CpGs that are related to gestational age (122).

4.6.2 Determining the tuning parameters

The two main tuning parameters for the stability selection procedure is q , the number of selected variables by the variable selection method, and π_{thresh} , the threshold for stably selected variables. There is currently no consensus regarding how to define q , but it should be high enough so that, in theory, all predictive variables can be chosen (261). If q is too low, only a small subset of the predictive variables will be selected in the set of stably selected variables. One approach for determining q involves using permuted values to calculate the average number of variables selected when there is no association between the variables and the outcome (262, 263). In our study, we repeated the stability selection procedure with permuted gestational age values and calculated the average number of CpGs selected, setting this number as q .

Regarding π_{thresh} , any value between 0.5 and 1 is potentially acceptable, meaning that a variable should be selected in more than half of the fitted models to be included in the stably selected set (261). To compute π_{thresh} , Hofner *et al.* (261) recommended choosing an upper bound for the expected number of false discoveries in the stably selected set ($E(V)$), specifying q , and then including these parameters in Equation (2).

The number of subsamples is not as important, as long as it is large enough. Meinshausen and Bühlmann (260) proposed to use 100 subsamples. The size of the subsamples should always be $n/2$, which is an essential requirement for the derivation of the error bound (Equation (2)) (260).

4.7 Generalized additive models for building epigenetic clocks

One of the limitations of the regression methods discussed in chapter 4.5.5 is that they are only able to model linear relationships between the variables and the outcome. Methods that take nonlinearity into account may thus improve prediction if the underlying relationship is nonlinear. Generalized additive models (GAMs) provide a general framework for allowing nonlinear functions of each of the variables included in the model while maintaining additivity (264). This involves calculating a nonlinear function, such as a smoothing spline, for each variable, and then summing up their contributions. The additive property means that it is possible to examine the effect of each variable on the outcome individually while holding the other variables fixed (264). Furthermore, the degree of nonlinearity in the relationship between each variable and the outcome is indicated by the effective degrees of freedom estimated from the GAM (265). Thus, GAMs can be used to build a prediction model that takes nonlinear relationships between the variables and the outcome into account, and, concurrently, reveal which variables exhibit such nonlinear relationships and the degree of nonlinearity. However, GAMs do not perform variable selection, meaning that, to be able to use them for building prediction models from high-dimensional data, variable selection must be performed separately. Therefore, we combined stability selection with GAM regression to build parsimonious epigenetic gestational age

clocks as presented in **Paper 3**.

4.8 Genome annotation & enrichment

4.8.1 Annotation methods

Genome annotation is the process of identifying the location of genes and coding regions in the genome and determining their functions (266). Annotation databases also include information about noncoding RNAs, regulatory regions, DNA methylation sites, and more. An annotation search downstream of epigenetic analyses may provide useful information about the genomic locations of significant CpGs. Determining whether a CpG is in or near a specific gene, CpG island or regulatory region (and its relative location) may help unravel the biological underpinnings of the findings. In the work of this thesis, we have used three different sources/methods for genome annotation: Illumina's manifest files, the Ensembl database (267), and the Genomic Regions Enrichment of Annotation Tool (GREAT) (268).

Illumina provides manifest files, which consist of annotation data specifically for the CpGs that are targeted by their arrays. These manifest files include information on genomic location, probe type and sequence, information on nearby genes and regulatory regions, and the relative location of CpGs to genes and CpG islands. The manifest files can be downloaded from Illumina's website, and it is easy to extract information on specific CpGs of interest since the information is linked to the Illumina CpG IDs. However, even though Illumina has occasionally released updated versions of its manifest files, the information contained in those files has not necessarily been completely up to date. Furthermore, the EPIC manifest file significantly underestimates the number of enhancers and long noncoding RNAs (lncRNAs) covered by the array (269). A recent study proposed an updated annotation approach for EPIC, which greatly improved the annotation of enhancers and lncRNA transcripts (269).

Genome annotation provided by Ensembl includes detailed and comprehensive annotation of gene structures, regulatory elements, and variants for a range of vertebrate species (267). Ensembl includes both automatic annotations based on mRNA and protein sequences from publicly available databases, and manual curation of specific transcripts. The Ensembl database and associated software are frequently updated and freely available to researchers.

GREAT is an online annotation and enrichment tool that associates input genomic regions such as CpG sites with their putative target genes using annotations from numerous ontologies (268). A great advantage of GREAT is that it considers both proximal and distal regulatory elements as opposed to other methods that only take proximal regions into account. This might be especially beneficial when investigating significant CpGs on the EPIC array since EPIC covers more distal regulatory elements than previous arrays.

4.8.2 Gene-set enrichment analysis

A gene-set enrichment analysis is performed to determine whether a list of genes or CpGs is enriched for specific biological annotations or pathways. This may provide additional biologically relevant insights from the CpGs found to be associated with the phenotype of interest. There are many publicly available software to conduct this type of analysis, including GREAT (used in **Paper 1**, (268)), Web-based Gene set analysis toolkit (WebGestalt) (used in **Paper 2**, (270)), Database for Annotation, Visualization and Integrated Discovery (DAVID) (271), Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) (272), among others. Most of the available methods test for an overrepresentation of specific annotations or pathways in the list of CpGs or genes and extract information from ontologies such as Gene Ontology (273, 274). Because the CpGs that are covered by the BeadChip arrays are not a random selection from the whole genome but are highly biased in terms of location and annotation, it is important to apply background correction to avoid spurious findings. Finally, it is important to bear in mind that the annotations and pathways that are recorded in ontologies are only based on the latest knowledge available about genes, functions and pathways at the time, and important insights about the findings may therefore be missed due to this knowledge not being completely up to date.

5. Result summary

5.1 Paper 1

In **Paper 1** “*Nucleated red blood cells explain most of the association between DNA methylation and gestational age*”, we conducted an EWAS of gestational age in heterogeneous cord blood samples from a randomly selected subsample of 953 newborns in MoBa. We identified 13,660 CpGs as being significantly associated with gestational age [Bonferroni p-value ($p_{\text{Bonferroni}} < 0.05$)], of which 7,639 (56%) were present exclusively on the EPIC array. We also discovered cell-type specific signatures of gestational age in seven main cell-types in cord blood. Most of the cell-type specific gestational-age associated CpGs ($n = 2,030$, 87%) were found in nRBCs, but only a few of the CpGs ($n = 31-157$, 1.3-67%) were identified in the other cell types. Most of the nRBC-specific CpGs ($n = 1,888$, 93%) were hypomethylated with increasing gestational age.

We found similar patterns of cell-type specific DNA methylation changes with gestational age as described above when we (i) used a different method for cell-type specific analysis, and (ii) analyzed a different DNA methylation dataset from 1,062 newborns in MoBa, where DNA methylation was measured using a different array (450K). In the second subsample, we identified fewer significant cell-type specific CpGs overall ($n = 373$, $p_{\text{Bonferroni}} < 0.05$), with only 62% ($n = 231$) being specific for nRBCs. Nevertheless, there was a high level of consistency across array types; notably, 174 nRBC-specific CpGs identified in the first (EPIC-based) subsample were significantly replicated in the second (450K-based) subsample. However, there was no overlap in CpGs between the two subsamples for any of the remaining six cell types.

nRBC-specific gestational-age associated CpGs were predominantly located in gene bodies (48% of the nRBC-specific CpGs versus 30% of all CpGs on EPIC, $p = 2.5 \times 10^{-67}$) and open sea (75% versus 56%, $p = 2.2 \times 10^{-69}$). By contrast, they were depleted in promoter regions (22% versus 38%, $p = 2.8 \times 10^{-55}$) and CpG islands (4.7% versus 19%, $p = 5.3 \times 10^{-77}$).

The 2,030 nRBC-specific gestational-age associated CpGs were mapped to 2,836 genes. In addition, they were significantly enriched in four clusters of Gene Ontology (GO) biological processes: (i) response to corticosteroid (75 CpGs/55 genes, $p_{\text{Bonferroni}} = 0.0001$), (ii) response to purine-containing compound (65 CpGs/45 genes, $p_{\text{Bonferroni}} = 0.002$), (iii) granulocyte migration (34 CpGs/23 genes, $p_{\text{Bonferroni}} = 0.006$), and (iv) stress-activated protein kinase signaling cascade (58 CpGs/32 genes, $p_{\text{Bonferroni}} = 0.01$). Many of the nRBC-specific CpGs were in or near genes implicated in erythropoiesis and hemoglobin switching.

Collectively, these results point to a cell-type specific association between DNA methylation and gestational age, with nRBCs as the main cell type driving this association. Furthermore, an epigenetic signature of erythropoiesis may be partly responsible for the association.

5.2 Paper 2

In **Paper 2** “*An EPIC predictor of gestational age and its application to newborns conceived by assisted reproductive technologies*”, we developed an EPIC-specific gestational age clock – the ‘EPIC GA clock’ – by training a lasso regression model on cord-blood DNA methylation data from 755 newborns in MoBa. 176 CpGs were selected for being predictive of gestational age. The clock showed a high performance in a test set of 200 newborns in MoBa ($R^2 = 0.713$, MAD = 3.59 days) and outperformed two previous gestational age clocks (Bohlin and Knight) when tested on an independent dataset of EPIC-derived DNA methylation data from 148 newborns in the Finnish PREDO cohort (EPIC GA clock: $R^2 = 0.724$, MAD = 3.42 days; Bohlin clock: $R^2 = 0.610$, MAD = 6.69 days; Knight clock: $R^2 = 0.406$, MAD = 4.55 days).

Furthermore, we built a separate clock – the ‘450K/EPIC overlap clock’ – using the same training set but only including CpGs that were present on both arrays. When comparing the performance of this overlap clock with that of the EPIC GA clock, we found no significant difference in R^2 (-0.0001 ; 95% CI: $-0.021, 0.018$) or MAD (0.162; 95% CI: $-0.375, 0.794$) between the two clocks, indicating that the additional probes on EPIC do not add much to the prediction of gestational age.

We also developed a third clock – the ‘ETD-clock’ – using the ETDs of 674 ART-conceived newborns for training. When compared to the EPIC GA clock in the non-ART test set from START, the ETD-based clock showed a similar performance, with an R^2 difference of 0.048 (95% CI: $-0.041, 0.123$) and a difference in MAD of 0.645 (95% CI: $-0.181, 1.209$). These results indicate that using the ETD of ART-conceived newborns for training the clock does not significantly improve the prediction of gestational age compared to using ultrasound-estimated gestational age.

The EPIC GA clock showed a similar performance in ART-conceived newborns and naturally conceived newborns when tested on a dataset of 838 ART-conceived newborns from MoBa ($R^2 = 0.767$, MAD = 3.80 days, ultrasound measurements). Gestational age estimated by ETD was predicted with a similar precision (R^2 difference of 0.015 (95% CI: $-0.003, 0.033$) and accuracy (MAD difference of -0.102 (95% CI: $-0.465, 0.174$) compared to ultrasound-estimated gestational age.

Finally, we assessed the association between GAA and ART by performing a logistic regression of ART on GAA, with the latter calculated from the residuals of a regression of epigenetic gestational age (as predicted by the EPIC GA clock) on ultrasound-estimated gestational age in the 838 ART-newborns and 200 naturally conceived newborns. We found no difference in GAA between the ART and non-ART newborns ($p = 0.388$), nor when taking different ART procedures into account.

To summarize, our newly developed EPIC specific gestational age clock is highly performant in both ART-conceived and naturally conceived newborns. Using ETD for training the clock did not increase predictive performance. Moreover, ART-conceived newborns do not seem to differ from naturally

conceived newborns in terms of their epigenetic gestational age or GAA.

5.3 Paper 3

In **Paper 3** “*Stability selection enhances feature selection and enables accurate prediction of gestational age using only five DNA methylation sites*”, we combined stability selection with lasso regression to identify CpGs that are stably predictive of gestational age in an EPIC-derived DNA methylation dataset of 2,138 newborns from two different subsamples in MoBa. Of the 770,000 CpGs included in the analysis, 24 were identified as stably predictive of gestational age (i.e., they were selected more than 73% of the time). The stably selected CpGs were enriched in promoter regions ($n = 11$, 46%), which is an interesting finding given that the opposite pattern was found for nRBC-specific gestational-age associated CpGs in **Paper 1**.

Further, we explored the stability of CpGs previously selected for gestational age prediction by investigating CpGs in three previously developed clocks: (i) the EPIC GA clock we developed in **Paper 2** (also referred to as the ‘Haftorn clock’), (ii) the Bohlin clock, and (iii) the Knight clock. Eighteen (10.2%) of the Haftorn clock CpGs, eight (9.3%) of the Bohlin clock CpGs, and none of the Knight clock CpGs were found to be stably predictive of gestational age.

The stably selected CpGs were then used to build a new gestational age clock trained on 80% ($n = 1709$) of our total sample. To do this, we first reran the stability selection analysis on the training sample and ran GAM regressions of gestational age on DNA methylation levels in the training set for each of the top 15 stably selected CpGs. We then used GAM again to develop clocks that included from one to 15 of the stably selected CpGs based on the strength of their relationship with gestational age (determined by the R^2 value). These 15 clocks were subsequently used to predict gestational age in the test set ($n = 429$). We then compared their prediction performances as well as the performance of a clock that was developed on the same training set but that was based on a standard framework with lasso.

Very few CpGs were needed to attain a good prediction of gestational age. The top CpG (cg04347477) alone predicted gestational age with an R^2 of 0.52 and a median absolute deviation (MAD) of 5.09 days. When including five CpGs, we obtain an R^2 of 0.674 and a MAD of 4.4 days, a performance that is virtually indistinguishable from that of the original Bohlin clock that required 96 CpGs for prediction (122), suggesting that these five CpGs explain most of the variance related to gestational age. Furthermore, using GAM regression to create the clocks improved the prediction of gestational age, particularly in preterm newborns, suggesting that at least some of the predictive CpGs exhibit a nonlinear relationship with gestational age. Finally, many of the stably selected CpGs were mapped to genes and regulatory regions that are relevant for immune responses, metabolism, and developmental

processes, including changes in hemoglobin expression and metabolic processes that occur in the transition from pre- to postnatal life.

6. Discussion

6.1 Summary of key findings

The overarching aim of this thesis was to investigate the relationship between DNA methylation levels in newborns and their gestational age at birth, and to explore mechanisms that may explain this relationship. In **Paper 1**, we identified epigenome-wide associations between DNA methylation levels of CpGs present on the EPIC array. We found specific signatures of gestational age for all the seven main cell-types in cord blood and discovered that most of the cell-type specific CpGs associated with gestational age were confined to nRBCs. Many of the nRBC-specific CpGs were in or near genes that are implicated in erythropoiesis and hemoglobin switching.

In **Paper 2**, we developed an EPIC-specific gestational age clock that outperformed previous clocks. We also showed that the additional probes on EPIC did not improve the prediction of gestational age compared to using the probes that are also included on 450K. Furthermore, we discovered that using the embryo transfer date of ART-conceived newborns for training the clock did not significantly improve the prediction of gestational age compared to using ultrasound-based estimates. In addition, we found no differences in GAA between newborns conceived by ART and those conceived naturally.

Finally, in **Paper 3**, we introduced a methodological framework for identifying CpGs that are stably predictive of an outcome and demonstrated its utility for gestational age prediction. We discovered that the majority of CpGs selected in previous gestational age clocks, including those in the EPIC-specific clock we developed in **Paper 2**, were not stably predictive of gestational age. We also used the stably predictive CpGs to create yet another highly performant gestational age clock consisting of only five CpGs, and showed that there is a non-linear relationship between some of the CpGs and gestational age.

6.2 Implications of key findings

6.2.1 Using EPIC-derived DNA methylation data for studying gestational age

In all three papers, we used DNA methylation data quantified on the EPIC array. EPIC provides information on almost twice as many CpG sites as the preceding 450K array and has a higher coverage of intragenic and regulatory regions. Since most previous studies of DNA methylation and gestational age have used data generated on the 450K or 27K array, we sought to investigate whether the additional CpGs on EPIC could shed new light on the relationship between DNA methylation and gestational age. In **Paper 1**, we identified 2.5 times as many significantly associated CpGs compared to what Bohlin *et al.* (122) had found in a similar study using 450K data from a different subsample of MoBa. Also, slightly more of the significant CpGs we found were specific for the EPIC array (56%), although only 48% of the CpGs included in the analyses were EPIC-specific.

Another interesting finding from **Paper 1** was that we only found enrichment of biological pathways in the nRBC-specific CpGs when we included all the CpGs in the analysis, and not when restricting the analysis to only those CpGs that are present on both 450K and EPIC. Taken together, these findings suggest that more of the additional CpGs on EPIC were associated with gestational age than what would be expected just from the increased number of CpGs. Hence, assessing the additional CpGs on EPIC may be a key step forward to learn more about the biological mechanisms underlying the relationship between DNA methylation and gestational age.

A large proportion of the EPIC-specific probes target CpGs in gene bodies, intergenic, and non-CpG island regions ('open sea') (46). Interestingly, in **Paper 1**, we found an enrichment of gestational-age associated CpGs in these regions, particularly for the nRBC-specific CpGs. These observations further indicate that the EPIC-specific CpGs may be important in explaining the association between DNA methylation and gestational age. However, as we show in **Paper 2**, the EPIC-specific CpGs did not seem to improve the prediction of gestational age, indicating that the DNA methylation sites needed to create a precise gestational age clock are already present on the 450K array. Moreover, EPIC-specific CpGs were not enriched in the set of stable CpGs we identified in **Paper 3** (50% EPIC-specific CpGs). Despite the saturation of prediction performance with 450K, we did obtain a higher precision and accuracy when predicting gestational age with our EPIC-based clock compared to the previously published clocks based on 27K and 450K. This may be because some of the CpGs in these clocks are not covered on the EPIC array. Moreover, since 27K and 450K are no longer in use, most new array-based DNA methylation datasets are nowadays quantified on EPIC. Therefore, we anticipate that our EPIC-based gestational age clocks will be useful in predicting gestational age in future studies.

6.2.2 Cell-type specific DNA methylation signatures of gestational age in cord blood

Most previously published EWASes of gestational age have adjusted for cell-type proportions in the analyses, but no study had yet investigated cell-type specific associations between DNA methylation and gestational age. Shedding light on the specific cell type(s) in which those associations are present may help uncover biological mechanisms underlying the associations. In **Paper 1**, we discovered significantly associated CpGs that are specific for all the seven main cell types in cord blood, indicating that DNA methylation patterns change with gestational age in all the cell types. Some CpGs overlapped between several cell types, whereas some were unique for a specific cell type, indicating that there are some lineage-specific signals related to gestational age as well as some overall changes. Notably, we found a clear overrepresentation of gestational age-related CpGs in nRBCs, a finding that was robust across both EPIC and 450K datasets as well as when using two different cell-type specific methods for detecting significant associations. Because of this remarkable finding, we decided to focus on the nRBC-specific CpGs in the downstream analyses.

6.2.3 nRBCs in cord blood

nRBCs are immature red blood cells that have not yet extruded their nucleus. In healthy adults, red blood cells are enucleated before they enter circulation. In the fetus, however, a proportion of the circulating red blood cells still contains a nucleus (236). Although the number of nRBCs in the fetus gradually declines as gestation progresses, a proportion (typically 0-10 nRBCs per 100 white blood cells) of these cells are often still present at birth (236). The presence of nRBCs in the fetal and neonatal circulation is likely due to the high demand for red blood cells (275), but immunoregulatory functions of nRBCs may also be important in this respect (276). Either way, the nRBCs decline rapidly and disappear from the circulation during the first few days after birth (236). Therefore, our findings in **Paper 1** showing that nRBCs are the primary drivers behind the association between DNA methylation and gestational age may partly explain why gestational age-related DNA methylation changes in cord blood do not persist through childhood and adolescence (123, 125). Moreover, there is minimal overlap in CpGs between clocks for gestational age and adult age (122), and adult age clocks do not predict gestational age accurately (126), and vice versa. These differences between epigenetic clocks for adults and newborns may, in light of our discoveries, be partly explained by the absence of nRBCs in the adult bloodstream.

6.2.4 Erythropoiesis

Many of the nRBC-specific gestational-age associated CpGs found in **Paper 1** mapped to genes or regulatory regions that are related to erythropoiesis (the development of red blood cells). Red blood cells are produced in the adult and fetal bone marrow, fetal liver, and the embryonic yolk sac. They arise from hematopoietic stem cells that sequentially give rise to common myeloid progenitor, megakaryocyte-erythrocyte progenitor, burst-forming unit-erythroid (BFU-E), colony-forming unit-erythroid (CFU-E) cells, and, finally, to proerythroblasts (277). Further differentiation of proerythroblasts is commonly referred to as terminal erythropoiesis. A simple overview of this process is provided in **Figure 6**.

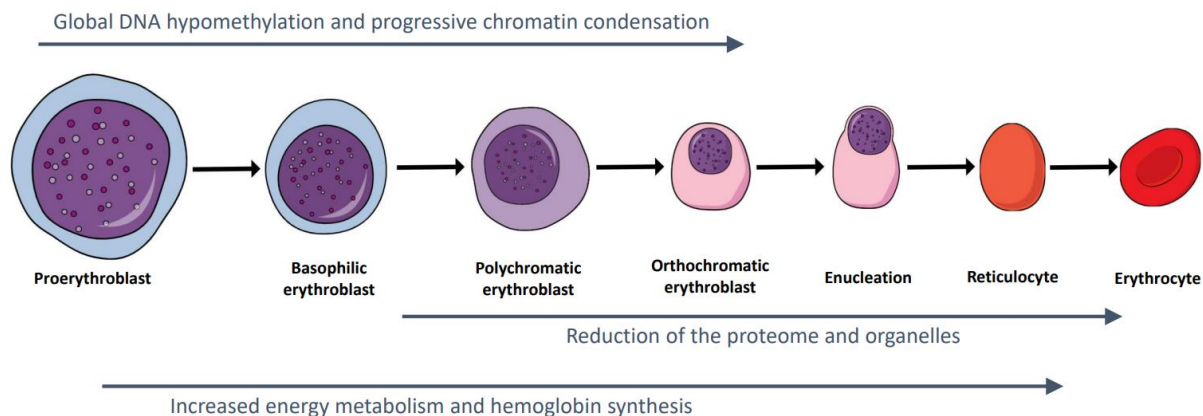


Figure 6. Simplified overview of terminal erythropoiesis. The process of differentiation from proerythroblasts into basophilic, polychromatic, and orthochromatic erythroblasts is characterized by global DNA hypomethylation and progressive chromatin condensation, increased energy metabolism and hemoglobin synthesis, and reduction of the proteome and organelles. Tightly condensed nuclei are extruded from the orthochromatic erythroblasts to generate enucleated reticulocytes in the process of enucleation. Reticulocytes further develop into mature erythrocytes. Created with [mindthegraph.com](https://www.mindthegraph.com/).

Proerythroblasts undergo mitosis to produce basophilic, polychromatic, and orthochromatic erythroblasts. This process is characterized by global DNA hypomethylation (278), progressive chromatin condensation (279), increasing energy metabolism (280), iron storage (281) and hemoglobin synthesis (282), as well as degradation of proteins and organelles (283, 284). Orthochromatic erythroblasts extrude their tightly condensed nuclei in a process called enucleation to generate reticulocytes. Finally, the reticulocytes go through extensive membrane remodeling and loss of organelles to develop into mature erythrocytes (285, 286). Terminal erythropoiesis mainly takes place in erythroblastic islands, consisting of developing erythroblasts surrounding a central macrophage that provides iron to the developing erythroblasts and removes extruded nuclei by phagocytosis (287).

In **Paper 1**, we identified several gestational-age associated nRBC-specific CpGs located in or near genes involved in each of the abovementioned processes. In the discussion below, I highlight some of these CpGs and include their CpG IDs in parentheses next to the relevant genes in which they either reside or are located close to. Examples include the DNA methyltransferase 3A gene (*DNMT3A*, cg05945668 and cg18169886) which is required for genome-wide *de novo* methylation (288), and the Tet methylcytosine dioxygenase 2 gene (*TET2*, cg22794775) which plays a key role in active DNA demethylation (19).

Previous studies by de Goede *et al.* (244, 289) demonstrated a general hypomethylated DNA methylation profile of nRBCs and, more specifically, a consistent shift of more hypomethylated sites in these cells in term newborns compared to preterm newborns. The authors hypothesized that the

DNA methylation profile of nRBCs is partially due to stochastic loss of DNA methylation, and partly due to DNA methylation changes specific for erythropoiesis. Our finding from **Paper 1**, of DNA methylation in nRBCs shifting towards more hypomethylation with increasing gestational age, supports the findings of de Goede *et al.* This shift may reflect a change in the nRBC population's maturity, thereby representing a signature of erythropoiesis. These results are also in line with those of previous studies showing a consistently higher proportion of hypomethylated CpGs amongst those associated with gestational age (119, 122, 123, 125).

Further, we identified many CpGs in or near genes involved in chromatin condensation, such as *NCOR2* (cg03406367) (290), *GATAD2B* (cg07704502) (291), *HDAC2* (cg08601899), *HDAC4* (cg10037204, cg11556929) (292), and *EED* (cg02149136) (290, 293). Chromatin condensation is also an important prerequisite for the enucleation process, and a few genes, such as *HDAC2*, are important contributors to both processes (294). Enucleation is a complex process requiring a highly orchestrated coordination between various transcription factors, cytoskeletal proteins, and other signaling molecules (295). One example is the transcription factor Forkhead box O3, encoded by *FOXO3* (cg03398073), which is essential for regulating the enucleation process and for subsequent mitochondrial clearance in reticulocytes (296). Enucleation is characterized by nuclear polarization and the formation of a contractile actin ring, in which a range of genes have been implicated, including *DIAPH3* (cg08942545) (295, 297, 298), *RAC2* (cg02860019) (299), and *CDC42EP4* (cg02234489, cg16916549, cg10305337) (300).

Other genes that are important for erythropoiesis and that are linked to CpGs that were identified in our nRBC-specific results include *PHOSPHO1* (cg21924438) which encodes Phosphoethanolamine/phosphocholine phosphatase, a key regulator of lipid metabolism during erythroid differentiation (301); *CCND3* (cg08545995) which encodes Cyclin D3, a key regulator of the number of cell divisions during terminal erythropoiesis (302), and *SOX6* (cg05275596, cg22115204) which encodes SRY-box transcription factor 6. This transcription factor stimulates erythroblast and reticulocyte maturation and ensures the long-term stability of the erythrocyte cytoskeleton (303).

During terminal erythropoiesis, the erythroblast proteome is rapidly reduced to 2-5% via bulk degradation of many cellular proteins and organelles by the ubiquitin proteasome system and the autophagy pathway (283). Many of the gestational-age associated nRBC-specific CpGs are linked to genes implicated in these processes. Examples include *HSPA12A* (cg12485738) and *HSPA13* (cg13793211, cg06118712), both of which encode 70-kDa heat shock proteins (Hsp70s). Hsp70s are molecular chaperones involved in maintaining protein homeostasis, promoting red blood cell differentiation, and ensuring cell survival throughout erythropoiesis (283, 304). Among our nRBC-specific results, we also found CpGs in or near numerous ubiquitin conjugating enzymes that are

implicated in protein degradation [*UBE2D4* (cg23624016), *UBE2E1* (cg23480273, cg18587674), *UBE2E2* (cg23480273), *UBE2E3* (cg18940067), *UBE2F* (cg23438758), *UBE2I* (cg05422590), *UBE2J2* (cg22699977), *UBE2V1* (cg16648971), *UBE3A* (cg22490346)] (305). Further, we identified CpGs linked to *BNIP3L* (cg10260596), which is implicated in mitochondrial clearance (284), and also to several genes involved in autophagy [e.g. *ATG5* (cg12216009) (284), *ATG16LI* (cg14134851) (306) and *ATG12* (cg10387289) (307)].

The reason for the degradation of proteins and organelles in developing erythroblasts is to make space for the steadily increasing levels of hemoglobin. Sufficient iron uptake and heme synthesis is crucial for appropriate hemoglobinization during terminal erythropoiesis (282). We identified gestational-age associated nRBC-specific CpGs linked to several genes encoding hemoglobin subunits [*HBB* (cg12485738, cg25660811), *HBD* (cg12485738), *HBE1* (cg12485738) and *HBG1* (cg12485738)] (308), as well as genes implicated in cellular iron uptake (*TFRC*, cg24730676, cg21259588) (309, 310) and heme biosynthesis (*FECH*, cg18365938) (282).

In summary, many of the CpGs that are associated with gestational age specifically in nRBCs are in or near genes implicated in a range of processes occurring during erythropoiesis, supporting the hypothesis that gestational-age associated DNA methylation changes represent a signature of erythropoiesis.

6.2.5 The switch from fetal to adult hemoglobin

Red blood cells are specialized cells that produce hemoglobin for oxygen transportation. Hemoglobin is a tetramer consisting of α -chain and β -chain subunits. During early embryogenesis in humans, α -globin ζ and β -globin ϵ (comprising embryonic hemoglobin) are the first globin genes to be expressed. They are subsequently replaced by α and γ genes (comprising fetal hemoglobin) along with fetal development. While α genes continue to be expressed throughout the lifetime, γ genes are subsequently replaced by the β and δ (adult hemoglobin) genes after birth. This is commonly known as the ‘fetal to adult hemoglobin switch’ (**Figure 7**, see also (302)).

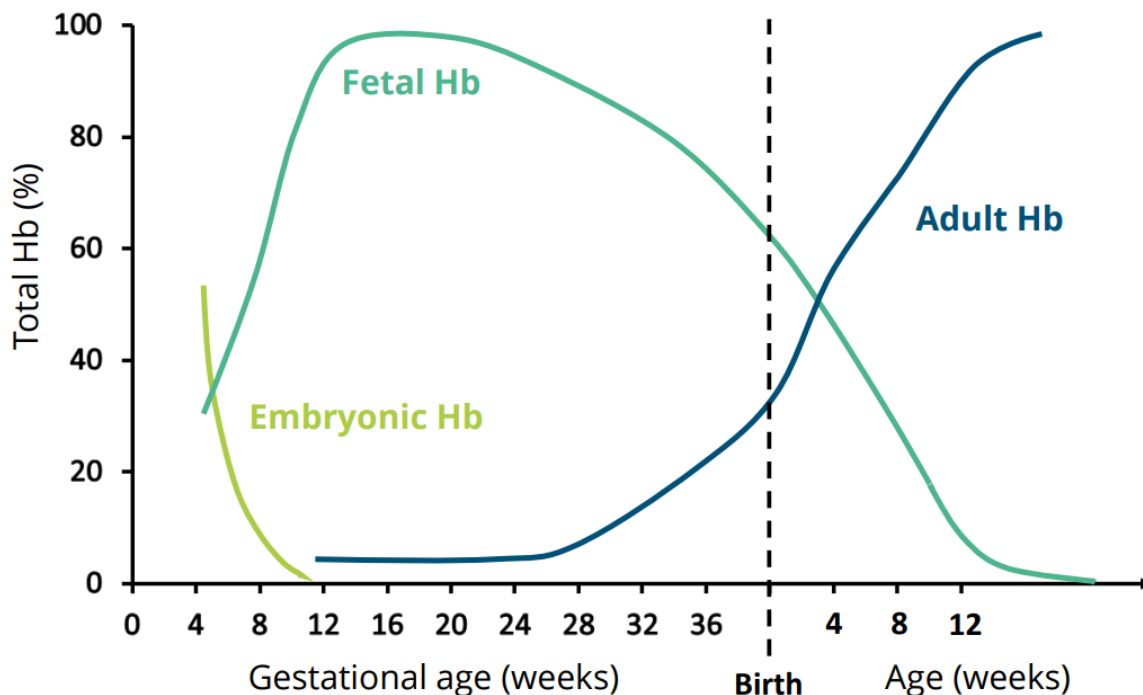


Figure 7. The ‘fetal to adult hemoglobin switch’. A schematic representation of the hemoglobin switches from embryonic to fetal hemoglobin (Hb) and from fetal to adult hemoglobin. The type of hemoglobin is displayed as the percentage of total hemoglobin synthesized throughout gestation. Adapted from Sankaran and Nathan (308).

Epigenetic mechanisms, including DNA methylation, are important for controlling the expression of human globin genes (311, 312). In addition to CpGs linked to genes encoding hemoglobin subunits, we identified several gestational-age associated nRBC-specific CpGs in or near genes known to be involved in the fetal to adult hemoglobin switch. Some examples include *KDM1A* (cg12847374) (290), *SOX6* (cg05275596, cg22115204) (313), *CDH4* (cg00037187) (290), *HDAC2* (cg08601899) (290, 302, 314), and *IGF2BP1* (cg11783901, cg12382333, cg14484274, cg22254242) (315, 316). *IGF2BP1* is particularly interesting given that a CpG in the promoter region of this gene was one of the most stably selected CpGs in our study and was also included in the ‘5 stable CpG GA clock’ we developed in **Paper 3**. However, the stably selected CpG (cg18183624) did not map to nRBCs in the cell-type specific analysis in **Paper 1**, but rather to granulocytes, indicating that mechanisms other than hemoglobin switching may underlie the predictive ability of this CpG.

6.2.6 Glucocorticoids and gestational age

The underlying reason for the fetal to adult hemoglobin switch is likely due to the difference in oxygen affinity between embryonic/fetal globins and adult globins. Embryonic/fetal globins have higher affinity for oxygen, which enables a more efficient maternal-fetal gas exchange in the placental environment (317). A sufficient supply of oxygen to the fetus is vital during the second and third

trimester of pregnancy when most of the fetal growth occurs (318). A lack of oxygen *in utero*, termed intrauterine hypoxia, is a huge prenatal stressor that can have major consequences for the mother and fetus, such as preeclampsia, fetal growth restriction, and fetal demise (319). A study on rats showed that hypoxia also led to an increase in fetal nRBC count (320).

An important control mechanism for the erythroid response to stress such as hypoxia is the glucocorticoid response (321, 322). Interestingly, the response to glucocorticoids was one of the main enriched processes of nRBC-specific gestational age-related CpGs in **Paper 1**. Indeed, one of the CpGs (cg22114534) is located near the glucocorticoid receptor gene *NR3C1*. We also identified nRBC-specific CpGs in or near other genes that are strongly implicated in the glucocorticoid receptor-dependent erythroid response to stress, including *CXCR4* (cg03733145) and *BMP4* (cg07193115) (323, 324). These are key contributors to the BMP4-dependent stress erythropoiesis pathway, in which many new erythrocytes are generated to maintain homeostasis (323, 324). There are marked similarities between stress erythropoiesis and fetal erythropoiesis, highlighting the link between fetal erythropoiesis and the response to glucocorticoids (325, 326). In addition, glucocorticoids play central roles in pregnancy and normal fetal growth and development (327, 328, 329)

6.2.7 The role of nRBCs in gestational age and fetal development

To sum up the previous subchapters, nRBCs and erythropoiesis play crucial roles in fetal growth and development, and, referring to our results in **Paper 1**, some of these mechanisms seem to be detected in DNA methylation patterns that change over the last part of gestation. Interestingly, a higher nRBC count in cord blood at birth is associated with a variety of maternal and fetal health factors, including intrauterine growth-restriction (330) and preeclampsia (331, 332). Moreover, the persistence of nRBCs after birth is predictive of newborn morbidity and mortality, as well as the risk of neonatal sepsis (333, 334). Whether or not the gestational age clocks might be picking up DNA methylation changes that are related to these outcomes remains to be elucidated.

In addition to their essential roles in oxygen and nutrient transportation, red blood cells also have immunoregulatory functions (335, 336, 337, 338). Particularly, neonatal red blood cells, including nRBCs, exert immunoregulatory functions in pregnancy, presumably contributing to the suppression of immune response to infections and the maintenance of fetomaternal tolerance (276, 339, 340, 341), which brings us to the role of immune cells in gestational age.

6.2.8 The role of leukocytes in gestational age

As mentioned in chapter 4.4.1, all nucleated cell types in cord blood, except for nRBCs, are leukocytes – cells of the immune system. Although most of the gestational-age associated CpGs identified in **Paper 1** were specific to nRBCs, we also found significant associations for all of the

other six cell types, although none of these associations were replicated in the second dataset that was based on a different array (450K). Because the nRBC-specific findings were more robust and conspicuous, we focused our downstream analyses on those CpGs. However, there are several properties of the fetal and newborn immune system that also warrant a closer investigation of the leukocyte-specific CpGs.

After spending fetal life within the protective environment of the mother's womb, the newborn's immune responses are considered immature and exhibit significant differences compared to those of adults (342, 343). Particularly, the fetal and newborn immune system is dominated by innate immunity and suppression of immune responses to maintain fetomaternal tolerance (342). There is a strong association between gestational age and the immune profile of newborns, showing particularly strong correlations with granulocyte and T-cell proportions (344). Moreover, infections and inflammation during pregnancy have been linked to preterm birth (345).

Leukocytes, including neutrophils (the largest subgroup of granulocytes), have been shown to migrate and infiltrate reproductive tissues around the time of parturition (346, 347, 348). Interestingly, one of the significantly enriched pathways in our nRBC-specific results in **Paper 1** was granulocyte migration, and, more specifically, neutrophil migration. This is particularly intriguing when considering the immunoregulatory functions of red blood cells mentioned in the previous chapter. One of the gestational-age associated nRBC-specific CpGs is located near *IL1B* (cg10486274), which encodes the proinflammatory cytokine interleukine-1 β (IL-1 β). IL-1 β has many immunoregulatory functions, such as regulating neutrophil influx and activation. IL-1 β also induces the synthesis of prostaglandin PGE2 through interaction with COX-2 (349, 350). COX-2 is an enzyme encoded by the *PTGS2* gene, which also mapped to one of the nRBC-specific CpGs associated with gestational age (cg06107544). Prostaglandins, including PGE2, are central in stimulating uterine contractions during labor (351, 352). Notably, six of the gestational-age associated nRBC-specific CpGs are annotated to *PTGER4* (cg01091117, cg03769161, cg07725759, cg08485587, cg16104076, cg26597539) which encodes a PGE2-receptor.

There are several reasons for investigating the association between granulocyte-specific changes and gestational age more closely. At term, granulocytes are the largest group of cells in cord blood and the cell type with the largest shift in proportion across gestational age (344). Moreover, the top two CpGs we identified as stably predictive of gestational age in **Paper 3**, cg04347477 and cg18183624, both mapped to granulocytes in **Paper 1**. Thus, further studies are warranted to elucidate the role of leukocyte DNA methylation changes in gestational age.

6.2.9 ART-children and epigenetic gestational age

In **Paper 2**, we sought to investigate the epigenetic gestational age in ART-conceived newborns for two main reasons. First, the known ETD for ART-newborns provides a more direct estimate of gestational age compared to both LMP and ultrasound, which may be advantageous for developing and validating gestational age clocks. Second, ART procedures coincide with the extensive epigenetic reprogramming that takes place during early embryonic life, and DNA methylation perturbations have been reported in ART newborns, suggesting that their epigenetic gestational age could differ from newborns conceived naturally.

To address the first point, we used the EPIC GA clock for predicting gestational age in ART-conceived newborns and found that the predictions were equally precise and accurate when compared to gestational age estimated from either ultrasound measurements or ETD. When using ETD estimated gestational age for training the clock, we did not observe a significantly improved prediction compared to using ultrasound-estimated gestational age. These results indicate that ultrasound estimations alone provide a good enough representation of the gestational age in ART-children, in line with a previous study showing a high agreement between ultrasound and ETD (118). When compared to ultrasound estimates, gestational age estimated from ETD has previously been referred to as the ‘actual’ gestational age of the newborn (117). However, as described in chapter 2.4.3, ETD-based estimates may also be prone to sources of error. Our results indicate that the level of measurement error is similar when using ETD to estimate gestational age compared to using ultrasound measurements, implying that both types of estimates are equally useful for training and validating gestational age clocks.

To address the second point, we compared the GAA of ART-conceived newborns to those conceived naturally and found no significant differences between the groups, suggesting that DNA methylation perturbations in ART-conceived newborns do not affect the epigenetic gestational age or GAA of these newborns. Moreover, we did not find any significant differences between different ART procedures (IVF with or without ICSI) or an effect of embryo cryopreservation on GAA, in accordance with previous studies showing no significant differences in gestational age between these groups (117, 118).

Nevertheless, it is important to bear in mind that the children in the MoBa cohort were born in the period between 1999 and 2009. There have been several major developments in ART since that time, e.g., with respect to embryo culturing and cryopreservation techniques. In particular, the length of culture time *in vitro* before transfer has changed from typically 2-3 days to 5-6 days. Thus, for the MoBa cohort, most embryos were transferred when they were in the cleavage stage, whereas nowadays most embryos are transferred at the blastocyst stage. These and other changes in procedures and/or technology could have had an impact on the DNA methylation patterns in the embryo, and

consequently rendered our results less relevant for ART children born today.

6.2.10 Utility of gestational age clocks

Epigenetic clocks, such as the ones we developed in **Paper 2** and **Paper 3**, show that it is possible to accurately predict the gestational age of the newborn from DNA methylation data. A natural follow-up question is: how is this information useful? We can divide the utility of gestational age clocks into two main categories: (i) the accurate prediction of gestational age, and (ii) the assessment of developmental maturity.

The accurate prediction of gestational age is particularly important to inform clinical decisions on proper perinatal care. Examples include if (and when) to induce labor, in addition to decisions on specialized care and interventions for babies born preterm or after pregnancy complications. Postnatal assessment of epigenetic gestational age may be useful for some of these scenarios. Future studies may reveal possibilities for assessing epigenetic gestational age also during pregnancy, by analyzing DNA methylation patterns either from circulating fetal DNA or fetal cells such as nRBCs isolated from the mother's peripheral blood (353). In addition, accurate prediction of epigenetic gestational age may be useful in forensic applications.

Epigenetic gestational age estimates may also be useful as biomarkers of developmental maturity through the assessment of GAA. As a surrogate marker for developmental maturity, GAA may provide a complement to clinical estimations of gestational age and thus aid in clinical decision-making beyond being an accurate estimate of gestational age. In addition to the applications already mentioned, GAA may be useful for tracking the efficacy of clinical interventions. In epidemiological and clinical research, GAA can be used to assess the relationship between developmental maturity and different pregnancy exposures or neonatal outcomes. However, this all depends on the appropriateness of epigenetic gestational age and GAA as markers of developmental maturity. Different clocks include different CpGs and may thus capture only part of the underlying biological mechanisms. Moreover, all gestational age clocks developed thus far have been trained only on clinically-estimated gestational age and have not considered any clinical biomarkers or phenotypic traits such as perinatal outcomes or measures of growth. Studies from the aging field have shown that the second-generation clocks, such as PhenoAge or GrimAge, can capture the biological and phenotypic part of the aging process significantly better than first-generation clocks that were only trained on chronological age (162, 163). It is conceivable that this may also be the case for the utility of gestational age clocks as markers of developmental maturity.

This type of research is still in its infancy, and there is still a long way to go before clinical use of gestational age clocks becomes a reality. Currently, the options for quantifying DNA methylation in specific CpGs are either laborious, costly, or both. Even custom arrays typically contain thousands of

probes or more and are relatively costly. Quantifying a smaller number of CpGs with qPCR or creating assays/kits using only a few probes are viable options. However, a sufficiently sensitive qPCR detection requires high specificity to be able to discriminate between cytosine and thymine bases derived from methylated and unmethylated cytosines following bisulfite conversion. This specificity depends on appropriate primer design, which, again, is influenced by the positions of CpGs. In general, a single CpG that is situated close to other CpGs in the genome, e.g., in a CpG island, is difficult to detect with qPCR. Nevertheless, the first step in enhancing the clinical utility of gestational age clocks is to identify CpGs that explain a large part of the variance in gestational age and that generate good prediction power while minimizing noise. Thus, the clocks we developed in **Paper 3** represent a significant leap forward for the future clinical utility of gestational age clocks.

6.2.11 Stability of CpGs in gestational age clocks

As there are large discrepancies regarding which CpGs are selected for prediction in different gestational age clocks, the clocks may capture different biological signals. Due to the limitations of penalized regression methods described in chapter 4.5.5, it is also conceivable that several of the CpGs selected in clocks are mere noise variables that are not truly associated with gestational age. This begs the question of how to ‘separate the wheat from the chaff’ and figure out which CpGs are most important for gestational age prediction, which are expendable, and which are most biologically relevant.

Taken together, our results from **Papers 1 and 3** show that many CpGs are weakly associated with gestational age individually, but only a few selected CpGs are able to explain a remarkably large proportion of the variance in gestational age. The most striking example is cg04347477 which had a 100% selection probability in our analysis. Alone, this CpG predicted gestational age with an R^2 of 0.52 and a MAD of 5.09 days in our test set. The stably selected CpGs were enriched in promoter regions, whereas the nRBC-specific gestational-age associated CpGs and those from the conventional EWAS were relatively depleted in promoter regions. Furthermore, whereas the results from **Paper 1** point to changes in nRBCs as the main driver behind the epigenome-wide associations between DNA methylation and gestational age, most of the stably selected CpGs identified in **Paper 3** do not map to any specific cell type. Moreover, the two most stably predictive CpGs map to granulocytes. Thus, CpGs that are important for gestational age prediction seem to have distinctive characteristics compared to the genome-wide patterns associated with gestational age.

It is difficult to establish the extent to which gestational-age associated CpGs are correlated. If several predictive CpGs are highly correlated with each other, or are part of a larger network, it is possible that only a few or even one of these CpGs are selected in a given clock. Similarly, if different CpGs from a group of correlating CpGs are selected in different subsamples during stability selection, this may lead

to low selection probabilities even though these CpGs might be important for gestational age prediction. Thus, there is a need for more studies to explore the level of correlation between gestational-age associated CpGs.

6.2.12 Biological relevance of CpGs that are stably predictive of gestational age

Many of the CpGs identified as stably predictive of gestational age in **Paper 3** are located in genes or regulatory regions that are relevant for fetal development and growth. The top stably selected CpG, cg04347477, lies in the promoter region of the *NCOR2* gene (formerly known as *SMRT*). In addition to being implicated in chromatin condensation (see chapter 6.2.4), *NCOR2* is involved in a range of biological processes related to mammalian development (354, 355), as well as in metabolic homeostasis and aging (356, 357, 358). Another stably selected CpG, cg18183624, is located in the promoter region of *IGF2BP1*, which, as mentioned in chapter 6.2.5, is involved in the fetal to adult hemoglobin switch. Furthermore, *IGF2BP1* regulates the translation of IGF2, a growth factor highly expressed *in utero* and playing an essential role in fetal and placental growth (359). Other examples of stably selected CpGs include cg20320200 and cg01833485, which were both mapped to the estrogen-related receptor gamma gene (*ESRRG*). *ESRRG* is involved in directing and maintaining the metabolic switch from a predominant dependence on carbohydrates during prenatal life to a greater reliance on oxidative metabolism after birth (360, 361). Another of the stably selected CpGs, cg21180953, is in the promoter flanking region of *SETBP1*, a gene implicated in visceral organ and brain development (362, 363). Hence, the stably selected CpGs identified in **Paper 3** are not only highly predictive of gestational age, but many of them are also linked to genes and regions that are relevant for biological processes coinciding with gestational age.

6.2.13 Linearity of the relationship between DNA methylation and gestational age

As described in chapter 4.5, the penalized regression methods typically used for developing epigenetic clocks are linear methods. Judging from the good predictive performance of the resulting gestational age clocks, the changes in DNA methylation with gestational age follows, at least partly, a linear trajectory. However, the Knight and Bohlin clocks as well as the clocks developed in **Paper 2** all *overestimate* the gestational age of preterm newborns. This is an issue that has not been adequately addressed in the epigenetic gestational age literature. We hypothesized that this could indicate that different CpGs are driving the association between DNA methylation and gestational age in different stages of pregnancy, and/or that the association between gestational age and epigenetic gestational age is not as linear throughout pregnancy.

Whereas lasso (and elastic net in general, except ridge regression) perform variable selection and prediction simultaneously, we performed these tasks separately when developing stable CpG clocks in

Paper 3. As we had already selected a small subset of CpGs with stability selection, we could use GAM regression when developing the clocks. As described in chapter 4.7, GAM includes smoothing splines and is thus able to account for nonlinearities in the relationship between DNA methylation and gestational age. Interestingly, as we showed in **Paper 3**, the clocks built using GAM did not overestimate the gestational age of preterm newborns to the same extent as clocks built using linear methods. These results suggest that at least a proportion of the predictive CpGs exhibit a nonlinear relationship with gestational age, and that this is especially important to account for when applying the clocks to preterm newborns.

6.3 What makes the gestational age clock tick?

Why is there such a strong association between DNA methylation and gestational age, and why is it possible to use DNA methylation levels to predict gestational age so precisely? Based on the results presented in this thesis, as well as those from previous studies, I propose some main hypotheses to explain the underlying mechanisms for this strong association and briefly address each of the following: (i) signatures of cell type development, (ii) preparation for birth and postnatal life, (iii) developmental maturity, (iv) aging, (v) circadian rhythms, and, finally, (vi) epigenetic drift.

The identification of gestational-age associated CpGs specific for all the seven main cell types in cord blood and the overrepresentation of nRBC-specific CpGs suggest that part of the gestational age signal represent signatures of lineage-specific development and changes in the relative maturity and developmental stages of the different cell-type populations. For nRBCs specifically, these signatures may reflect fetal erythropoiesis, but also changes in immunoregulatory activity.

DNA methylation changes in specific cell types may also indicate an essential step in the *priming*, or preparation, for birth and postnatal life. The fetal to adult hemoglobin switch is one example of a key process that takes place around birth, which may explain some of the extensive DNA methylation changes observed in nRBCs. Moreover, DNA methylation alterations that occur in the later stages of gestation may represent pivotal changes in metabolism and the immune system that help prepare for the transition from fetal to postnatal life. Examples include the migration of granulocytes and other leukocytes, and prostaglandin signaling in preparation for birth (346, 347, 348, 351, 352).

As previously hypothesized, part of the gestational age-specific DNA methylation signal may represent a measure of developmental maturity (126). Several gestational-age associated CpGs, including CpGs that are stably predictive of gestational age, are in or near genes implicated in developmental processes. Furthermore, previous studies have shown associations between positive or negative GAA and a range of maternal and perinatal exposures and traits (126, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204).

It has been hypothesized that the aging methylome reflects an innate process that is intricately linked to development and differentiation (152), and that the epigenetic aging process already starts shortly after conception (364). Thus, part of the gestational age-related DNA methylation signal may reflect a continuous aging process extending from early development throughout the entire life course. Some features that have been related to the aging methylome may also be relevant for gestational age, including the maintenance of genetic and epigenetic stability, which involve methylation and demethylation enzymes, such as DNMT3B and TET2 (152). However, there is a substantial lack of overlap in CpGs when comparing EWASes and clocks for aging and gestational age. Developmental processes *in utero* are also vastly different from those that occur in postnatal life.

Aging is also accompanied by the loss of robust circadian oscillations and the desynchronization of these processes (365). Circadian clocks maintain periodicity in internal cycles of many physiological activities in the body, such as metabolism, blood pressure, sleep, and immune responses (366, 367, 368). Furthermore, CpGs exhibit circadian behavior, which is facilitated by the rhythmic actions of DNMT and TET enzymes (365). Importantly, the fetal circadian system develops and gains autonomy toward term (369). Moreover, the circadian system is tightly connected to glucocorticoid regulation (367, 370). These characteristics of the circadian system warrant an exploration of its potential role in epigenetic gestational age.

Another established mechanism underlying aging is *epigenetic drift*. Some CpGs exhibit increased variability in DNA methylation with age. This was first shown in twin studies, where the methylomes of monozygotic twins seemed to diverge as the twins became older. These sites, therefore, capture the stochastic changes in DNA methylation that accumulate with time. On the other hand, there is a proportion of CpG sites that decrease in variance with age, with a tendency to approach fully methylated or unmethylated states (371). These types of changes in variability have not been adequately explored in the context of gestational age, owing perhaps to the lack of large datasets with enough variability in gestational age. Another measure of epigenetic drift is *entropy*, a term used to describe a state of chaos/disorder, randomness, and uncertainty. In the mammalian genome, most CpG sites are either highly methylated or unmethylated, with few sites showing intermediate levels of methylation. However, at many CpG sites, DNA methylation levels shift over time from states of high or low methylation to an intermediate fraction of ~50%, thereby increasing the level of entropy. The level of entropy is intimately linked to predictiveness. With low entropy, it is easy to predict the information stored in a given variable because of the presence of less uncertainty (372). Thus, increasing entropy cannot explain the predictiveness of epigenetic clocks. Moreover, because development is a strictly controlled process, there is likely less epigenetic drift during early development compared to later in life. However, when looking specifically at nRBCs, intermediate levels of methylation are much more common than in the other cord blood cell types (289). Increased entropy and loss of information may thus be a part of the epigenetic changes observed with the

progression of erythropoiesis and, therefore, may be part of the gestational age signal captured by DNA methylation. One hypothesis is that the majority of extensive and epigenome-wide DNA methylation changes observed in nRBCs are due to stochastic loss of DNA methylation, while the predictive CpGs are involved in more strictly controlled processes.

In summary, a range of different processes may explain the ticking of the gestational age clock and the association between DNA methylation and gestational age. It is not likely, however, that only one among these is adequate to fully explain the whole association. It is more plausible that several of them, or perhaps all of them, are intricately linked together and work in concert in various biological pathways. Moreover, other mechanisms that are yet to be identified may also be implicated in this complex interplay.

6.4 Strengths and limitations

6.4.1 Sample size and statistical power

A considerable strength of this study is the large sample size. This is particularly important for achieving enough power and maintaining a high degree of sensitivity in the analyses when studying epigenome-wide patterns and when including several interaction terms (e.g., in the cell-type specific analyses). Moreover, having a large training set is crucial to create accurate prediction models.

6.4.2 Array-based DNA methylation data

A feature that made it possible to include a large number of samples was the use of BeadChip arrays for quantification of DNA methylation, instead of more costly and laborious methods such as whole-genome bisulfite sequencing. Array-based data are also widely used in epidemiological studies of DNA methylation, which make it easier to compare our results to those of other studies and thus contribute to increasing the applicability of the epigenetic clocks we have developed in this study. By using the EPIC array for quantifying DNA methylation in our main datasets, we were able to investigate DNA methylation patterns that had not previously been scrutinized for gestational age due to the use of older and less comprehensive arrays in previous studies. Furthermore, the random allocation of cases (ART-conceived newborns) and controls (naturally-conceived newborns) on array plates in the START dataset minimized potential bias between the groups.

However, the CpGs targeted by these arrays were selected based on prior knowledge and are therefore biased in terms of coverage. Moreover, the coverage of CpG sites is marginal compared to the whole methylome. Even the EPIC array covers only approximately 3% of all CpG sites in the epigenome (estimated at ~29 million). This lack of coverage is especially pertinent to distal regulatory elements (46), although newer annotation methods have revealed a higher coverage of enhancer elements and lncRNAs than previously assumed (269). Furthermore, the arrays were designed without consideration

of interindividual variation in DNA methylation, and generally target CpGs that do not exhibit such variation (32). When developing epigenetic clocks with the goal of making them as precise as possible in all individuals, this lack of interindividual variation may be an advantage. However, it can be a limitation in other settings e.g., when studying differences between groups in an EWAS, or when treating the epigenetic clocks as biomarkers, such as in studies of GAA. Another limitation of the arrays is that, due to the bisulfite conversion step, they are unable to distinguish between 5-mC and other cytosine modifications such as 5-hmC (373).

6.4.3 Tissue specificity

As previously mentioned, DNA methylation is highly tissue-specific, which means that the choice of tissue to be analyzed may have a large impact on the results and influence their interpretation. Cord blood is a widely used tissue for studying the associations between DNA methylation and neonatal traits such as gestational age. It is readily available, is non-invasive, and is thus convenient for use in large cohort studies. However, it is not clear how performant it is as a surrogate for other types of neonatal tissues, especially in the context of gestational age. Furthermore, cord blood samples can be contaminated by maternal blood during sample collection, which may negatively impact epigenetic gestational age estimations (374). We did not assess the levels of maternal blood contamination in cord blood samples used in this study.

Alternative tissues to cord blood include dried peripheral blood spots (375), buccal swabs (376), and placenta (377). Dried blood spots have been used in some studies of epigenetic gestational age, including the development of the Knight clock (126). However, such blood spots are commonly collected several days after birth, and the timing may vary from 24 hours to more than 5 days post-delivery. It is conceivable that this variation in sampling may negatively affect epigenetic gestational age analyses, especially when considering the role of nRBCs and their rapid decline within the first few days after birth. Placental samples have also been used in studies of gestational age, and a couple of placenta-specific clocks for gestational age have been developed (124, 188). A recent study suggested that epigenetic gestational age deviations do not correspond well between placenta and cord blood, highlighting the tissue-specific aspect of epigenetic analyses of gestational age (195).

Another potential limitation of cord blood samples is that they represent a heterogeneous mixture of different cell types, which can complicate the interpretation of results and in pinpointing biological mechanisms underlying the observed associations. The use of statistical methods to deconvolute the cell-type specific signals may partly alleviate this problem, but it would be far better to analyze the different cell types separately.

6.4.4 Using reference data for inferring cell type proportions

Since individual cell-type counts were not available for our samples, it was advantageous to have cord blood specific references for calculating cell-type proportions. There are several physiological differences between cord blood and adult peripheral blood, such as the presence of nRBCs. This makes references constructed for adult peripheral blood suboptimal for newborn cord blood (236, 242).

However, reference-based methods are not without caveats. The cord blood references that have been published thus far only account for seven main cell types, many of which have different subtypes. A higher resolution of cell-type deconvolution could provide a more detailed and complete picture of how cell-type differences impact DNA methylation patterns (378). A recently published enhanced reference library of adult peripheral blood included 12 leukocyte subtypes (379), which could inspire the construction of a reference with a similar resolution in cord blood. However, smaller subsets of cell types would also require larger datasets to obtain enough power in the analyses.

There are also several biological and technical differences between the datasets used for generating the cord-blood references. In addition, each of the datasets consists of relatively few samples (between 4-11 samples, depending on the cell-type and dataset), with varying purity and separation of cell types (242). Moreover, there is little to no information on the gestational age range in these datasets, and it is not known whether differences in gestational age would have an impact on these references. However, large differences in DNA methylation, like those that are typical for cell-type-specific differentially methylated CpGs, are unlikely to be the result of confounding factors such as age, sex or genetics, that are typically associated with relatively smaller shifts in DNA methylation (380).

Notably, due to the relative immaturity of the adaptive immune system in newborns, DNA methylation patterns in cord-blood derived CD4+ and CD8+ T cells are substantially more similar to each other than those from adult blood (242, 343). These two cell types are therefore more difficult to differentiate from each other in cord blood samples. Furthermore, interactions between nRBCs and other cell types during FACS may induce significant cross-contamination of cell populations if not properly accounted for during cell sorting (381). Additionally, nRBCs were only included in two of the four datasets used in the combined cord-blood reference (242). The inferred proportions of nRBCs in our datasets showed more intraindividual variation and were generally higher than expected based on previously reported normal newborn values (236), which indicate that part of the cell-type proportion that is assigned to nRBCs may actually belong to other cell-types.

The accuracy of cell-type deconvolution may greatly influence downstream analyses, especially cell-type specific analyses such as CellDMC and TCA. At the time of writing, a method to assess the validity of cell-type proportion estimates for datasets where the true cellular proportions are unknown has been published (382). Such a method may be particularly useful for assessing the accuracy of

cellular deconvolution before using the proportion estimates in future cell-type specific DNA methylation analyses.

6.4.5 Reliability of cell-type specific DNA methylation analysis methods

A considerable strength of our study is the use of cell-type specific methods to learn more about how the different cell types impact the association between DNA methylation and gestational age in cord blood. As mentioned in the previous chapter, the results from these cell-type specific methods are largely dependent on the accuracy of cell-type deconvolution. Furthermore, the power of the cell-type specific analyses depends on several different characteristics of the data. These characteristics include the number of cell types in each sample, the distribution of relative cell-type proportions, whether DNA methylation changes occur in one or more cell types, and, if the latter is the case, the direction of these changes (i.e., if a CpG is hypermethylated in one cell type but hypomethylated in another, or if the direction is the same in both/all cell types) (248). We included seven main cell types in our study, which is more than what was used in most of the validation scenarios provided by the CellDMC and TCA authors (three or six cell types, see (248, 249)). Moreover, some of the cell types had relatively low estimated proportions in our datasets (e.g., a median of 5.1% NK cells, 5.6% B-cells, and 7.4% monocytes in the START dataset). This may result in less robust results, especially for the cell types that have lower proportions, although the large sample size in our analyses may mitigate this issue. One option to further increase robustness of the results would be to reduce the number of cell types by combining some of them based on their lineage, e.g., one may merge the B-cells, T-cells, and NK cells into a common category of lymphoid cells, and the granulocytes, monocytes, and nRBCs into a common category of myeloid cells. However, this could limit the resolution and interpretability of the cell-type specific analyses.

The compositional feature of cell-type proportions leads to some degree of correlation between the cell-type proportions. This might, in turn, lead to mapping differentially methylated sites to the wrong cell type (249). However, the use of a marginal conditional approach (i.e., a model is first fitted to all cell types jointly and the effect of each cell type is then tested separately for its statistical significance) may help alleviate this problem because the effect of other cell types is then accounted for.

In our implementation of cell-type specific methods, assuming that GA affects DNA methylation and not vice versa, the approaches of CellDMC and (one-stage) TCA are very similar. However, whereas CellDMC assumes a fixed effect of the phenotype on cell-type variation, the two-stage implementation of TCA is, in theory, able to take interindividual variation of cell-type specific methylation into account. However, this version of TCA has not yet been externally validated, and although the differences between CellDMC and TCA have been discussed thoroughly in several preprints (383, 384, 385), no objective comparison of these methods has thus far been published. Because CellDMC

has been more thoroughly validated in several datasets and EWASes, it was chosen as the main method in **Paper 1**.

6.4.6 Phenotypic information

The MoBa study provides a wealth of phenotyped data as well as linkage opportunities to several national registries, such as the MBRN. This combination provides critically important phenotypic information. In particular, the mandatory reporting of ART-procedures in MBRN and the specific details regarding the procedures used to achieve pregnancy were pivotal for (i) identifying ART-conceived newborns, (ii) conducting the ETD-specific analyses of epigenetic gestational age, and (iii) discriminating between different ART procedures.

Ethnicity may be a confounding factor of the association between gestational age and DNA methylation, as it is associated with both (386, 387). However, due to the lack of information regarding ethnicity in MoBa questionnaires and MBRN, we were unable to address ethnic-specific differences in our analyses. According to a recent preprint documenting genotype data from MoBa, approximately 95% of the ~235,000 genotyped individuals were identified as having European ancestry (388). Therefore, it is unlikely that differences in ethnicity have confounded our analyses. Homogeneity in terms of ethnicity may, however, limit the relevance of our results to other ethnic groups especially in relation to the performance and transferability of the epigenetic gestational clocks we have developed.

6.4.7 Range of gestational age

The gestational age range of the samples we used in our analyses is relatively narrow, with a clear overrepresentation of term newborns and an underrepresentation of preterm newborns. As a result, conclusions drawn from these studies may not necessarily apply to newborns born preterm. To investigate this further, it would be necessary to have a dataset that includes more preterm babies, although this might in turn lead to other methodological pitfalls and biases in prediction due to clinical differences between preterm and term newborns (187). However, as discussed in chapter 6.2.13, using nonlinear methods for building the prediction models may help resolve some of these issues.

7. Conclusions and future perspectives

Overall, our findings contribute to an increased understanding of the relationship between DNA methylation and gestational age. Plausible underlying mechanisms of this association include signatures of cell-type development, in particular erythropoiesis, preparation for birth and postnatal life, developmental maturity, maintenance of genetic and epigenetic stability, circadian oscillations, and stochastic changes in DNA methylation that accumulate with time. Furthermore, we have developed several accurate epigenetic gestational age clocks that may prove useful in future clinical applications and will be valuable tools for further research on gestational age and developmental maturity. These findings contribute to the formation of new research questions and further studies that should be pursued. Below, I outline some prospects for future research in this field.

First, the cell-type specific relationship between DNA methylation and gestational age should be further explored. Increasing our understanding of gestational-age associated DNA methylation changes in specific cell types as well as those that are common to several or all cell types may be valuable to gain further mechanistic insights into the epigenetic regulation of fetal growth and development. Furthermore, since our study was restricted to cord blood, it would be intriguing to investigate the stability of predictive CpGs and cell-type specific patterns in other neonatal tissues, such as placenta or buccal cells.

To learn more about the biological mechanisms underlying the association between DNA methylation and gestational age, it would also be highly relevant to study the impact of epigenetic drift on gestational age in terms of changes in variability and entropy. Furthermore, little is known about the correlation of DNA methylation patterns related to gestational age. Although some clusters of co-methylated CpGs have been associated with adult aging (389, 390), such studies on gestational age are currently lacking. It would be of great interest to identify co-methylation networks specifically for the different cell types in cord blood and other tissues. Additionally, investigating other epigenetic marks such as histone modifications is important to obtain a more complete picture of how the epigenetic machinery is associated with gestational age. The role of 5hmC in several cellular and developmental processes has recently been highlighted (391, 392), and further studies should explore the potential role of this modification in the context of gestational age. Finally, combining insights gained from epigenetic studies of gestational age with other ‘-omics’ data types, such as genomics, transcriptomics and proteomics, would be valuable in illuminating the functional roles of DNA methylation in gestational age and fetal development.

The nonlinear relationship between certain CpGs and gestational age that was demonstrated in Paper 3 should be validated and further explored, for example, by applying deep learning methods as has been demonstrated for adult epigenetic age (167). To fully understand the changes in DNA methylation throughout pregnancy, future studies should include a larger range of gestational age than was possible

in the work of this thesis. There is, for example, a pressing need to include more preterm newborns in gestational age studies. When studying preterm newborns, however, it is important to be cautious of pathology and other factors correlating with preterm birth that could potentially confound the analyses.

Epidemiological studies will continue to be important to investigate the link between GAA and different exposures and outcomes of interest. However, to draw valid conclusions from such studies, it is pivotal to discern the relevance of the GAA measure. In addition to learning more about the biological mechanisms being tagged by the gestational age-related CpGs, it would be of interest to develop second-generation epigenetic clocks for gestational age, taking phenotypic markers of perinatal development into account.

Finally, investigating the feasibility of using fetal cell-free DNA or fetal cells, such as nRBCs extracted from maternal blood for investigating the relationship between DNA methylation and gestational age during pregnancy, offers great promise for future translational applications, as it could open for longitudinal studies of DNA methylation changes throughout pregnancy.

Appendix

Relevant papers that are not part of this thesis:

Lee, Y., **Haftorn, K.L.**, Denault, W.R.P., Nustad, H.E., Page, C.M, Lyle, R., Lee-Ødegård, S., Moen, G.-H., Prasad, R.B., Groop, L.C., Sletner, L., Sommer, C., Magnus, M.C., Gjessing, H.K., Harris, J.R., Magnus, P., Håberg, S.E., Jugessur, A. and Bohlin, J. **Blood-based epigenetic estimators of chronological age in human adults using DNA methylation data from the Illumina MethylationEPIC array.** *BMC Genomics* 21, 747 (2020). <https://doi.org/10.1186/s12864-020-07168-8>

Håberg, S.E., Page, C.M., Lee, Y., Nustad, H.E., Magnus, M.C, **Haftorn, K.L.**, Carlsen, E.Ø., Denault, W.R.P., Bohlin, J., Jugessur, A., Magnus, P., Gjessing, H.K., Lyle, R. **DNA methylation in newborns conceived by assisted reproductive technology.** *Nat Commun* 13, 1896 (2022). <https://doi.org/10.1038/s41467-022-29540-w>

Romanowska J., Nustad H.E., Page C.M., Denault W.R.P., Bohlin J., Lee Y., Magnus M.C., **Haftorn K.L.**, Gjerdevik M., Novakovic B., Saffery R., Gjessing H.K., Lyle R., Magnus P., Håberg S.E., and Jugessur A. **The X-factor in ART: does the use of Assisted Reproductive Technologies influence DNA methylation on the X chromosome?** *bioRxiv* 2022.2010.2006.510603 (2022). <https://www.biorxiv.org/content/biorxiv/early/2022/10/07/2022.10.06.510603.full.pdf>. Accepted for publication in *Human Genomics* (2023).

Sammallahti, S., Koopman-Verhoeff, M.E., Binter, AC., Mulder, R.H., Cabré-Riera, A., Kvist, T., Malmberg, A.L.K., Pesce, G., Plancoulaine, S., Heiss, J.A., Rifas-Shiman, S.L., Röder, S.W., Starling, A.P., Wilson R., Guerlich K., **Haftorn, K.L.**, Page, C.M., Luik, A.I., Tiemeier H., Felix, J.F., Raikkonen, K., Lahti, J., Relton, C.L., Sharp, G.C., Waldenberger, M., Grote, V., Heude, B., Annesi-Maesano, I., Hivert, M., Zenclussen, A.C., Herberth, G., Dabelea, D., Grazuleviciene, R., Vafeiadi, M., Håberg, S.E., London, S.J., Guxens, M., Richmond R.C., Cecil C.A.M. **Longitudinal associations of DNA methylation and sleep in children: a meta-analysis.** *Clin Epigenet* 14, 83 (2022). <https://doi.org/10.1186/s13148-022-01298-4>

References

1. Stricker SH, Köferle A, Beck S. From profiles to function in epigenomics. *Nature Reviews Genetics*. 2017;18(1):51-66.
2. Tiedemann RL, Liang G, Jones PA. The Human Epigenome. In: Michels KB, editor. *Epigenetic Epidemiology*. Cham: Springer International Publishing; 2022. p. 3-25.
3. Deans C, Maggert KA. What do you mean, "epigenetic"? *Genetics*. 2015;199(4):887-96.
4. Holliday R. Epigenetics: a historical overview. *Epigenetics*. 2006;1(2):76-80.
5. Yousefi PD, Suderman M, Langdon R, Whitehurst O, Davey Smith G, Relton CL. DNA methylation-based predictors of health: applications and statistical considerations. *Nature reviews Genetics*. 2022;23(6):369-83.
6. Martin EM, Fry RC. Environmental Influences on the Epigenome: Exposure- Associated DNA Methylation in Human Populations. *Annu Rev Public Health*. 2018;39:309-33.
7. Pogribny IP, Rusyn I. Environmental toxicants, epigenetics, and cancer. *Adv Exp Med Biol*. 2013;754:215-32.
8. Rakan V, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nature reviews Genetics*. 2011;12(8):529-41.
9. Li E, Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol*. 2014;6(5):a019133.
10. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*. 2012;13(7):484-92.
11. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*. 2009;41(2):178-86.
12. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-95.
13. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*. 2011;6(6):692-702.
14. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*. 2010;328(5980):916-9.
15. Zemach A, Zilberman D. Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Curr Biol*. 2010;20(17):R780-5.
16. Raddatz G, Guzzardo PM, Olova N, Fantappiè MR, Rampp M, Schaefer M, et al. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc Natl Acad Sci U S A*. 2013;110(21):8627-31.
17. Yi SV, Goodisman MAD. The impact of epigenetic information on genome evolution. *Philos Trans R Soc Lond B Biol Sci*. 2021;376(1826):20200114.
18. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*. 1999;99(3):247-57.
19. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*. 2013;502(7472):472-9.
20. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*. 2009;324(5929):930-5.
21. Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*. 2010;466(7310):1129-33.
22. Hashimoto H, Liu Y, Upadhyay AK, Chang Y, Howerton SB, Vertino PM, et al. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res*. 2012;40(11):4841-9.
23. Otani J, Kimura H, Sharif J, Endo TA, Mishima Y, Kawakami T, et al. Cell cycle-dependent turnover of 5-hydroxymethyl cytosine in mouse embryonic stem cells. *PloS one*. 2013;8(12):e82961.

24. He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*. 2011;333(6047):1303-7.
25. Maiti A, Drohat AC. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem*. 2011;286(41):35334-8.
26. Weber AR, Krawczyk C, Robertson AB, Kuśnierczyk A, Vågbø CB, Schuermann D, et al. Biochemical reconstitution of TET1-TDG-BER-dependent active DNA demethylation reveals a highly coordinated mechanism. *Nature communications*. 2016;7:10806.
27. Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Res*. 2016;44(11):5123-32.
28. Haerter JO, Lövkvist C, Dodd IB, Sneppen K. Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. *Nucleic Acids Res*. 2014;42(4):2235-44.
29. Nustad HE, Steinsland I, Ollikainen M, Cazaly E, Kaprio J, Benjamini Y, et al. Modeling dependency structures in 450k DNA methylation data. *Bioinformatics (Oxford, England)*. 2022;38(4):885-91.
30. Li G, Liu Y, Zhang Y, Kubo N, Yu M, Fang R, et al. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat Methods*. 2019;16(10):991-3.
31. Shayevitch R, Askayo D, Keydar I, Ast G. The importance of DNA methylation of exons on alternative splicing. *Rna*. 2018;24(10):1351-62.
32. Gunasekara CJ, MacKay H, Scott CA, Li S, Laritsky E, Baker MS, et al. Systemic interindividual epigenetic variation in humans is associated with transposable elements and under strong genetic control. *Genome Biol*. 2023;24(1):2.
33. de Mendoza A, Nguyen TV, Ford E, Poppe D, Buckberry S, Pflueger J, et al. Large-scale manipulation of promoter DNA methylation reveals context-specific transcriptional responses and stability. *Genome Biology*. 2022;23(1):163.
34. Borgel J, Guibert S, Li Y, Chiba H, Schübeler D, Sasaki H, et al. Targets and dynamics of promoter DNA methylation during early mouse development. *Nature genetics*. 2010;42(12):1093-100.
35. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. *Nature*. 1993;366(6453):362-5.
36. Bartolomei MS, Webber AL, Brunkow ME, Tilghman SM. Epigenetic mechanisms underlying the imprinting of the mouse H19 gene. *Genes & development*. 1993;7(9):1663-73.
37. Mohandas T, Sparkes RS, Shapiro LJ. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science*. 1981;211(4480):393-6.
38. Suelves M, Carrió E, Núñez-Álvarez Y, Peinado MA. DNA methylation dynamics in cellular commitment and differentiation. *Brief Funct Genomics*. 2016;15(6):443-53.
39. Colaneri A, Wang T, Pagadala V, Kittur J, Staffa NG, Jr., Peddada SD, et al. A minimal set of tissue-specific hypomethylated CpGs constitute epigenetic signatures of developmental programming. *PloS one*. 2013;8(9):e72670.
40. Song F, Mahmood S, Ghosh S, Liang P, Smiraglia DJ, Nagase H, et al. Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development. *Genomics*. 2009;93(2):130-9.
41. Kulis M, Queirós AC, Beekman R, Martín-Subero JI. Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochim Biophys Acta*. 2013;1829(11):1161-74.
42. Nator KL, Altun G, Lynch C, Tran H, Harness JV, Slavin I, et al. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell*. 2012;10(5):620-34.
43. Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. *Epigenetics Chromatin*. 2016;9(1):26.
44. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, et al. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res*. 2006;16(3):383-93.
45. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, et al. Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics*. 2009;1(1):177-200.

46. 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 Biol.* 2016;17(1):208.
47. Noguera-Castells A, García-Prieto CA, Álvarez-Errico D, Esteller M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics.* 2023;18(1):2185742.
48. Zhou W, Hinoue T, Barnes B, Mitchell O, Iqbal W, Lee SM, et al. DNA methylation dynamics and dysregulation delineated by high-throughput profiling in the mouse. *Cell Genom.* 2022;2(7).
49. Wilcox AJ. *Fertility and Pregnancy : an epidemiologic perspective.* New York, NY: Oxford University Press; 2010.
50. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nature reviews Genetics.* 2013;14(8):585-94.
51. Schoenwolf GC, Larsen WJ. *Larsen's human embryology.* 4th ed. Philadelphia, PA: Churchill Livingstone; 2009.
52. Stiles J. Brain development and the nature versus nurture debate. *Prog Brain Res.* 2011;189:3-22.
53. Jena A, Montoya CA, Mullaney JA, Dilger RN, Young W, McNabb WC, et al. Gut-Brain Axis in the Early Postnatal Years of Life: A Developmental Perspective. *Front Integr Neurosci.* 2020;14:44.
54. Nauta AJ, Ben Amor K, Knol J, Garssen J, van der Beek E. Relevance of pre- and postnatal nutrition to development and interplay between the microbiota and metabolic and immune systems. *The American Journal of Clinical Nutrition.* 2013;98(2):586S-93S.
55. Rakoff-Nahoum S, Kong Y, Kleinstein SH, Subramanian S, Ahern PP, Gordon JI, et al. Analysis of gene–environment interactions in postnatal development of the mammalian intestine. *Proceedings of the National Academy of Sciences.* 2015;112(7):1929-36.
56. Abreu AP, Kaiser UB. Pubertal development and regulation. *Lancet Diabetes Endocrinol.* 2016;4(3):254-64.
57. Arain M, Haque M, Johal L, Mathur P, Nel W, Rais A, et al. Maturation of the adolescent brain. *Neuropsychiatr Dis Treat.* 2013;9:449-61.
58. Somerville LH. Searching for Signatures of Brain Maturity: What Are We Searching For? *Neuron.* 2016;92(6):1164-7.
59. Girgis F, Lee DJ, Goodarzi A, Ditterich J. Toward a Neuroscience of Adult Cognitive Developmental Theory. *Front Neurosci.* 2018;12:4.
60. Blagosklonny MV, Hall MN. Growth and aging: a common molecular mechanism. *Aging.* 2009;1(4):357-62.
61. Walker RF. A Mechanistic Theory of Development-Aging Continuity in Humans and Other Mammals. *Cells.* 2022;11(5).
62. de Magalhães JP, Sandberg A. Cognitive aging as an extension of brain development: A model linking learning, brain plasticity, and neurodegeneration. *Mechanisms of Ageing and Development.* 2005;126(10):1026-33.
63. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol.* 2019;20(10):590-607.
64. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell.* 1992;69(6):915-26.
65. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nature reviews Genetics.* 2013;14(3):204-20.
66. Monk M, Boubelik M, Lehnert S. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. *Development.* 1987;99(3):371-82.
67. Sanford JP, Clark HJ, Chapman VM, Rossant J. Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse. *Genes & development.* 1987;1(10):1039-46.
68. Wang L, Zhang J, Duan J, Gao X, Zhu W, Lu X, et al. Programming and inheritance of parental DNA methylomes in mammals. *Cell.* 2014;157(4):979-91.

69. Zhu P, Guo H, Ren Y, Hou Y, Dong J, Li R, et al. Single-cell DNA methylome sequencing of human preimplantation embryos. *Nature genetics*. 2018;50(1):12-9.
70. Guo H, Zhu P, Yan L, Li R, Hu B, Lian Y, et al. The DNA methylation landscape of human early embryos. *Nature*. 2014;511(7511):606-10.
71. Proudhon C, Duffié R, Ajjan S, Cowley M, Iranzo J, Carbajosa G, et al. Protection against de novo methylation is instrumental in maintaining parent-of-origin methylation inherited from the gametes. *Mol Cell*. 2012;47(6):909-20.
72. Greenberg MV, Glaser J, Borsos M, Marjou FE, Walter M, Teissandier A, et al. Transient transcription in the early embryo sets an epigenetic state that programs postnatal growth. *Nature genetics*. 2017;49(1):110-8.
73. Glaser J, Iranzo J, Borensztein M, Marinucci M, Gualtieri A, Jouhanneau C, et al. The imprinted *Zdbf2* gene finely tunes control of feeding and growth in neonates. *Elife*. 2022;11.
74. Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, et al. The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol Cell*. 2012;48(6):849-62.
75. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res*. 2013;23(3):555-67.
76. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*. 2015;523(7559):212-6.
77. He Y, Hariharan M, Gorkin DU, Dickel DE, Luo C, Castanon RG, et al. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. *Nature*. 2020;583(7818):752-9.
78. Sliker RC, Roost MS, van Iperen L, Suchiman HE, Tobi EW, Carlotti F, et al. DNA Methylation Landscapes of Human Fetal Development. *PLoS Genet*. 2015;11(10):e1005583.
79. Bianco-Miotto T, Mayne BT, Buckberry S, Breen J, Rodriguez Lopez CM, Roberts CT. Recent progress towards understanding the role of DNA methylation in human placental development. *Reproduction*. 2016;152(1):R23-30.
80. Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res*. 1982;10(8):2709-21.
81. Fuke C, Shimabukuro M, Petronis A, Sugimoto J, Oda T, Miura K, et al. Age related changes in 5-methylcytosine content in human peripheral leukocytes and placentas: an HPLC-based study. *Ann Hum Genet*. 2004;68(Pt 3):196-204.
82. Novakovic B, Yuen RK, Gordon L, Penaherrera MS, Sharkey A, Moffett A, et al. Evidence for widespread changes in promoter methylation profile in human placenta in response to increasing gestational age and environmental/stochastic factors. *BMC genomics*. 2011;12:529.
83. 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;7(6):652-63.
84. Banister CE, Koestler DC, Maccani MA, Padbury JF, Houseman EA, Marsit CJ. Infant growth restriction is associated with distinct patterns of DNA methylation in human placentas. *Epigenetics*. 2011;6(7):920-7.
85. Filiberto AC, Maccani MA, Koestler D, Wilhelm-Benartzi C, Avissar-Whiting M, Banister CE, et al. Birthweight is associated with DNA promoter methylation of the glucocorticoid receptor in human placenta. *Epigenetics*. 2011;6(5):566-72.
86. Wilhelm-Benartzi CS, Houseman EA, Maccani MA, Poage GM, Koestler DC, Langevin SM, et al. In utero exposures, infant growth, and DNA methylation of repetitive elements and developmentally related genes in human placenta. *Environ Health Perspect*. 2012;120(2):296-302.
87. Constância M, Hemberger M, Hughes J, Dean W, Ferguson-Smith A, Fundele R, et al. Placental-specific IGF-II is a major modulator of placental and fetal growth. *Nature*. 2002;417(6892):945-8.
88. Frank D, Fortino W, Clark L, Musalo R, Wang W, Saxena A, et al. Placental overgrowth in mice lacking the imprinted gene *Ipl*. *Proc Natl Acad Sci U S A*. 2002;99(11):7490-5.
89. Goisis A, Håberg SE, Hanevik HI, Magnus MC, Kravdal Ø. The demographics of assisted reproductive technology births in a Nordic country. *Human reproduction (Oxford, England)*. 2020;35(6):1441-50.

90. Steptoe PC, Edwards RG. Birth after the reimplantation of a human embryo. *Lancet*. 1978;2(8085):366.
91. Calhaz-Jorge C, De Geyter CH, Kupka MS, Wyns C, Mocanu E, Motrenko T, et al. Survey on ART and IUI: legislation, regulation, funding and registries in European countries: The European IVF-monitoring Consortium (EIM) for the European Society of Human Reproduction and Embryology (ESHRE). *Hum Reprod Open*. 2020;2020(1):hoz044.
92. Mani S, Ghosh J, Coutifaris C, Sapienza C, Mainigi M. Epigenetic changes and assisted reproductive technologies. *Epigenetics*. 2020;15(1-2):12-25.
93. Håberg SE, Page CM, Lee Y, Nustad HE, Magnus MC, Haftorn KL, et al. DNA methylation in newborns conceived by assisted reproductive technology. *Nature communications*. 2022;13(1):1896.
94. Melamed N, Choufani S, Wilkins-Haug LE, Koren G, Weksberg R. Comparison of genome-wide and gene-specific DNA methylation between ART and naturally conceived pregnancies. *Epigenetics*. 2015;10(6):474-83.
95. Fauque P, De Mouzon J, Devaux A, Epelboin S, Gervoise-Boyer MJ, Levy R, et al. Reproductive technologies, female infertility, and the risk of imprinting-related disorders. *Clin Epigenetics*. 2020;12(1):191.
96. Lazaraviciute G, Kauser M, Bhattacharya S, Haggarty P, Bhattacharya S. A systematic review and meta-analysis of DNA methylation levels and imprinting disorders in children conceived by IVF/ICSI compared with children conceived spontaneously. *Human reproduction update*. 2014;20(6):840-52.
97. Berntsen S, Söderström-Anttila V, Wennerholm UB, Laivuori H, Loft A, Oldereid NB, et al. The health of children conceived by ART: 'the chicken or the egg?'. *Human reproduction update*. 2019;25(2):137-58.
98. Henningsen AA, Gissler M, Skjaerven R, Bergh C, Tiitinen A, Romundstad LB, et al. Trends in perinatal health after assisted reproduction: a Nordic study from the CoNARTaS group. *Human reproduction (Oxford, England)*. 2015;30(3):710-6.
99. McDonald SD, Han Z, Mulla S, Murphy KE, Beyene J, Ohlsson A. Preterm birth and low birth weight among in vitro fertilization singletons: a systematic review and meta-analyses. *Eur J Obstet Gynecol Reprod Biol*. 2009;146(2):138-48.
100. Pandey S, Shetty A, Hamilton M, Bhattacharya S, Maheshwari A. Obstetric and perinatal outcomes in singleton pregnancies resulting from IVF/ICSI: a systematic review and meta-analysis. *Human reproduction update*. 2012;18(5):485-503.
101. Quinn JA, Munoz FM, Gonik B, Frau L, Cutland C, Mallett-Moore T, et al. Preterm birth: Case definition & guidelines for data collection, analysis, and presentation of immunisation safety data. *Vaccine*. 2016;34(49):6047-56.
102. Nguyen TH, Larsen T, Engholm G, Møller H. Evaluation of ultrasound-estimated date of delivery in 17,450 spontaneous singleton births: do we need to modify Naegele's rule? *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 1999;14(1):23-8.
103. Spong CY. Defining "term" pregnancy: recommendations from the Defining "Term" Pregnancy Workgroup. *Jama*. 2013;309(23):2445-6.
104. Lavender T, Hofmeyr GJ, Neilson JP, Kingdon C, Gyte GM. Caesarean section for non-medical reasons at term. *Cochrane Database Syst Rev*. 2012;2012(3):Cd004660.
105. Knight AK, Conneely KN, Smith AK. Gestational age predicted by DNA methylation: potential clinical and research utility. *Epigenomics*. 2017.
106. Kerstjens JM, de Winter AF, Bocca-Tjeertes IF, Bos AF, Reijneveld SA. Risk of developmental delay increases exponentially as gestational age of preterm infants decreases: a cohort study at age 4 years. *Developmental medicine and child neurology*. 2012;54(12):1096-101.
107. Bhutta AT, Cleves MA, Casey PH, Cradock MM, Anand KJ. Cognitive and behavioral outcomes of school-aged children who were born preterm: a meta-analysis. *Jama*. 2002;288(6):728-37.
108. Kajantie E, Osmond C, Barker DJ, Eriksson JG. Preterm birth--a risk factor for type 2 diabetes? The Helsinki birth cohort study. *Diabetes care*. 2010;33(12):2623-5.

109. Boyle EM, Poulsen G, Field DJ, Kurinczuk JJ, Wolke D, Alfirovic Z, et al. Effects of gestational age at birth on health outcomes at 3 and 5 years of age: population based cohort study. *BMJ (Clinical research ed)*. 2012;344:e896.
110. El Marroun H, Zeegers M, Steegers EA, van der Ende J, Schenk JJ, Hofman A, et al. Post-term birth and the risk of behavioural and emotional problems in early childhood. *International journal of epidemiology*. 2012;41(3):773-81.
111. Gleason JL, Gilman SE, Sundaram R, Yeung E, Putnick DL, Vafai Y, et al. Gestational age at term delivery and children's neurocognitive development. *International journal of epidemiology*. 2022;50(6):1814-23.
112. Lynch CD, Zhang J. The research implications of the selection of a gestational age estimation method. *Paediatric and perinatal epidemiology*. 2007;21 Suppl 2:86-96.
113. Mongelli M, Wilcox M, Gardosi J. Estimating the date of confinement: ultrasonographic biometry versus certain menstrual dates. *Am J Obstet Gynecol*. 1996;174(1 Pt 1):278-81.
114. Skalkidou A, Kullinger M, Georgakis MK, Kieler H, Kesmodel US. Systematic misclassification of gestational age by ultrasound biometry: implications for clinical practice and research methodology in the Nordic countries. *Acta obstetrica et gynecologica Scandinavica*. 2018;97(4):440-4.
115. Simic M, Wåhlin IA, Marsál K, Källén K. Maternal obesity is a potential source of error in mid-trimester ultrasound estimation of gestational age. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 2010;35(1):48-53.
116. Beydoun H, Ugwu B, Oehninger S. Assisted reproduction for the validation of gestational age assessment methods. *Reprod Biomed Online*. 2011;22(4):321-6.
117. Wennerholm UB, Bergh C, Hagberg H, Sultan B, Wennergren M. Gestational age in pregnancies after in vitro fertilization: comparison between ultrasound measurement and actual age. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 1998;12(3):170-4.
118. Tunón K, Eik-Nes SH, Grøttum P, Von Düring V, Kahn JA. Gestational age in pregnancies conceived after in vitro fertilization: a comparison between age assessed from oocyte retrieval, crown-rump length and biparietal diameter. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 2000;15(1):41-6.
119. Schroeder JW, Conneely KN, Cubells JC, Kilaru V, Newport DJ, Knight BT, et al. Neonatal DNA methylation patterns associate with gestational age. *Epigenetics*. 2011;6(12):1498-504.
120. Lee H, Jaffe AE, Feinberg JI, Tryggvadottir R, Brown S, Montano C, et al. DNA methylation shows genome-wide association of NFIX, RAPGEF2 and MSRB3 with gestational age at birth. *International journal of epidemiology*. 2012;41(1):188-99.
121. Parets SE, Conneely KN, Kilaru V, Fortunato SJ, Syed TA, Saade G, et al. Fetal DNA Methylation Associates with Early Spontaneous Preterm Birth and Gestational Age. *PloS one*. 2013;8(6):e67489.
122. Bohlin J, Håberg SE, Magnus P, Reese SE, Gjessing HK, Magnus MC, et al. Prediction of gestational age based on genome-wide differentially methylated regions. *Genome Biol*. 2016;17(1):207.
123. Merid SK, Novoloaca A, Sharp GC, Küpers LK, Kho AT, Roy R, et al. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. *Genome medicine*. 2020;12(1):25.
124. 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*. 2019;11(12):4238-53.
125. Simpkin AJ, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, Ring SM, et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum Mol Genet*. 2015;24(13):3752-63.
126. Knight AK, Craig JM, Theda C, Bækvad-Hansen M, Bybjerg-Grauholm J, Hansen CS, et al. An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biol*. 2016;17(1):206.

127. Hannon E, Schendel D, Ladd-Acosta C, Grove J, Hansen CS, Hougaard DM, et al. Variable DNA methylation in neonates mediates the association between prenatal smoking and birth weight. *Philos Trans R Soc Lond B Biol Sci.* 2019;374(1770):20180120.
128. York TP, Latendresse SJ, Jackson-Cook C, Lapato DM, Moyer S, Wolen AR, et al. Replicated umbilical cord blood DNA methylation loci associated with gestational age at birth. *Epigenetics.* 2020;15(11):1243-58.
129. Cruickshank MN, Oshlack A, Theda C, Davis PG, Martino D, Sheehan P, et al. Analysis of epigenetic changes in survivors of preterm birth reveals the effect of gestational age and evidence for a long term legacy. *Genome medicine.* 2013;5(10):96.
130. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *International journal of epidemiology.* 2013;42(1):111-27.
131. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, et al. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* 2015;16(1):25.
132. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. Hallmarks of aging: An expanding universe. *Cell.* 2023;186(2):243-78.
133. Booth LN, Brunet A. The Aging Epigenome. *Mol Cell.* 2016;62(5):728-44.
134. Sen P, Shah PP, Nativio R, Berger SL. Epigenetic Mechanisms of Longevity and Aging. *Cell.* 2016;166(4):822-39.
135. Unnikrishnan A, Hadad N, Masser DR, Jackson J, Freeman WM, Richardson A. Revisiting the genomic hypomethylation hypothesis of aging. *Annals of the New York Academy of Sciences.* 2018;1418(1):69-79.
136. Seale K, Horvath S, Teschendorff A, Eynon N, Voisin S. Making sense of the ageing methylome. *Nature reviews Genetics.* 2022;23(10):585-605.
137. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.* 2009;5(8):e1000602.
138. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* 2010;20(4):440-6.
139. Day K, Waite LL, Thalacker-Mercer A, West A, Bamman MM, Brooks JD, et al. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol.* 2013;14(9):R102.
140. Palou-Márquez G, Subirana I, Nonell L, Fernández-Sanlés A, Elosua R. DNA methylation and gene expression integration in cardiovascular disease. *Clin Epigenetics.* 2021;13(1):75.
141. Bacos K, Gillberg L, Volkov P, Olsson AH, Hansen T, Pedersen O, et al. Blood-based biomarkers of age-associated epigenetic changes in human islets associate with insulin secretion and diabetes. *Nature communications.* 2016;7:11089.
142. Qazi TJ, Quan Z, Mir A, Qing H. Epigenetics in Alzheimer's Disease: Perspective of DNA Methylation. *Mol Neurobiol.* 2018;55(2):1026-44.
143. Chatsirisupachai K, Lesluyes T, Paraoan L, Van Loo P, de Magalhães JP. An integrative analysis of the age-associated multi-omic landscape across cancers. *Nature communications.* 2021;12(1):2345.
144. Pellegrini C, Pirazzini C, Sala C, Sambati L, Yusipov I, Kalyakulina A, et al. A Meta-Analysis of Brain DNA Methylation Across Sex, Age, and Alzheimer's Disease Points for Accelerated Epigenetic Aging in Neurodegeneration. *Front Aging Neurosci.* 2021;13:639428.
145. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 2013;14(10):R115.
146. Zhu T, Zheng SC, Paul DS, Horvath S, Teschendorff AE. Cell and tissue type independent age-associated DNA methylation changes are not rare but common. *Aging.* 2018;10(11):3541-57.
147. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.* 2010;20(4):434-9.
148. Pérez RF, Tejedor JR, Bayón GF, Fernández AF, Fraga MF. Distinct chromatin signatures of DNA hypomethylation in aging and cancer. *Aging Cell.* 2018;17(3):e12744.

149. Simpson DJ, Olova NN, Chandra T. Cellular reprogramming and epigenetic rejuvenation. *Clin Epigenetics*. 2021;13(1):170.
150. Zhang W, Qu J, Liu GH, Belmonte JCI. The ageing epigenome and its rejuvenation. *Nat Rev Mol Cell Biol*. 2020;21(3):137-50.
151. Lu Y, Brommer B, Tian X, Krishnan A, Meer M, Wang C, et al. Reprogramming to recover youthful epigenetic information and restore vision. *Nature*. 2020;588(7836):124-9.
152. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature reviews Genetics*. 2018;19(6):371-84.
153. Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol*. 2014;15(2):R24.
154. Bocklandt S, Lin W, Sehl ME, Sánchez FJ, Sinsheimer JS, Horvath S, et al. Epigenetic predictor of age. *PloS one*. 2011;6(6):e14821.
155. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*. 2013;49(2):359-67.
156. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging*. 2018;10(7):1758-75.
157. Garagnani P, Bacalini MG, Pirazzini C, Gori D, Giuliani C, Mari D, et al. Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*. 2012;11(6):1132-4.
158. Simpkin AJ, Hemani G, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal and early life influences on epigenetic age in children: a study of mother-offspring pairs from two cohort studies. *Hum Mol Genet*. 2016;25(1):191-201.
159. Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, et al. Age-associated DNA methylation in pediatric populations. *Genome Res*. 2012;22(4):623-32.
160. McEwen LM, O'Donnell KJ, McGill MG, Edgar RD, Jones MJ, MacIsaac JL, et al. The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proc Natl Acad Sci U S A*. 2020;117(38):23329-35.
161. Wu X, Chen W, Lin F, Huang Q, Zhong J, Gao H, et al. DNA methylation profile is a quantitative measure of biological aging in children. *Aging*. 2019;11(22):10031-51.
162. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging*. 2018;10(4):573-91.
163. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging*. 2019;11(2):303-27.
164. Belsky DW, Caspi A, Arseneault L, Baccarelli A, Corcoran DL, Gao X, et al. Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. *Elife*. 2020;9.
165. Belsky DW, Caspi A, Corcoran DL, Sugden K, Poulton R, Arseneault L, et al. DunedinPACE, a DNA methylation biomarker of the pace of aging. *Elife*. 2022;11.
166. Trapp A, Kerepesi C, Gladyshev VN. Profiling epigenetic age in single cells. *Nat Aging*. 2021;1(12):1189-201.
167. de Lima Camillo LP, Lapierre LR, Singh R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging*. 2022;8(1):4.
168. Lemaître JF, Rey B, Gaillard JM, Régis C, Gilot-Fromont E, Débias F, et al. DNA methylation as a tool to explore ageing in wild roe deer populations. *Mol Ecol Resour*. 2022;22(3):1002-15.
169. Horvath S, Haghani A, Zoller JA, Raj K, Sinha I, Robeck TR, et al. Epigenetic clock and methylation studies in marsupials: opossums, Tasmanian devils, kangaroos, and wallabies. *Geroscience*. 2022;44(3):1825-45.
170. Horvath S, Haghani A, Macoretta N, Ablava J, Zoller JA, Li CZ, et al. DNA methylation clocks tick in naked mole rats but queens age more slowly than nonbreeders. *Nat Aging*. 2022;2(1):46-59.
171. Jasinska AJ, Haghani A, Zoller JA, Li CZ, Arneson A, Ernst J, et al. Epigenetic clock and methylation studies in vervet monkeys. *Geroscience*. 2022;44(2):699-717.
172. Horvath S, Lu AT, Haghani A, Zoller JA, Li CZ, Lim AR, et al. DNA methylation clocks for dogs and humans. *Proc Natl Acad Sci U S A*. 2022;119(21):e2120887119.

173. Raj K, Szladovits B, Haghani A, Zoller JA, Li CZ, Black P, et al. Epigenetic clock and methylation studies in cats. *Geroscience*. 2021;43(5):2363-78.
174. Horvath S, Haghani A, Peng S, Hales EN, Zoller JA, Raj K, et al. DNA methylation aging and transcriptomic studies in horses. *Nature communications*. 2022;13(1):40.
175. Lu AT, Fei Z, Haghani A, Robeck TR, Zoller JA, Li CZ, et al. Universal DNA methylation age across mammalian tissues. *bioRxiv*. 2021:2021.01.18.426733.
176. Roshandel D, Chen Z, Canty AJ, Bull SB, Natarajan R, Paterson AD, et al. DNA methylation age calculators reveal association with diabetic neuropathy in type 1 diabetes. *Clin Epigenetics*. 2020;12(1):52.
177. Horvath S, Garagnani P, Bacalini MG, Pirazzini C, Salvioli S, Gentilini D, et al. Accelerated epigenetic aging in Down syndrome. *Aging Cell*. 2015;14(3):491-5.
178. Xu K, Li S, Muskens IS, Elliott N, Myint SS, Pandey P, et al. Accelerated epigenetic aging in newborns with Down syndrome. *Aging Cell*. 2022;21(7):e13652.
179. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. *Aging*. 2015;7(12):1130-42.
180. Köhler F, Bormann F, Raddatz G, Gutekunst J, Corless S, Musch T, et al. Epigenetic deregulation of lamina-associated domains in Hutchinson-Gilford progeria syndrome. *Genome medicine*. 2020;12(1):46.
181. Levine ME, Lu AT, Bennett DA, Horvath S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging*. 2015;7(12):1198-211.
182. Teschendorff AE. A comparison of epigenetic mitotic-like clocks for cancer risk prediction. *Genome medicine*. 2020;12(1):56.
183. Daunay A, Hardy LM, Bouyacoub Y, Sahbatou M, Touvier M, Blanché H, et al. Centenarians consistently present a younger epigenetic age than their chronological age with four epigenetic clocks based on a small number of CpG sites. *Aging*. 2022;14(19):7718-33.
184. Marioni RE, Shah S, McRae AF, Ritchie SJ, Muniz-Terrera G, Harris SE, et al. The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *International journal of epidemiology*. 2015;44(4):1388-96.
185. Föhr T, Waller K, Viljanen A, Sanchez R, Ollikainen M, Rantanen T, et al. Does the epigenetic clock GrimAge predict mortality independent of genetic influences: an 18 year follow-up study in older female twin pairs. *Clin Epigenetics*. 2021;13(1):128.
186. Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *International journal of epidemiology*. 2016;45(2):382-8.
187. Simpkin AJ, Suderman M, Howe LD. Epigenetic clocks for gestational age: statistical and study design considerations. *Clin Epigenetics*. 2017;9:100.
188. Mayne BT, Leemaqz SY, Smith AK, Breen J, Roberts CT, Bianco-Miotto T. Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation. *Epigenomics*. 2017;9(3):279-89.
189. Falick Michaeli T, Spiro A, Sabag O, Karavani G, Yagel S, Eventov-Friedman S, et al. Determining gestational age using genome methylation profile: A novel approach for fetal medicine. *Prenat Diagn*. 2019;39(11):1005-10.
190. Steg LC, Shireby GL, Imm J, Davies JP, Franklin A, Flynn R, et al. Novel epigenetic clock for fetal brain development predicts prenatal age for cellular stem cell models and derived neurons. *Mol Brain*. 2021;14(1):98.
191. Graw S, Camerota M, Carter BS, Helderman J, Hofheimer JA, McGowan EC, et al. NEOage clocks - epigenetic clocks to estimate post-menstrual and postnatal age in preterm infants. *Aging*. 2021;13(20):23527-44.
192. Girchenko P, Lahti J, Czamara D, Knight AK, Jones MJ, Suarez A, et al. Associations between maternal risk factors of adverse pregnancy and birth outcomes and the offspring epigenetic clock of gestational age at birth. *Clin Epigenetics*. 2017;9:49.
193. Monasso GS, Voortman T, Felix JF. Maternal plasma fatty acid patterns in mid-pregnancy and offspring epigenetic gestational age at birth. *Epigenetics*. 2022;17(11):1562-72.

194. Khouja JN, Simpkin AJ, O'Keeffe LM, Wade KH, Houtepen LC, Relton CL, et al. Epigenetic gestational age acceleration: a prospective cohort study investigating associations with familial, sociodemographic and birth characteristics. *Clin Epigenetics*. 2018;10:86.
195. Dieckmann L, Lahti-Pulkkinen M, Kvist T, Lahti J, DeWitt PE, Cruceanu C, et al. Characteristics of epigenetic aging across gestational and perinatal tissues. *Clin Epigenetics*. 2021;13(1):97.
196. Clark J, Bulka CM, Martin CL, Roell K, Santos HP, O'Shea TM, et al. Placental epigenetic gestational aging in relation to maternal sociodemographic factors and smoking among infants born extremely preterm: a descriptive study. *Epigenetics*. 2022;17(13):2389-403.
197. Bright HD, Howe LD, Khouja JN, Simpkin AJ, Suderman M, O'Keeffe LM. Epigenetic gestational age and trajectories of weight and height during childhood: a prospective cohort study. *Clin Epigenetics*. 2019;11(1):194.
198. Knight AK, Smith AK, Conneely KN, Dalach P, Loke YJ, Cheong JL, et al. Relationship between Epigenetic Maturity and Respiratory Morbidity in Preterm Infants. *The Journal of pediatrics*. 2018;198:168-73.e2.
199. Daredia S, Huen K, Van Der Laan L, Collender PA, Nwanaji-Enwerem JC, Harley K, et al. Prenatal and birth associations of epigenetic gestational age acceleration in the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) cohort. *Epigenetics*. 2022;17(13):2006-21.
200. Chen L, Wagner CL, Dong Y, Wang X, Shary JR, Huang Y, et al. Effects of Maternal Vitamin D3 Supplementation on Offspring Epigenetic Clock of Gestational Age at Birth: A Post-hoc Analysis of a Randomized Controlled Trial. *Epigenetics*. 2020;15(8):830-40.
201. Song AY, Feinberg JI, Bakulski KM, Croen LA, Fallin MD, Newschaffer CJ, et al. Prenatal Exposure to Ambient Air Pollution and Epigenetic Aging at Birth in Newborns. *Front Genet*. 2022;13:929416.
202. Monasso GS, Küpers LK, Jaddoe VWV, Heil SG, Felix JF. Associations of circulating folate, vitamin B12 and homocysteine concentrations in early pregnancy and cord blood with epigenetic gestational age: the Generation R Study. *Clin Epigenetics*. 2021;13(1):95.
203. Suarez A, Lahti J, Czamara D, Lahti-Pulkkinen M, Knight AK, Girchenko P, et al. The Epigenetic Clock at Birth: Associations With Maternal Antenatal Depression and Child Psychiatric Problems. *J Am Acad Child Adolesc Psychiatry*. 2018;57(5):321-8.e2.
204. Appleton AA, Lin B, Kennedy EM, Holdsworth EA. Maternal depression and adverse neighbourhood conditions during pregnancy are associated with gestational epigenetic age deceleration. *Epigenetics*. 2022;17(13):1905-19.
205. Monasso GS, Jaddoe VWV, Küpers LK, Felix JF. Epigenetic age acceleration and cardiovascular outcomes in school-age children: The Generation R Study. *Clin Epigenetics*. 2021;13(1):205.
206. Zhang Q, Vallerga CL, Walker RM, Lin T, Henders AK, Montgomery GW, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome medicine*. 2019;11(1):54.
207. Girchenko P, Lahti M, Tuovinen S, Savolainen K, Lahti J, Binder EB, et al. Cohort Profile: Prediction and prevention of preeclampsia and intrauterine growth restriction (PREDO) study. *International journal of epidemiology*. 2017;46(5):1380-1g.
208. Paltiel L, Anita H, Skjerden T, Harbak K, Bækken S, Nina Kristin S, et al. The biobank of the Norwegian Mother and Child Cohort Study – present status. *Norsk Epidemiologi*. 2014;24(1-2).
209. Håberg SE, London SJ, Nafstad P, Nilsen RM, Ueland PM, Vollset SE, et al. Maternal folate levels in pregnancy and asthma in children at age 3 years. *J Allergy Clin Immunol*. 2011;127(1):262-4, 4.e1.
210. Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect*. 2012;120(10):1425-31.
211. McEwen LM, Jones MJ, Lin DTS, Edgar RD, Husquin LT, MacIsaac JL, et al. Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. *Clin Epigenetics*. 2018;10(1):123.

212. Wu MC, Joubert BR, Kuan PF, Håberg SE, Nystad W, Peddada SD, et al. A systematic assessment of normalization approaches for the Infinium 450K methylation platform. *Epigenetics*. 2014;9(2):318-29.
213. Campagna MP, Xavier A, Lechner-Scott J, Maltby V, Scott RJ, Butzkueven H, et al. Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clin Epigenetics*. 2021;13(1):214.
214. Mansell G, Gorrie-Stone TJ, Bao Y, Kumari M, Schalkwyk LS, Mill J, et al. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC genomics*. 2019;20(1):366.
215. Tsai PC, Bell JT. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *International journal of epidemiology*. 2015;44(4):1429-41.
216. Graw S, Henn R, Thompson JA, Koestler DC. pwrEWAS: a user-friendly tool for comprehensive power estimation for epigenome wide association studies (EWAS). *BMC Bioinformatics*. 2019;20(1):218.
217. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*. 1995;57(1):289-300.
218. van Iterson M, van Zwet EW, Heijmans BT, the BC. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology*. 2017;18(1):19.
219. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. 2007;3(9):1724-35.
220. Pearce N, Lawlor DA. Causal inference-so much more than statistics. *International journal of epidemiology*. 2016;45(6):1895-903.
221. Bozack AK, Colicino E, Just AC, Wright RO, Baccarelli AA, Wright RJ, et al. Associations between infant sex and DNA methylation across umbilical cord blood, artery, and placenta samples. *Epigenetics*. 2022;17(10):1080-97.
222. Broere-Brown ZA, Baan E, Schalekamp-Timmermans S, Verburg BO, Jaddoe VWV, Steegers EAP. Sex-specific differences in fetal and infant growth patterns: a prospective population-based cohort study. *Biology of Sex Differences*. 2016;7(1):65.
223. Alur P. Sex Differences in Nutrition, Growth, and Metabolism in Preterm Infants. *Front Pediatr*. 2019;7:22.
224. Bukowski R, Smith GCS, Malone FD, Ball RH, Nyberg DA, Comstock CH, et al. Human Sexual Size Dimorphism in Early Pregnancy. *American Journal of Epidemiology*. 2007;165(10):1216-8.
225. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am J Hum Genet*. 2016;98(4):680-96.
226. Stock SJ, Bauld L. Maternal smoking and preterm birth: An unresolved health challenge. *PLoS Med*. 2020;17(9):e1003386.
227. Markunas CA, Wilcox AJ, Xu Z, Joubert BR, Harlid S, Panduri V, et al. Maternal Age at Delivery Is Associated with an Epigenetic Signature in Both Newborns and Adults. *PloS one*. 2016;11(7):e0156361.
228. Fuchs F, Monet B, Ducruet T, Chaillet N, Audibert F. Effect of maternal age on the risk of preterm birth: A large cohort study. *PloS one*. 2018;13(1):e0191002.
229. Küpers LK, Monnerau C, Sharp GC, Yousefi P, Salas LA, Ghantous A, et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. *Nature communications*. 2019;10(1):1893.
230. Oken E, Kleinman KP, Rich-Edwards J, Gillman MW. A nearly continuous measure of birth weight for gestational age using a United States national reference. *BMC Pediatr*. 2003;3:6.
231. Kozuki N, Lee ACC, Silveira MF, Sania A, Vogel JP, Adair L, et al. The associations of parity and maternal age with small-for-gestational-age, preterm, and neonatal and infant mortality: a meta-analysis. *BMC Public Health*. 2013;13(3):S2.

232. Campagna MP, Xavier A, Stankovich J, Maltby VE, Slee M, Yeh WZ, et al. Parity is associated with long-term differences in DNA methylation at genes related to neural plasticity in multiple sclerosis. *Clin Epigenetics*. 2023;15(1):20.
233. Chen Q, Ming Y, Gan Y, Huang L, Zhao Y, Wang X, et al. The impact of cesarean delivery on infant DNA methylation. *BMC Pregnancy and Childbirth*. 2021;21(1):265.
234. Chirumbolo S, Ortolani R, Veneri D, Raffaelli R, Peroni D, Pigozzi R, et al. Lymphocyte phenotypic subsets in umbilical cord blood compared to peripheral blood from related mothers. *Cytometry B Clin Cytom*. 2011;80(4):248-53.
235. López MC, Palmer BE, Lawrence DA. Phenotypic differences between cord blood and adult peripheral blood. *Cytometry B Clin Cytom*. 2009;76(1):37-46.
236. Hermansen MC. Nucleated red blood cells in the fetus and newborn. *Arch Dis Child Fetal Neonatal Ed*. 2001;84(3):F211-5.
237. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*. 2014;15(2):R31.
238. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum Mol Genet*. 2017;26(R2):R216-r24.
239. 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*. 2012;13(1):86.
240. 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*. 2016;17(1):259.
241. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nature Reviews Genetics*. 2018;19(3):129-47.
242. 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. *Clin Epigenetics*. 2019;11(1):125.
243. Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, S LM, et al. DNA methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics*. 2016;11(5):354-62.
244. de Goede OM, Lavoie PM, Robinson WP. Cord blood hematopoietic cells from preterm infants display altered DNA methylation patterns. *Clin Epigenetics*. 2017;9:39.
245. Gervin K, Page CM, Aass HC, Jansen MA, Fjeldstad HE, Andreassen BK, et al. Cell type specific DNA methylation in cord blood: A 450K-reference data set and cell count-based validation of estimated cell type composition. *Epigenetics*. 2016;11(9):690-8.
246. 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*. 2018;13(9):941-58.
247. Houseman EA, Kim S, Kelsey KT, Wiencke JK. DNA Methylation in Whole Blood: Uses and Challenges. *Curr Environ Health Rep*. 2015;2(2):145-54.
248. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods*. 2018;15(12):1059-66.
249. Rahmani E, Schweiger R, Rhead B, Criswell LA, Barcellos LF, Eskin E, et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature communications*. 2019;10(1):3417.
250. Rahmani E, Jew B, Halperin E. The Effect of Model Directionality on Cell-Type-Specific Differential DNA Methylation Analysis. *Front Bioinform*. 2021;1:792605.
251. Breiman L. Statistical Modeling: The Two Cultures. *Statistical Science*. 2001;16(3):199-231.
252. Engebretsen S, Bohlin J. Statistical predictions with glmnet. *Clin Epigenetics*. 2019;11(1):123.
253. Teschendorff AE. Avoiding common pitfalls in machine learning omic data science. *Nat Mater*. 2019;18(5):422-7.
254. T. H, R. T, J. F. *The Elements of Statistical Learning*. 2 ed. New York, NY: Springer; 2009.
255. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;67(2):301-20.
256. Varin S, Panagiotakos DB. A review of robust regression in biomedical science research. *Arch Med Sci*. 2020;16(5):1267-9.

257. Sørensen Ø, Frigessi A, Thoresen M. MEASUREMENT ERROR IN LASSO: IMPACT AND LIKELIHOOD BIAS CORRECTION. *Statistica Sinica*. 2015;25(2):809-29.
258. Sørensen Ø, Hellton KH, Frigessi A, Thoresen M. Covariate Selection in High-Dimensional Generalized Linear Models With Measurement Error. *Journal of Computational and Graphical Statistics*. 2018;27(4):739-49.
259. Dugué PA, English DR, MacInnis RJ, Joo JE, Jung CH, Milne RL. The repeatability of DNA methylation measures may also affect the power of epigenome-wide association studies. *International journal of epidemiology*. 2015;44(4):1460-1.
260. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010;72(4):417-73.
261. Hofner B, Boccuto L, Göker M. Controlling false discoveries in high-dimensional situations: boosting with stability selection. *BMC Bioinformatics*. 2015;16(1):144.
262. BESAG J, CLIFFORD P. Sequential Monte Carlo p-values. *Biometrika*. 1991;78(2):301-4.
263. Gandy A. Sequential Implementation of Monte Carlo Tests With Uniformly Bounded Resampling Risk. *Journal of the American Statistical Association*. 2009;104(488):1504-11.
264. Gareth James DWTHRT. *An introduction to statistical learning : with applications in R*. New York: Springer; 2013.
265. Wood SN. *Generalized Additive Models: An Introduction with R*. 2nd ed: Chapman and Hall/CRC; 2017.
266. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. *Genome Biology*. 2019;20(1):92.
267. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amodè MR, et al. Ensembl 2021. *Nucleic Acids Research*. 2020;49(D1):D884-D91.
268. McLean CY, Bristol D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*. 2010;28(5):495-501.
269. Bizet M, Defrance M, Calonne E, Bontempi G, Sotiriou C, Fuks F, et al. Improving Infinium MethylationEPIC data processing: re-annotation of enhancers and long noncoding RNA genes and benchmarking of normalization methods. *Epigenetics*. 2022;17(13):2434-54.
270. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*. 2019;47(W1):W199-W205.
271. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*. 2022;50(W1):W216-W21.
272. Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications*. 2015;6:5890.
273. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*. 2000;25(1):25-9.
274. The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res*. 2021;49(D1):D325-d34.
275. Palis J, Segel GB. Developmental biology of erythropoiesis. *Blood Rev*. 1998;12(2):106-14.
276. Elahi S, Ertelt JM, Kinder JM, Jiang TT, Zhang X, Xin L, et al. Immunosuppressive CD71+ erythroid cells compromise neonatal host defence against infection. *Nature*. 2013;504(7478):158-62.
277. Li J, Hale J, Bhagia P, Xue F, Chen L, Jaffray J, et al. Isolation and transcriptome analyses of human erythroid progenitors: BFU-E and CFU-E. *Blood*. 2014;124(24):3636-45.
278. Yu Y, Mo Y, Ebenezer D, Bhattacharyya S, Liu H, Sundaravel S, et al. High resolution methylome analysis reveals widespread functional hypomethylation during adult human erythropoiesis. *J Biol Chem*. 2013;288(13):8805-14.
279. Mei Y, Liu Y, Ji P. Understanding terminal erythropoiesis: An update on chromatin condensation, enucleation, and reticulocyte maturation. *Blood Rev*. 2021;46:100740.
280. Richard A, Vallin E, Romestaing C, Roussel D, Gandrillon O, Gonin-Giraud S. Erythroid differentiation displays a peak of energy consumption concomitant with glycolytic metabolism rearrangements. *PloS one*. 2019;14(9):e0221472.
281. Camaschella C, Pagani A, Silvestri L, Nai A. The mutual crosstalk between iron and erythropoiesis. *Int J Hematol*. 2022;116(2):182-91.

282. Sen T, Chen J, Singbrant S. Decreased PGC1 β expression results in disrupted human erythroid differentiation, impaired hemoglobinization and cell cycle exit. *Scientific reports*. 2021;11(1):17129.
283. Mathangasinghe Y, Fauvet B, Jane SM, Goloubinoff P, Nillegoda NB. The Hsp70 chaperone system: distinct roles in erythrocyte formation and maintenance. *Haematologica*. 2021;106(6):1519-34.
284. Moras M, Lefevre SD, Ostuni MA. From Erythroblasts to Mature Red Blood Cells: Organelle Clearance in Mammals. *Front Physiol*. 2017;8:1076.
285. Kundu M, Lindsten T, Yang CY, Wu J, Zhao F, Zhang J, et al. Ulk1 plays a critical role in the autophagic clearance of mitochondria and ribosomes during reticulocyte maturation. *Blood*. 2008;112(4):1493-502.
286. Chasis JA, Prenant M, Leung A, Mohandas N. Membrane assembly and remodeling during reticulocyte maturation. *Blood*. 1989;74(3):1112-20.
287. Chasis JA, Mohandas N. Erythroblastic islands: niches for erythropoiesis. *Blood*. 2008;112(3):470-8.
288. Chen Z, Zhang Y. Role of Mammalian DNA Methyltransferases in Development. *Annu Rev Biochem*. 2020;89:135-58.
289. de Goede OM, Lavoie PM, Robinson WP. Characterizing the hypomethylated DNA methylation profile of nucleated red blood cells from cord blood. *Epigenomics*. 2016;8(11):1481-94.
290. Xu J, Bauer DE, Kerenyi MA, Vo TD, Hou S, Hsu YJ, et al. Corepressor-dependent silencing of fetal hemoglobin expression by BCL11A. *Proc Natl Acad Sci U S A*. 2013;110(16):6518-23.
291. Low JK, Webb SR, Silva AP, Saathoff H, Ryan DP, Torrado M, et al. CHD4 Is a Peripheral Component of the Nucleosome Remodeling and Deacetylase Complex. *J Biol Chem*. 2016;291(30):15853-66.
292. Varricchio L, Dell'Aversana C, Nebbioso A, Migliaccio G, Altucci L, Mai A, et al. Identification of NuRSERY, a new functional HDAC complex composed by HDAC5, GATA1, EKLF and pERK present in human erythroid cells. *Int J Biochem Cell Biol*. 2014;50:112-22.
293. Yu W, Zhang F, Wang S, Fu Y, Chen J, Liang X, et al. Depletion of polycomb repressive complex 2 core component EED impairs fetal hematopoiesis. *Cell Death Dis*. 2017;8(4):e2744.
294. Ji P, Yeh V, Ramirez T, Murata-Hori M, Lodish HF. Histone deacetylase 2 is required for chromatin condensation and subsequent enucleation of cultured mouse fetal erythroblasts. *Haematologica*. 2010;95(12):2013-21.
295. Menon V, Ghaffari S. Erythroid enucleation: a gateway into a "bloody" world. *Exp Hematol*. 2021;95:13-22.
296. Liang R, Campreciós G, Kou Y, McGrath K, Nowak R, Catherman S, et al. A Systems Approach Identifies Essential FOXO3 Functions at Key Steps of Terminal Erythropoiesis. *PLoS Genet*. 2015;11(10):e1005526.
297. Li X, Mei Y, Yan B, Vitriol E, Huang S, Ji P, et al. Histone deacetylase 6 regulates cytokinesis and erythrocyte enucleation through deacetylation of formin protein mDia2. *Haematologica*. 2017;102(6):984-94.
298. Ji P. New insights into the mechanisms of mammalian erythroid chromatin condensation and enucleation. *Int Rev Cell Mol Biol*. 2015;316:159-82.
299. Ji P, Jayapal SR, Lodish HF. Enucleation of cultured mouse fetal erythroblasts requires Rac GTPases and mDia2. *Nat Cell Biol*. 2008;10(3):314-21.
300. Ubukawa K, Goto T, Asanuma K, Sasaki Y, Guo YM, Kobayashi I, et al. Cdc42 regulates cell polarization and contractile actomyosin rings during terminal differentiation of human erythroblasts. *Scientific reports*. 2020;10(1):11806.
301. Huang NJ, Lin YC, Lin CY, Pishesha N, Lewis CA, Freinkman E, et al. Enhanced phosphocholine metabolism is essential for terminal erythropoiesis. *Blood*. 2018;131(26):2955-66.
302. Sankaran VG, Orkin SH. The switch from fetal to adult hemoglobin. *Cold Spring Harb Perspect Med*. 2013;3(1):a011643.
303. Dumitriu B, Patrick MR, Petschek JP, Cherukuri S, Klingmuller U, Fox PL, et al. Sox6 cell-autonomously stimulates erythroid cell survival, proliferation, and terminal maturation and is thereby an important enhancer of definitive erythropoiesis during mouse development. *Blood*. 2006;108(4):1198-207.

304. Rosenzweig R, Nillegoda NB, Mayer MP, Bukau B. The Hsp70 chaperone network. *Nat Rev Mol Cell Biol.* 2019;20(11):665-80.
305. Liu W, Tang X, Qi X, Fu X, Ghimire S, Ma R, et al. The Ubiquitin Conjugating Enzyme: An Important Ubiquitin Transfer Platform in Ubiquitin-Proteasome System. *Int J Mol Sci.* 2020;21(8).
306. Gammoh N. The multifaceted functions of ATG16L1 in autophagy and related processes. *J Cell Sci.* 2020;133(20).
307. Walczak M, Martens S. Dissecting the role of the Atg12-Atg5-Atg16 complex during autophagosome formation. *Autophagy.* 2013;9(3):424-5.
308. Sankaran VG, Nathan DG. Reversing the Hemoglobin Switch. *New England Journal of Medicine.* 2010;363(23):2258-60.
309. Jennifer B, Berg V, Modak M, Puck A, Seyerl-Jiresch M, König S, et al. Transferrin receptor 1 is a cellular receptor for human heme-albumin. *Commun Biol.* 2020;3(1):621.
310. Chen K, Liu J, Heck S, Chasis JA, An X, Mohandas N. Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis. *Proc Natl Acad Sci U S A.* 2009;106(41):17413-8.
311. Mavilio F, Giampaolo A, Carè A, Migliaccio G, Calandrini M, Russo G, et al. Molecular mechanisms of human hemoglobin switching: selective undermethylation and expression of globin genes in embryonic, fetal, and adult erythroblasts. *Proc Natl Acad Sci U S A.* 1983;80(22):6907-11.
312. Lee WS, McColl B, Maksimovic J, Vadolas J. Epigenetic interplay at the β -globin locus. *Biochim Biophys Acta Gene Regul Mech.* 2017;1860(4):393-404.
313. Xu J, Sankaran VG, Ni M, Menne TF, Puram RV, Kim W, et al. Transcriptional silencing of $\{\gamma\}$ -globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes & development.* 2010;24(8):783-98.
314. Cao H, Stamatoyannopoulos G, Jung M. Induction of human gamma globin gene expression by histone deacetylase inhibitors. *Blood.* 2004;103(2):701-9.
315. de Vasconcellos JF, Tumburu L, Byrnes C, Lee YT, Xu PC, Li M, et al. IGF2BP1 overexpression causes fetal-like hemoglobin expression patterns in cultured human adult erythroblasts. *Proc Natl Acad Sci U S A.* 2017;114(28):E5664-e72.
316. Chambers CB, Gross J, Pratt K, Guo X, Byrnes C, Lee YT, et al. The mRNA-Binding Protein IGF2BP1 Restores Fetal Hemoglobin in Cultured Erythroid Cells from Patients with β -Hemoglobin Disorders. *Mol Ther Methods Clin Dev.* 2020;17:429-40.
317. Wells RM, Brittain T. Transition to cooperative oxygen-binding by embryonic haemoglobin in mice. *J Exp Biol.* 1981;90:351-5.
318. Hutter D, Kingdom J, Jaeggi E. Causes and mechanisms of intrauterine hypoxia and its impact on the fetal cardiovascular system: a review. *Int J Pediatr.* 2010;2010:401323.
319. L. Thompson SC, B. Telugu, S. Turan. Intrauterine hypoxia: clinical consequences and therapeutic perspectives. *Research and Reports in Neonatology.* 2015;2015:5:79-89.
320. Minior VK, Levine B, Ferber A, Guller S, Divon MY. Nucleated Red Blood Cells as a Marker of Acute and Chronic Fetal Hypoxia in a Rat Model. *Rambam Maimonides Med J.* 2017;8(2).
321. Leonard MO, Godson C, Brady HR, Taylor CT. Potentiation of glucocorticoid activity in hypoxia through induction of the glucocorticoid receptor. *J Immunol.* 2005;174(4):2250-7.
322. Bauer A, Tronche F, Wessely O, Kellendonk C, Reichardt HM, Steinlein P, et al. The glucocorticoid receptor is required for stress erythropoiesis. *Genes & development.* 1999;13(22):2996-3002.
323. Varricchio L, Migliaccio AR. The role of glucocorticoid receptor (GR) polymorphisms in human erythropoiesis. *Am J Blood Res.* 2014;4(2):53-72.
324. Paulson RF, Ruan B, Hao S, Chen Y. Stress Erythropoiesis is a Key Inflammatory Response. *Cells.* 2020;9(3).
325. Porayette P, Paulson RF. BMP4/Smad5 dependent stress erythropoiesis is required for the expansion of erythroid progenitors during fetal development. *Dev Biol.* 2008;317(1):24-35.
326. Xiang J, Wu DC, Chen Y, Paulson RF. In vitro culture of stress erythroid progenitors identifies distinct progenitor populations and analogous human progenitors. *Blood.* 2015;125(11):1803-12.
327. Solano ME, Arck PC. Steroids, Pregnancy and Fetal Development. *Front Immunol.* 2019;10:3017.

328. Emin Turkay K, Aslı O, Gozde U, Inanc M. The Effects of Glucocorticoids on Fetal and Placental Development. In: Xiaoxiao Q, editor. *Glucocorticoids*. Rijeka: IntechOpen; 2012. p. Ch. 13.
329. Moisiadis VG, Matthews SG. Glucocorticoids and fetal programming part 1: Outcomes. *Nat Rev Endocrinol*. 2014;10(7):391-402.
330. Davari-Tanha F, Kaveh M, Nemati S, Javadian P, Salmanian B. Nucleated red blood cells count in pregnancies with idiopathic intra-uterine growth restriction. *J Family Reprod Health*. 2014;8(2):77-81.
331. Gasparović VE, Ahmetasević SG, Colić A. Nucleated red blood cells count as first prognostic marker for adverse neonatal outcome in severe preeclamptic pregnancies. *Coll Antropol*. 2012;36(3):853-7.
332. Hebbar S, Misha M, Rai L. Significance of maternal and cord blood nucleated red blood cell count in pregnancies complicated by preeclampsia. *J Pregnancy*. 2014;2014:496416.
333. Sokou R, Ioakeimidis G, Lampridou M, Pouliakis A, Tsantes AG, Tsantes AE, et al. Nucleated Red Blood Cells: Could They Be Indicator Markers of Illness Severity for Neonatal Intensive Care Unit Patients? *Children (Basel)*. 2020;7(11).
334. Morton SU, Brettin K, Feldman HA, Leeman KT. Association of nucleated red blood cell count with mortality among neonatal intensive care unit patients. *Pediatr Neonatol*. 2020;61(6):592-7.
335. Grzywa TM, Nowis D, Golab J. The role of CD71(+) erythroid cells in the regulation of the immune response. *Pharmacol Ther*. 2021;228:107927.
336. Arosa FA, Pereira CF, Fonseca AM. Red blood cells as modulators of T cell growth and survival. *Curr Pharm Des*. 2004;10(2):191-201.
337. Schäkel K, von Kietzell M, Hänsel A, Ebling A, Schulze L, Haase M, et al. Human 6-sulfo LacNAc-expressing dendritic cells are principal producers of early interleukin-12 and are controlled by erythrocytes. *Immunity*. 2006;24(6):767-77.
338. Fonseca AM, Porto G, Uchida K, Arosa FA. Red blood cells inhibit activation-induced cell death and oxidative stress in human peripheral blood T lymphocytes. *Blood*. 2001;97(10):3152-60.
339. Delyea C, Bozorgmehr N, Koleva P, Dunsmore G, Shahbaz S, Huang V, et al. CD71(+) Erythroid Suppressor Cells Promote Fetomaternal Tolerance through Arginase-2 and PDL-1. *J Immunol*. 2018;200(12):4044-58.
340. Elahi S. New insight into an old concept: role of immature erythroid cells in immune pathogenesis of neonatal infection. *Front Immunol*. 2014;5:376.
341. Cui L, Takada H, Takimoto T, Fujiyoshi J, Ishimura M, Hara T. Immunoregulatory function of neonatal nucleated red blood cells in humans. *Immunobiology*. 2016;221(8):853-61.
342. Tsafaras GP, Ntontsi P, Xanthou G. Advantages and Limitations of the Neonatal Immune System. *Front Pediatr*. 2020;8:5.
343. Jones MJ, Dinh L, Razzaghian HR, Goede Od, MacIsaac JL, Morin AM, et al. Differences in DNA methylation of white blood cell types at birth and in adulthood reflect postnatal immune maturation and influence accuracy of cell type prediction. *bioRxiv*. 2018:399279.
344. Peterson LS, Hedou J, Ganio EA, Stelzer IA, Feyaerts D, Harbert E, et al. Single-Cell Analysis of the Neonatal Immune System Across the Gestational Age Continuum. *Front Immunol*. 2021;12:714090.
345. Romero R, Espinoza J, Gonçalves LF, Kusanovic JP, Friel L, Hassan S. The role of inflammation and infection in preterm birth. *Semin Reprod Med*. 2007;25(1):21-39.
346. Gomez-Lopez N, StLouis D, Lehr MA, Sanchez-Rodriguez EN, Arenas-Hernandez M. Immune cells in term and preterm labor. *Cell Mol Immunol*. 2014;11(6):571-81.
347. Thomson AJ, Telfer JF, Young A, Campbell S, Stewart CJ, Cameron IT, et al. Leukocytes infiltrate the myometrium during human parturition: further evidence that labour is an inflammatory process. *Human reproduction (Oxford, England)*. 1999;14(1):229-36.
348. Shynlova O, Nedd-Roderique T, Li Y, Dorogin A, Nguyen T, Lye SJ. Infiltration of myeloid cells into decidua is a critical early event in the labour cascade and post-partum uterine remodelling. *J Cell Mol Med*. 2013;17(2):311-24.
349. Noguchi K, Iwasaki K, Shitashige M, Endo H, Kondo H, Ishikawa I. Cyclooxygenase-2-dependent prostaglandin E2 down-regulates intercellular adhesion molecule-1 expression via EP2/EP4 receptors in interleukin-1beta-stimulated human gingival fibroblasts. *J Dent Res*. 2000;79(12):1955-61.

350. Brown NL, Alvi SA, Elder MG, Bennett PR, Sullivan MH. Regulation of prostaglandin production in intact fetal membranes by interleukin-1 and its receptor antagonist. *The Journal of endocrinology*. 1998;159(3):519-26.
351. Brown NL, Alvi SA, Elder MG, Bennett PR, Sullivan MH. A spontaneous induction of fetal membrane prostaglandin production precedes clinical labour. *The Journal of endocrinology*. 1998;157(2):R1-6.
352. Olson DM, Ammann C. Role of the prostaglandins in labour and prostaglandin receptor inhibitors in the prevention of preterm labour. *Front Biosci*. 2007;12:1329-43.
353. Sabbatinelli G, Fantasia D, Palka C, Morizio E, Alfonsi M, Calabrese G. Isolation and Enrichment of Circulating Fetal Cells for NIPD: An Overview. *Diagnostics (Basel)*. 2021;11(12).
354. Jepsen K, Solum D, Zhou T, McEvelly RJ, Kim HJ, Glass CK, et al. SMRT-mediated repression of an H3K27 demethylase in progression from neural stem cell to neuron. *Nature*. 2007;450(7168):415-9.
355. Jepsen K, Gleiberman AS, Shi C, Simon DI, Rosenfeld MG. Cooperative regulation in development by SMRT and FOXP1. *Genes & development*. 2008;22(6):740-5.
356. Pei L, Leblanc M, Barish G, Atkins A, Nofsinger R, Whyte J, et al. Thyroid hormone receptor repression is linked to type I pneumocyte-associated respiratory distress syndrome. *Nat Med*. 2011;17(11):1466-72.
357. Nofsinger RR, Li P, Hong SH, Jonker JW, Barish GD, Ying H, et al. SMRT repression of nuclear receptors controls the adipogenic set point and metabolic homeostasis. *Proc Natl Acad Sci U S A*. 2008;105(50):20021-6.
358. Reilly SM, Bhargava P, Liu S, Gangl MR, Gorgun C, Nofsinger RR, et al. Nuclear receptor corepressor SMRT regulates mitochondrial oxidative metabolism and mediates aging-related metabolic deterioration. *Cell Metab*. 2010;12(6):643-53.
359. Sandovici I, Georgopoulou A, Pérez-García V, Hufnagel A, López-Tello J, Lam BYH, et al. The imprinted Igf2-Igf2r axis is critical for matching placental microvasculature expansion to fetal growth. *Developmental Cell*. 2022;57(1):63-79.e8.
360. Sakamoto T, Matsuura TR, Wan S, Ryba DM, Kim JU, Won KJ, et al. A Critical Role for Estrogen-Related Receptor Signaling in Cardiac Maturation. *Circ Res*. 2020;126(12):1685-702.
361. Alaynick WA, Kondo RP, Xie W, He W, Dufour CR, Downes M, et al. ERRgamma directs and maintains the transition to oxidative metabolism in the postnatal heart. *Cell Metab*. 2007;6(1):13-24.
362. Piazza R, Magistroni V, Redaelli S, Mauri M, Massimino L, Sessa A, et al. SETBP1 induces transcription of a network of development genes by acting as an epigenetic hub. *Nature communications*. 2018;9(1):2192.
363. Antonyan L, Ernst C. Putative Roles of SETBP1 Dosage on the SET Oncogene to Affect Brain Development. *Front Neurosci*. 2022;16:813430.
364. Hoshino A, Horvath S, Sridhar A, Chitsazan A, Reh TA. Synchrony and asynchrony between an epigenetic clock and developmental timing. *Scientific reports*. 2019;9(1):3770.
365. Oh ES, Petronis A. Origins of human disease: the chrono-epigenetic perspective. *Nature reviews Genetics*. 2021;22(8):533-46.
366. Marcheva B, Ramsey KM, Peek CB, Affinati A, Maury E, Bass J. Circadian clocks and metabolism. *Handb Exp Pharmacol*. 2013(217):127-55.
367. Shimba A, Ikuta K. Glucocorticoids Regulate Circadian Rhythm of Innate and Adaptive Immunity. *Front Immunol*. 2020;11:2143.
368. Fuller PM, Gooley JJ, Saper CB. Neurobiology of the sleep-wake cycle: sleep architecture, circadian regulation, and regulatory feedback. *J Biol Rhythms*. 2006;21(6):482-93.
369. Astiz M, Oster H. Feto-Maternal Crosstalk in the Development of the Circadian Clock System. *Front Neurosci*. 2020;14:631687.
370. Dickmeis T. Glucocorticoids and the circadian clock. *The Journal of endocrinology*. 2009;200(1):3-22.
371. Vershinina O, Bacalini MG, Zaikin A, Franceschi C, Ivanchenko M. Disentangling age-dependent DNA methylation: deterministic, stochastic, and nonlinear. *Scientific reports*. 2021;11(1):9201.

372. Shannon CE. The mathematical theory of communication. 1963. MD Comput. 1997;14(4):306-17.
373. Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PloS one*. 2010;5(1):e8888.
374. 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. *Clin Epigenetics*. 2017;9(1):75.
375. Zar Kyaw T, Yamaguchi S, Imai C, Uematsu M, Sato N. The utility of post-test newborn blood spot screening cards for epigenetic association analyses: association between HIF3A methylation and birth weight-for-gestational age. *Journal of Human Genetics*. 2019;64(8):795-801.
376. Gavriel G, Modi N, Stanier P, Moore GE. Neonatal buccal cell collection for DNA analysis. *Arch Dis Child Fetal Neonatal Ed*. 2005;90(2):F187.
377. Robinson WP, Price EM. The human placental methylome. *Cold Spring Harb Perspect Med*. 2015;5(5):a023044.
378. Bergstedt J, Azzou SAK, Tsuo K, Jaquaniello A, Urrutia A, Rotival M, et al. The immune factors driving DNA methylation variation in human blood. *Nature communications*. 2022;13(1):5895.
379. Salas LA, Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, et al. Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nature communications*. 2022;13(1):761.
380. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*. 2017;9(5):757-68.
381. de Goede OM, Razzaghian HR, Price EM, Jones MJ, Kobor MS, Robinson WP, et al. Nucleated red blood cells impact DNA methylation and expression analyses of cord blood hematopoietic cells. *Clin Epigenetics*. 2015;7(1):95.
382. Vellame DS, Shireby G, MacCalman A, Dempster EL, Burrage J, Gorrie-Stone T, et al. Uncertainty quantification of reference-based cellular deconvolution algorithms. *Epigenetics*. 2023;18(1):2137659.
383. Jing H, Zheng SC, Breeze CE, Beck S, Teschendorff AE. Calling differential DNA methylation at cell-type resolution: an objective status-quo. *bioRxiv*. 2019:822940.
384. Rahmani E, Jew B, Schweiger R, Rhead B, Criswell LA, Barcellos LF, et al. Calling differential DNA methylation at cell-type resolution: addressing misconceptions and best practices. *bioRxiv*. 2021:2021.02.14.431168.
385. Jing H, Zheng SC, Breeze CE, Beck S, Teschendorff AE. Calling differential DNA methylation at cell-type resolution: avoiding misconceptions and promoting best practices. *bioRxiv*. 2021:2021.02.28.433245.
386. Patel RR, Steer P, Doyle P, Little MP, Elliott P. Does gestation vary by ethnic group? A London-based study of over 122 000 pregnancies with spontaneous onset of labour. *International journal of epidemiology*. 2004;33(1):107-13.
387. 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*. 2011;91(8):728-36.
388. Corfield EC, Frei O, Shadrin AA, Rahman Z, Lin A, Athanasiu L, et al. The Norwegian Mother, Father, and Child cohort study (MoBa) genotyping data resource: MoBaPsychGen pipeline v.1. *bioRxiv*. 2022:2022.06.23.496289.

Paper 1







Nucleated red blood cells explain most of the association between DNA methylation and gestational age

1

<https://doi.org/10.1038/s42003-023-04584-w>

OPEN

Nucleated red blood cells explain most of the association between DNA methylation and gestational age

Kristine L. Haftorn ^{1,2✉}, William R. P. Denault ^{1,3}, Yunsung Lee ¹, Christian M. Page^{1,4}, Julia Romanowska ^{1,5}, Robert Lyle ^{1,6}, Øyvind E. Næss^{2,7}, Dana Kristjansson^{1,8}, Per M. Magnus¹, Siri E. Håberg¹, Jon Bohlin ^{1,9,10} & Astanand Jugessur^{1,5,10}

Determining if specific cell type(s) are responsible for an association between DNA methylation (DNAm) and a given phenotype is important for understanding the biological mechanisms underlying the association. Our EWAS of gestational age (GA) in 953 newborns from the Norwegian MoBa study identified 13,660 CpGs significantly associated with GA ($p_{\text{Bonferroni}} < 0.05$) after adjustment for cell type composition. When the CellDMC algorithm was applied to explore cell-type specific effects, 2,330 CpGs were significantly associated with GA, mostly in nucleated red blood cells [nRBCs; $n = 2,030$ (87%)]. Similar patterns were found in another dataset based on a different array and when applying an alternative algorithm to CellDMC called Tensor Composition Analysis (TCA). Our findings point to nRBCs as the main cell type driving the DNAm–GA association, implicating an epigenetic signature of erythropoiesis as a likely mechanism. They also explain the poor correlation observed between epigenetic age clocks for newborns and those for adults.

¹Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway. ²Institute of Health and Society, University of Oslo, Oslo, Norway. ³Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA. ⁴Department of Physical Health and Ageing, Division of Mental and Physical Health, Norwegian Institute of Public Health, Oslo, Norway. ⁵Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway. ⁶Department of Medical Genetics, Oslo University Hospital and University of Oslo, Oslo, Norway. ⁷Division of Mental and Physical Health, Norwegian Institute of Public Health, Oslo, Norway. ⁸Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway. ⁹Division for Infection Control and Environmental Health, Department of Infectious Disease Epidemiology and Modelling, Norwegian Institute of Public Health, Oslo, Norway. ¹⁰These authors jointly supervised this work: Jon Bohlin, Astanand Jugessur. ✉email: KristineLokas.Haftorn@fhi.no

Gestational age (GA) is intimately linked to fetal development. Even slight variations in GA at birth are associated with a wide variety of perinatal health outcomes, some of which have important clinical consequences^{1–5}. Epigenetic modifications, such as DNA methylation (DNAm), play a critical role in fetal development^{6–8}. DNAm has also been shown to be robustly associated with GA at thousands of CpG sites^{5,9–12}. The strong association between DNAm and GA probably reflects biological processes related to fetal development, but the specific mechanisms underlying this association are still unknown. Thus, elucidating these mechanisms may provide a deeper understanding of the molecular processes involved in normal as well as aberrant fetal growth and development.

Most of the previous epigenome-wide association studies (EWASs) of GA were based on DNAm data generated on the Illumina Infinium HumanMethylation450 array (450k) or its predecessor, the Illumina Infinium HumanMethylation27 array (27k)^{5,9,12}. These arrays were designed to cover mainly gene promoters and protein-coding regions^{13,14}. In December 2015, 450k was replaced by the more comprehensive Illumina Infinium MethylationEPIC array (EPIC), which employs the same technology as 450k for measuring DNAm but contains almost twice the number of CpG sites (~850,000) and has a higher coverage of CpGs in regulatory regions¹³. Despite the substantial improvement in genome-wide coverage of regulatory regions and the higher reproducibility and reliability of EPIC¹³, studies investigating the association between GA and DNAm data generated on EPIC are lacking. It is also uncertain whether the extra regulatory CpGs on EPIC are useful in explaining the association between GA and DNAm.

Most studies exploring the link between DNAm and GA are based on samples from cord blood, which comprises a mixture of cell types¹⁵. As cell-type proportions vary substantially across individuals and DNAm is highly cell-type specific¹⁶, it is customary to adjust for cell-type proportions in statistical models in order to avoid bias¹⁷. Several cellular deconvolution algorithms and cord-blood reference panels are available to infer cell-type proportions from heterogeneous samples and adjust for cord blood cell-type composition in newborn DNAm data^{18–20}. However, including cell-type proportions as covariates in the statistical model will not necessarily provide insight as to how cell types influence the association between the explanatory variable and DNAm. One solution is to perform an EWAS in isolated cell types. However, cell sorting of whole-blood samples is costly, especially in large cohort studies with hundreds of thousands of participants.

To counter this, statistical algorithms have been developed to allow the detection of cell-type specific differential DNAm within a heterogeneous mixture of cells without the need for cell sorting or single-cell methods^{21–24}. One example is CellDMC, by Zheng et al.²⁴, which incorporates interaction terms between the phenotype of interest and the estimated cell-type fractions in a linear modeling framework. Another example is Tensor Composition Analysis (TCA), by Rahmani et al.²³, which employs matrix factorization to infer cell-type specific DNAm signals that are subsequently used to search for associations in each cell type separately. Exploring cell-type specific associations can be essential to decipher the biological underpinnings of an association between DNAm and a specific phenotype of interest²⁵. Whilst changes in cord blood cell-type proportions have been reported for GA^{26,27}, studies on cell-type specific epigenetic associations with GA are lacking.

To bridge these knowledge gaps, we investigate the association between cord blood DNAm and GA using an EPIC-derived DNAm dataset comprising 953 newborns and a 450k-derived dataset comprising 1062 newborns. Both datasets are from the Norwegian Mother, Father, and Child Cohort Study (MoBa)²⁸. We apply CellDMC to these datasets to determine the relationship between cell-type specific DNAm and GA. We also apply TCA as an alternative method for cell-type-specific analysis. The results show many CpGs associated with GA, predominantly in nucleated red blood cells (nRBCs). This association reflects an epigenetic signature of erythropoiesis in fetal development and provides a biologically plausible rationale for the consistently observed strong association between DNAm and GA. It also helps explain the observed incompatibility between epigenetic age clocks for newborns and those for adults.

Results

Study sample characteristics. We analysed cord blood DNAm in newborns from two substudies in MoBa. The main study sample consisted of 953 naturally conceived newborns from the Study of Assisted Reproductive Technology (START), in which DNAm was measured using the EPIC array^{29,30}. We also used another dataset consisting of 1062 newborns (referred to as MoBa1 hereafter) with DNAm measured using the 450k array¹⁰. GA ranged from 216–305 days (mean 280.1 days, SD ± 10.7 days) in START and 209–301 days (mean 279.8 days, SD ± 10.8 days) in MoBa1. Table 1 summarizes the key demographic and clinical characteristics of these two datasets. More MoBa1 mothers continued to smoke during pregnancy compared to START mothers

Table 1 Characteristics of the mothers and newborns in START and MoBa1.

Characteristics	START <i>n</i> = 956	MoBa1 <i>n</i> = 1062	<i>p</i> value ^a
Mothers			
Age (years), mean (SD)	29.9 (4.7)	29.9 (4.3)	0.800
Smoking, <i>n</i> (%)			0.033
No smoking before or during pregnancy	478 (50%)	522 (49%)	
Smoked, but quit before pregnancy	245 (26%)	233 (22%)	
Smoked, but quit early in pregnancy	131 (14%)	154 (15%)	
Continued smoking during pregnancy	102 (11%)	153 (14%)	
Newborns			
GA in days, mean (SD)	280.1 (10.7)	279.8 (10.8)	0.400
GA in days, min	216	209	
GA in days, max	305	301	
Birth weight in grams, mean (SD)	3657 (521)	3643 (539)	0.500
Sex (male), <i>n</i> (%)	455 (47%)	569 (54%)	0.007

SD standard deviation, GA gestational age.

^aWilcoxon rank-sum test; Pearson's Chi-squared test.

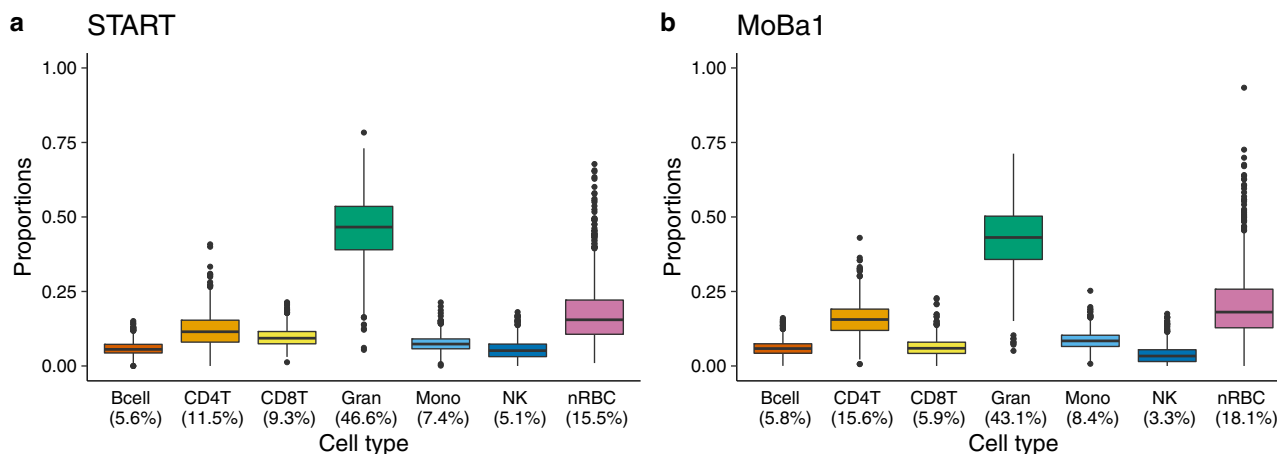


Fig. 1 Estimated proportions of seven main cell types in cord blood. **a** Estimated proportions of cell types in the START dataset ($n = 953$, EPIC-based). **b** Estimated proportions of cell types in the MoBa1 dataset ($n = 1062$, 450k-based). The upper and lower box limits correspond to the interquartile range (25 to 75% of the values for each cell type) and the horizontal line in the box represents the median value. The whiskers outstretch 1.5 times the box height from the top and bottom of the box. The dots outside the whiskers represent outliers beyond the interquartile range. The percentage below each cell type denotes the median proportion of that cell type. Bcell B-cell, CD4T CD4 + T-cell, CD8T CD8 + T-cell, Gran granulocyte, Mono monocyte, NK natural killer cell, nRBC nucleated red blood cell.

($p = 0.033$, Table 1). There were also more boys in MoBa1 than in START ($p = 0.007$, Table 1).

Analyses of cell-type composition. We estimated the proportion of each of the seven main cell types in cord blood (B-cells, CD4 + T-cells, CD8 + T-cells, granulocytes, monocytes, natural killer cells, and nRBCs) separately in START and MoBa1, using a combined reference dataset consisting of cell-type specific DNAm profiles in cord blood¹⁹ (Fig. 1 and Supplementary Data 1). As expected from the reference data, granulocytes and nRBCs were the two most abundant cell types in both datasets. The results of a principal component analysis (PCA) of cell-type proportions in START further confirmed that granulocytes and nRBCs explained most of the variance in cell-type composition (Supplementary Fig. 1 and Supplementary Table 1).

We examined the proportion of each cell type in START and found significant correlations with GA in B-cells (Pearson correlation $r = -0.21$, $p = 6.30 \times 10^{-11}$), CD4 + T-cells ($r = -0.10$, $p = 0.002$), granulocytes ($r = 0.20$, $p = 5.77 \times 10^{-10}$), and nRBCs ($r = -0.08$, $p = 0.010$; see Supplementary Fig. 2 for more details).

Conventional EWAS of GA. First, we applied a linear mixed effects regression model to the EPIC-derived START dataset where the outcome was DNAm level at each CpG, the exposure was GA, and the following were included as covariates: cell-type proportions, newborn sex, maternal age, maternal smoking, and batch (see Methods for details). This model is referred to as the conventional EWAS model throughout this paper, since this framework is routinely adopted in the majority of published EWASs. We identified 13,660 CpGs significantly associated with GA after applying a Bonferroni correction for multiple testing (Bonferroni-corrected p value (p_B) < 0.05 , Fig. 2a and Supplementary Data 2). About 7639 (56%) of the GA-associated CpGs were only present on the EPIC array and were distributed across the genome (Supplementary Fig. 3). Most of the GA-associated CpGs in the conventional EWAS were hypermethylated [$n = 9503$ (70%), Fig. 3a].

Cell-type specific analyses of the association between DNAm and GA. We applied CellDMC to investigate cell-type specific

DNAm in the START dataset and identified 2,330 CpGs significantly associated with GA ($p_B < 0.05$, Fig. 2b–h). Most of these CpGs ($n = 2030$, 87%) were specific for nRBCs (Fig. 2h), and only a few of the CpGs ($n = 31$ –157 and 1.3–6.7%) were identified in the other cell types. Moreover, 522 of the 2330 cell-type-specific CpGs associated with GA were also identified in the conventional EWAS. Detailed results of the CellDMC analyses are provided in Supplementary Data 3.

CpGs that were associated with GA in CD4 + T-cells and monocytes were predominantly hypermethylated [CD4 + T-cells: $n = 67$ (65%), Fig. 3c; monocytes: $n = 29$ (78%), Fig. 3f]. We found an almost equal number of hyper- and hypomethylated CpGs associated with GA in B-cells [hypermethylated $n = 29$ (55%); hypomethylated $n = 24$ (45%); Fig. 3b] and CD8 + T-cells [hypermethylated $n = 13$ (42%); hypomethylated $n = 18$ (58%); Fig. 3d]. In contrast, GA-associated CpGs specific for granulocytes, natural killer cells, and nRBCs were predominantly hypomethylated [granulocytes: $n = 97$ (71%), Fig. 3e; natural killer cells: $n = 97$ (62%), Fig. 3g; nRBCs: $n = 1888$ (93%), Fig. 3h].

Impact of the type of DNAm array: 450k versus EPIC. To determine whether the type of DNAm array had an impact on the cell-type specific results, given the lower coverage of regulatory CpGs on 450k compared to EPIC, we repeated the CellDMC analysis on MoBa1 ($n = 1062$ newborns) in which DNAm was measured using 450k. The results showed a similar pattern of cell-type specific DNAm associated with GA, despite fewer significant CpGs overall ($n = 373$, $p_B < 0.05$, Supplementary Data 4 and Supplementary Fig. 4). Specifically, 62% ($n = 231$) of the Bonferroni-significant CpGs mapped to nRBCs.

To further assess the robustness of our findings, we used the r value approach of ref. 31 to compare the results from START and MoBa1. This approach tests if a CpG is significantly associated in two separate studies and then computes the corresponding false discovery rate (FDR) value of this test, which is referred to as the r value (see Methods for details). If the r value was < 0.05 , we deemed a GA–CpG association detected in START as successfully replicated in MoBa1. Among 1129 nRBC-specific CpGs detected in START that were also available on the 450k array, 174 CpGs were significantly replicated in MoBa1 ($r < 0.05$, Fig. 4 and Supplementary Data 5). The results were also consistent in terms of the direction of effect, except for one CpG (cg13746414). Importantly, there was no overlap in CpGs between

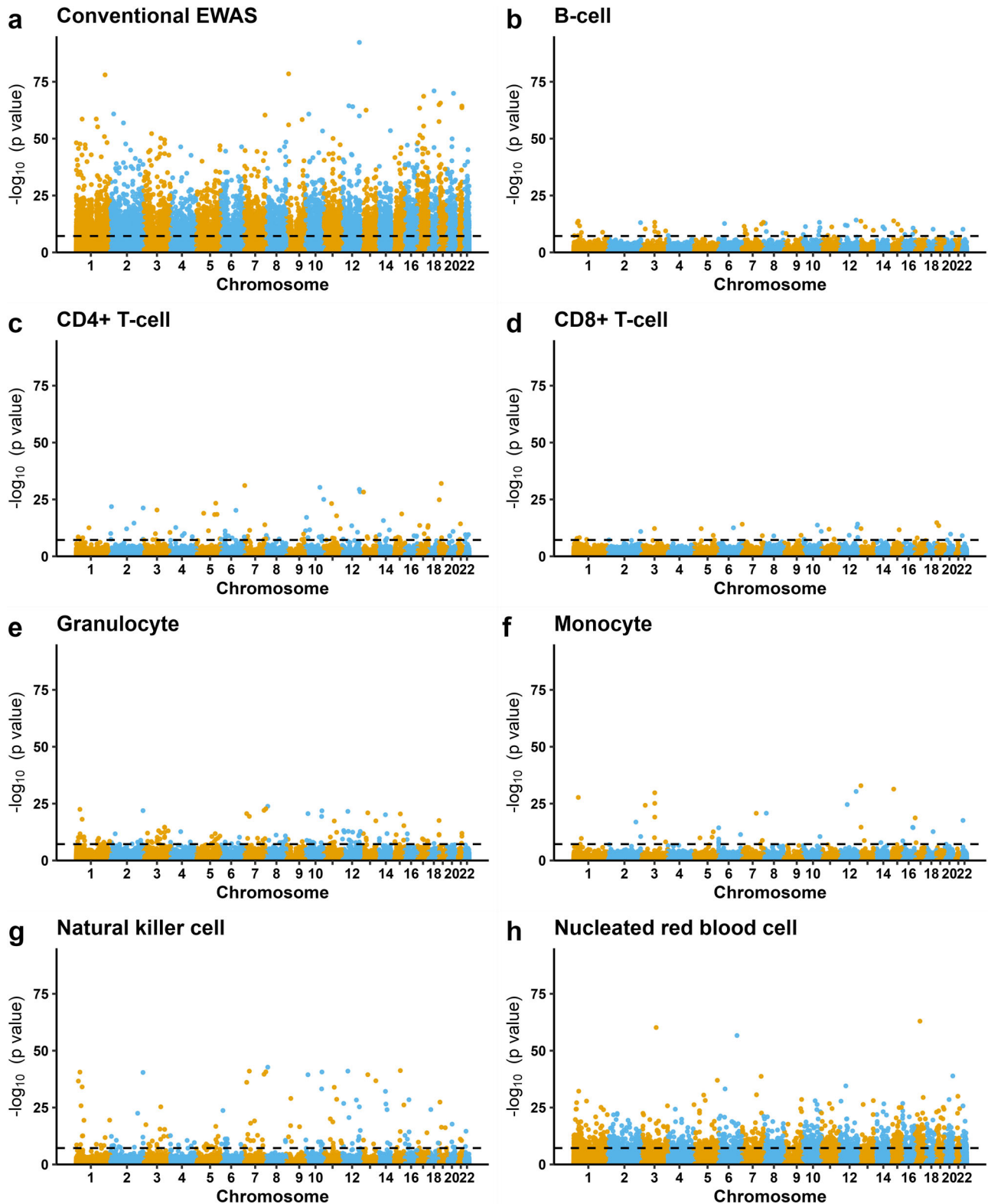


Fig. 2 Manhattan plots of the epigenome-wide DNAm associated with GA in START ($n = 953$). **a** Results from the conventional EWAS where we adjusted for the estimated cell-type proportions (see Methods for details of the statistical model). **b–h** Results for each of the seven cell types from the cell-type specific analysis using CellDMC. CpG loci are aligned on the x-axis according to their genomic coordinates. The y-axis represents the $-\log_{10} p$ values. The dashed black line denotes the Bonferroni-corrected genome-wide significance threshold ($p_B < 0.05$).

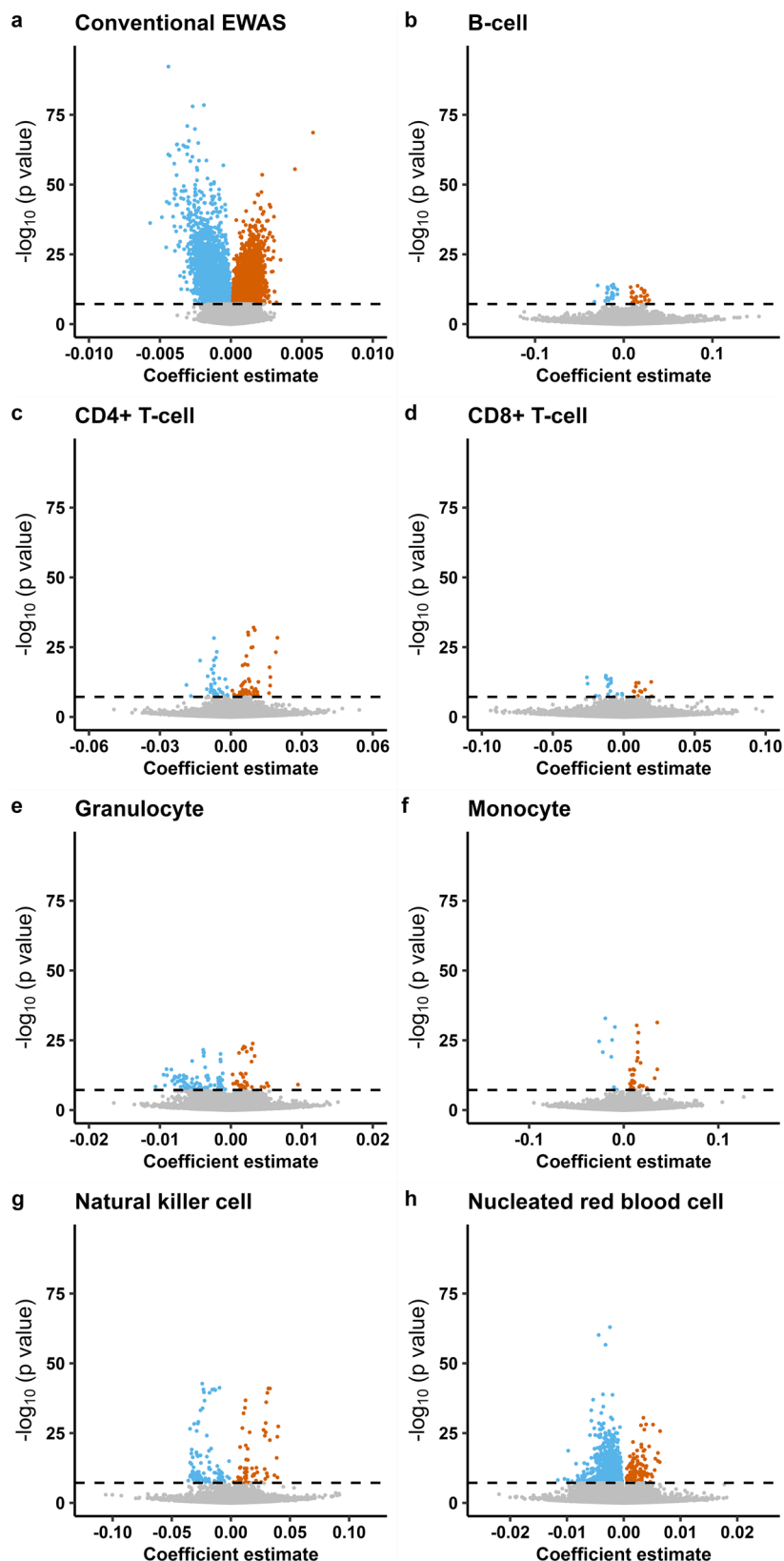


Fig. 3 Volcano plots of the epigenome-wide DNAm associated with GA in START ($n = 953$). **a** Results from the conventional EWAS in which we adjusted for estimated cell-type proportions (see Methods for details of the statistical model). **b-h** Results for each of the seven cell types from the cell-type specific analysis using CellDMC. Gray dots indicate nonsignificant associations and colored dots indicate those that are Bonferroni-significant ($p_B < 0.05$). Blue-colored dots show CpGs with a negative effect size and orange dots show CpGs with a positive effect size. The x-axis represents coefficient estimates (β -values) for the DNAm-GA association, and the y-axis the corresponding $-\log_{10} p$ values. The horizontal dashed line denotes the Bonferroni-corrected genome-wide significance threshold ($p_B < 0.05$).

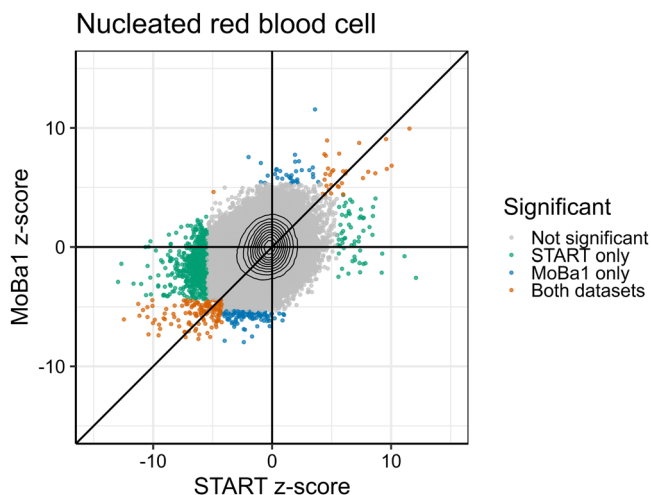


Fig. 4 Comparison of nRBC-specific CpGs associated with GA in the EPIC-based START dataset ($n = 953$) and the 450k-based MoBa1 dataset ($n = 1062$). Gray dots indicate nonsignificant CpGs, blue dots CpGs significantly associated only in MoBa1 ($p_B < 0.05$), green dots CpGs significantly associated only in START ($p_B < 0.05$), and orange dots CpGs significantly associated in both datasets ($r < 0.05$). Black isolines indicate the density of the points, increasing towards the crossing point of the axes. The x and y axes represent z-scores (i.e., the coefficient estimate divided by the standard error) for START and MoBa1, respectively.

START and MoBa1 for the remaining six cell types ($r < 0.05$, Supplementary Fig. 5).

Validation with a different cell-type specific method. To further validate the cell-type specific associations between DNAm and GA, we applied TCA to the START dataset using two different approaches. First, we applied a one-stage implementation of TCA which runs marginal conditional tests for each cell type, analogous to CellDMC. We then applied a two-stage implementation of TCA, by first extracting the cell-type tensors additionally adjusted for the above-mentioned covariates and then performing separate EWAS regressions on each tensor with respect to GA. With the one-stage approach, we identified 979 GA-associated CpGs ($p_B < 0.05$), whereas with the two-stage approach, we identified 4714 GA-associated CpGs ($p_B < 0.05$). Both approaches map most of the cell-type specific significant CpGs to nRBCs [$n = 836$ (85%) in the one-stage approach (Supplementary Fig. 6) and $n = 3130$ (66%) in the two-stage approach (Supplementary Fig. 7)]. For all cell types, more CpGs were statistically significant using the two-stage approach compared to the one-stage approach. In granulocytes specifically, 1668 CpGs were identified as significantly associated with GA, of which 829 were also mapped to nRBCs. The results from the one-stage and two-stage TCA analyses can be found in Supplementary Data 6 and 7, respectively.

Among the 2030 nRBC-specific CpGs detected by CellDMC, 623 CpGs were also detected when applying the one-stage TCA (Supplementary Fig. 8). Overall, 260 nRBC-specific CpGs were detected by both CellDMC and the two-stage TCA approach (Supplementary Fig. 9). The results from the one-stage TCA analysis were also generally consistent with those of the CellDMC analysis for the other six cell types (Supplementary Fig. 8), while the two-stage TCA results showed more divergent associations for the other cell types (Supplementary Fig. 9).

Location of GA-associated CpGs. We scrutinized the GA-associated CpGs identified by the conventional EWAS and

CellDMC analyses according to their location in the genome (Fig. 5 and Supplementary Data 2 and 3). The 2030 nRBC-specific CpGs that were significantly associated with GA in START were predominantly localized to gene bodies (48% of the nRBC-specific CpGs versus 30% of all CpGs on EPIC, $p = 2.5 \times 10^{-67}$, Fig. 5a), open sea (75% versus 56%, $p = 2.2 \times 10^{-69}$, Fig. 5b), and CpG island shelves (8.2% versus 7.1%, $p = 0.023$, Fig. 5b). Markedly fewer nRBC-specific CpGs were in promoter regions (22 versus 38%, $p = 2.8 \times 10^{-55}$, Fig. 5a), shores (12 versus 18%, $p = 1.0 \times 10^{-12}$, Fig. 5b), and CpG islands (4.7% versus 19%, $p = 5.3 \times 10^{-77}$, Fig. 5b). We discovered a similar pattern of CpG localization in the nRBC-specific MoBa1 results. The corresponding patterns for the other cell types showed more variation between the two datasets (Supplementary Fig. 10), which may be due to a substantially lower number of CpGs in each category.

Gene annotation and enrichment analysis of nRBC-specific CpGs associated with GA. We used the online Genomic Regions Enrichment of Annotations Tool (GREAT)³² to examine whether the 2030 GA-associated CpGs for nRBC were located near or within any gene of known pathway annotation. 2836 genes were identified using this approach (Supplementary Data 8), 198 of which were associated with more than three differentially methylated CpGs. A foreground/background hypergeometric test was performed on the 2030 GA-associated nRBC-specific CpGs. The results of this test revealed four clusters of Gene Ontology (GO) biological processes significantly enriched in our data (Supplementary Data 9). These processes were related to (i) response to corticosteroid (75 CpGs/55 genes, $p_B = 0.0001$), (ii) response to purine-containing compound (65 CpGs/45 genes, $p_B = 0.002$), (iii) granulocyte migration (34 CpGs/23 genes, $p_B = 0.006$), and (iv) stress-activated protein kinase signaling cascade (58 CpGs/32 genes, $p_B = 0.01$). When the analyses were restricted to only those CpGs that are present on both 450k and EPIC, we did not find any significantly enriched biological pathways.

Discussion

Although epigenome-wide associations between GA and DNAm in cord blood are now well established, little is known about the contribution of different cell types and the biological mechanisms underlying these associations. In this study, we explored the association between GA and DNAm using data from two types of DNAm arrays (EPIC and 450k) and conducted both a conventional EWAS as well an investigation of cell-type specific associations. We found that most of the cell-type-specific associations between DNAm and GA were restricted to nRBCs. These results were robust across different datasets, DNAm arrays, and analysis methods. Our results point to a strong link between red blood cell development (erythropoiesis) in fetal life and fetal growth as measured by GA, providing critical insights and implications for further studies on the relationship between DNAm and GA.

In the conventional EWAS, we identified 13,660 CpGs linked to 8669 genes as being differentially methylated with GA. Slightly more of the significant CpGs were specific for the EPIC array (56%), despite only 48% of the CpGs being EPIC-specific. Bohlin et al.¹⁰ previously applied a similar model to the MoBa1 dataset and identified 5474 CpGs associated with GA. 2556 of the CpGs and 1741 of the genes identified in that study overlap with our results in the START EPIC-based dataset. We also compared our results to the “all births model” from a recent meta-analysis by Merid et al.⁵ where the authors investigated GA and DNAm measured on 450k in cord-blood DNA from 6885 newborns in 20 different cohorts. The authors identified 17,095 CpGs

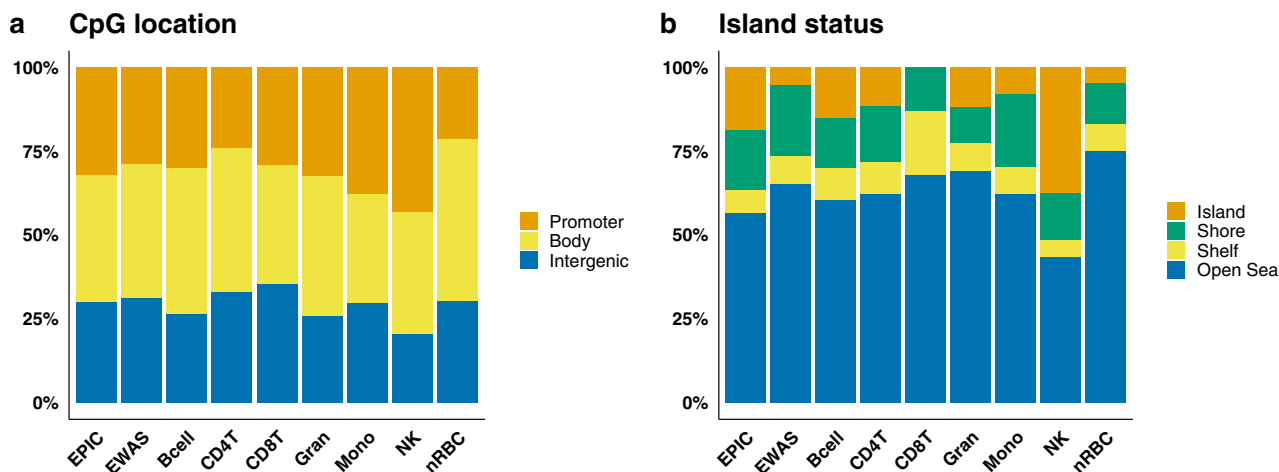


Fig. 5 Position enrichment results of CpGs associated with GA compared to all CpGs on the EPIC array. Position enrichment results of all the CpGs on the EPIC array ($n = 770,586$; denoted as EPIC on the x-axis), those specifically associated with GA in the conventional EWAS ($n = 13,660$; EWAS), and each cell type in the CellDMC analyses in START (Bcell, $n = 53$; CD4T, $n = 103$; CD8T, $n = 31$; Gran, $n = 136$; Mono, $n = 37$; NK, $n = 157$; nRBC, $n = 2030$). **a** The proportion of CpGs in the promoter (orange), gene body (yellow), and intergenic (blue) regions. **b** The proportion of CpGs in CpG islands (orange), shores (green), shelves (yellow), and open sea (blue). Bcell B-cell, CD4T CD4 + T-cell, CD8T CD8 + T-cell, Gran granulocyte, Mono monocyte, NK natural killer cell, nRBC nucleated red blood cell.

significantly associated with GA, of which 4688 CpGs and 4437 genes overlap with our results. Of note, MoBa1 and yet another MoBa-based dataset (MoBa2) were also included in the meta-analysis by Merid et al. Nevertheless, these comparisons show that the results from our conventional EWAS model are concordant with those of previous studies on DNAm and GA.

As a primary step to explore cell-type specific changes in DNAm with GA, we used the interaction-based algorithm CellDMC that has been validated in several EWAS datasets and data in which the actual cell-type composition is known^{24,33,34}. We identified 2330 differentially methylated CpGs associated with GA, with an overwhelming number of the significant CpGs confined to nRBCs (2030 CpGs linked to 2836 genes). This is particularly striking given that nRBCs are not the dominant cell type in terms of variation and abundance. Taken together, these findings strongly suggest that DNAm changes in nRBCs are responsible for the observed DNAm–GA association.

It is nevertheless important to account for the limited sensitivity of CellDMC when including seven different cell types in the analysis³³. To assess this limited sensitivity and verify that the nRBC-specific results were not an array-based artifact, we repeated the CellDMC analyses in MoBa1, which is a 450k-based dataset stemming from the same source population as the START dataset (MoBa). We observed a similar pattern of cell-type-specific association with GA as with the START dataset, although there were fewer significant CpGs in the MoBa1 dataset. Moreover, 174 nRBC-specific CpGs were significantly associated with GA in both datasets, as opposed to no such overlap in CpGs across the other six cell types. One option to further increase the power of the CellDMC analysis would have been to merge the two datasets over the common set of 450k CpGs. Even though this would have increased the sample size substantially, such an approach has several major drawbacks. First, one would lose the much greater coverage of the EPIC array and possibly miss important associations between GA and CpGs that are only detectable using EPIC-derived DNAm data. Second, merging the datasets would introduce a new batch variable that would need to be accounted for in the model. We thus opted to keep the analyses of the two datasets separate.

To further validate our results, we applied another method for cell-type specific analysis, TCA, to the START data. TCA utilizes

a statistical framework based on matrix factorization²³. The results from both the one-stage and two-stage applications of TCA showed a similar pattern of cell-type specific association with GA as observed with CellDMC. Our findings are also consistent with a previous study on nRBCs pointing to extensive DNAm changes in nRBCs between preterm and term newborns³⁵. In that study, the authors identified 9258 differentially methylated sites when comparing nRBCs from preterm and term newborns. These sites were predominantly hypomethylated and enriched in gene body and intergenic regions³⁵. Taken together, these results strengthen the interpretation that nRBCs are the primary cell type driving the association between DNAm and GA in cord blood.

nRBCs are an integral part of erythropoiesis, the process by which mature red blood cells (erythrocytes) are produced in adult and fetal bone marrow, fetal liver, and the embryonic yolk sac. Erythropoiesis is crucial for embryonic and fetal growth. During the third trimester of pregnancy, the production of erythrocytes is approximately three to five times that of the adult steady-state levels³⁶. Although nRBCs circulate in the fetal bloodstream throughout pregnancy, they stay in circulation for only a few days after birth³⁷. Several genes annotated to the nRBC-specific CpGs that we found to be associated with GA are implicated in a wide array of biological processes involved in erythropoiesis. A subset of the genes related to these processes are described in more detail in Supplementary Data 10. Briefly, these processes include cell-cycle progression and cytokinesis^{38,39}, chromatin condensation^{39,40}, hemoglobin synthesis³⁸, mitochondrial function and iron metabolism^{38,41,42}, degradation of proteins and organelles^{34,43}, erythroblastic island formation⁴⁴, and enucleation^{39,40}. Moreover, several of the genes are essential for the switch from fetal to adult hemoglobin, which occurs shortly after birth⁴⁵. Taken together, our findings provide strong support for fetal erythropoiesis representing an important biological mechanism underlying the association between DNAm and GA.

To learn more about the mechanisms contributing to the nRBC-specific association between DNAm and GA, we searched for the enrichment of specific biological pathways in the set of nRBC-specific CpGs. One of the main clusters of biological pathways was the response to corticosteroids, and more specifically, the response to glucocorticoids. Glucocorticoids are a class

of corticosteroids that are essential for a wide variety of biological processes, including proliferation, differentiation, and apoptosis of many cell types in response to stress. They also play a pivotal role in pregnancy and normal fetal development⁴⁶, even though prenatal overexposure to glucocorticoids has also been reported to be detrimental to fetal growth and postnatal physiology^{47,48}. Glucocorticoids are known regulators of erythroid progenitors^{49,50}, and the glucocorticoid receptor encoded by *NR3C1* controls several processes involved in erythropoiesis^{51–53}. In particular, the glucocorticoid receptor controls erythroid response to stress^{54–56}. Stress, such as hypoxia, leads to the glucocorticoid receptor-dependent activation of the *BMP4*-dependent stress erythropoiesis pathway, in which many new erythrocytes are generated to maintain homeostasis⁵⁷. Interestingly, stress erythropoiesis shares several similarities with fetal erythropoiesis⁵⁸.

The link between erythropoiesis and GA is not unprecedented. Several of the genes found to be relevant for erythropoiesis in our data have previously been identified in other studies of GA. A few examples include *NCOR2*^{5,10,59,60}, *HDAC4*^{5,10,60}, *CASP8*^{5,10,60,61}, and *RAPGEF2*^{5,60,62}. The nuclear receptor co-repressor encoded by *NCOR2* interacts with the transcription factor *BCL11A* in regulating the expression of fetal hemoglobin⁶³. *NCOR2* also promotes chromatin condensation, which is a crucial step during terminal erythropoiesis. Histone deacetylase 4 (*HDAC4*) also plays a key role in chromatin condensation and associates directly with the key erythroid transcription factor *GATA1*⁶⁴. *CASP8* encodes the protease Caspase 8, which is a key activator of effector caspases required for terminal erythroid differentiation⁶⁵. Finally, *RAPGEF2* encodes a guanine nucleotide exchange factor known to play an important role in embryonic hematopoiesis⁶⁶.

The results of our study, as well as those of others described above, strongly suggest that DNAm patterns related to erythropoiesis are at least partly responsible for the observed association between DNAm and GA. Our findings of predominantly hypomethylated nRBC-specific CpGs are in line with previous studies showing progressive global DNA hypomethylation involved in erythroid lineage commitment and differentiation as well as chromatin condensation and enucleation of nRBCs during erythropoiesis^{67,68}. Other studies have consistently shown a higher proportion of hypomethylated CpGs amongst those associated with GA^{5,10,59,61}.

Further, the findings that nRBCs are the primary drivers behind the association between DNAm and GA may help explain the poor correlation observed between epigenetic clocks for newborn GA and those for chronological age in adults^{10,11}. Indeed, GA-related changes in cord blood DNAm do not persist through childhood and adolescence, as shown in a longitudinal analysis of DNAm associated with GA⁵⁹ and a meta-analysis of several EWASs of GA⁵. This could be due to the rapid loss of nRBCs with increasing GA and its subsequent disappearance from the bloodstream of healthy newborns within the first few days after birth. In other words, the disappearance of nRBCs shortly after birth implies that the main driver behind the GA-related changes in cord blood DNAm also disappears. Moreover, the association between GA and specific DNAm changes in nRBCs, as demonstrated by our study, may also help explain why GA acceleration (GAA, defined as the discrepancy between GA predicted from DNAm data and GA determined by clinical measurements) has been linked to several adverse outcomes^{11,69,70}. In this regard, it is interesting to note that increased nRBC counts at birth are associated with a higher risk of mortality and adverse neonatal outcomes and have been suggested as a predictive marker for perinatal hypoxia, intrauterine growth restriction, and preeclampsia^{71–75}. Further studies are needed to determine if GAA is indeed related to these or other

adverse outcomes, and if differences in nRBCs may be driving these associations.

The results of our study may have important clinical implications. For instance, fetal nRBCs are routinely isolated from the mother's peripheral blood during pregnancy for prenatal diagnostics, and several experimental approaches are available for the rapid isolation of nRBCs^{76,77}. Our findings may help pave the way for the development of DNAm-based GA prediction during pregnancy based on nRBC-specific assays, which may provide a more targeted assessment of fetal growth and prenatal development.

One important limitation of our study is the use of *in silico* estimations of cell-type proportions. Although we have used a reference-based method with validated cord blood-specific reference data, it is important to bear in mind that the proportions we have used here are only estimates. In addition, since the cell-type proportions are essentially fractions that sum up to one, they are not independent of each other, and the correlation between them may impact our analyses. However, since our results were robust despite the use of different DNAm arrays, datasets, and methods, our findings are unlikely to be severely affected by these limitations.

In conclusion, the results of our study strongly indicate that nRBCs are the primary drivers behind the observed DNAm–GA association. Importantly, an epigenetic signature of erythropoiesis seems to be partly responsible for this association, providing a biologically compelling mechanism that links GA, DNAm, and nRBCs. Furthermore, our findings provide an explanation for the poor correlation observed between epigenetic clocks for newborn GA and those for chronological age in adults, contributing important mechanistic insights into the epigenetic regulation of fetal growth and development.

Methods

Study population. MoBa is a population-based pregnancy cohort study in which ~114,500 newborns, 95,200 mothers, and 75,200 fathers were recruited from all over Norway from 1999 to 2008²⁸. The mothers consented to participation in 41% of the pregnancies. The study participants have been followed at different time points via self-administered questionnaires and linkage to the Medical Birth Registry of Norway (MBRN). Further details on MoBa have been provided in our previous publications^{28,78}.

For this study specifically, we used two non-overlapping subsamples: (i) the Study of Assisted Reproductive Technology (START; $n = 953$ newborns) and (ii) MoBa1 ($n = 1062$ newborns). Both datasets are based on cord blood samples from the same source population (MoBa). However, they differ in the methylation array used to generate the DNAm data: START used EPIC whereas MoBa1 used 450k (see below for details). An overview of the sample selection and analysis flow is provided in Supplementary Fig. 11. Detailed characteristics and eligibility criteria for the START and MoBa1 datasets have been provided in our previous work^{29,79}.

Sample processing, DNAm measurement, and quality control. The sample processing, DNAm measurement, and quality control pipeline used for data cleaning have been extensively detailed in our previous works^{29,79}. Briefly, cord blood samples taken by a midwife immediately after birth were frozen. For the START dataset, DNAm was measured at 885,000 CpG sites using the Illumina Infinium MethylationEPIC BeadChip (Illumina, San Diego, USA). The raw iDAT files were processed in four batches using the R package *RnBeads*⁸⁰. Cross-hybridizing probes⁸¹ and probes that had a detection p value above 0.01 were removed using the *greedy* algorithm in *RnBeads*. We also excluded probes in which the last three bases overlapped with a single-nucleotide polymorphism (SNP). The remaining DNAm signal was processed using *BMIQ*⁸² to normalize the type I and type II probe chemistries⁸³. The *RnBeads* output of control probes were visually inspected for all samples, and those with low overall signals were removed. The *greedy* algorithm was used to remove outliers with markedly different DNAm signals than the rest of the samples, resulting in the removal of 58 samples in total. For consistency, CpG sites excluded from one batch due to poor quality and low detection p value were also removed from all subsequent batches.

For the MoBa1 samples, DNAm was measured at 485,577 CpG sites using the Illumina Infinium HumanMethylation450 BeadChip (Illumina, San Diego, USA). Arrays not fulfilling the 5% detection p value were removed together with all duplicates. Within-array normalization was carried out using *BMIQ* from the *wateRmelon* R package⁸⁴.

Variables. Information on GA, newborn sex and birth weight, maternal age, parity, and whether the birth was induced was extracted from MBRN. GA at birth was estimated by ultrasound measurements around week 18 of pregnancy. Since newborn sex may occasionally be incorrectly recorded in MBRN, we inferred sex from the DNAm data. As a result, one female was reclassified as male, and five males were reclassified as females. Information on maternal smoking was derived from the MoBa questionnaires and was included as a four-level categorical variable: (i) no smoking before or during pregnancy; (ii) smoked, but quit before pregnancy; (iii) smoked, but quit early in pregnancy; and (iv) continued smoking during pregnancy.

Estimation of cell-type proportions. To estimate cell-type proportions in our samples, we used the filtered and combined reference dataset “FlowSorted.Cord-BloodCombined.450k” from ref. 19, which specifies seven main cell types in cord blood (B-cells, CD4 + T-cells, CD8 + T-cells, granulocytes, monocytes, natural killer cells, and nRBCs). We used the `estimateCellCounts2` function in the `FlowSorted.Blood.EPIC` R package⁸⁵ and the Identifying Optimal Libraries (IDOL) probe selection to perform cellular deconvolution and noob preprocessing.

Statistics and reproducibility. After quality control, the sample available for the current analyses in the START dataset consisted of 770,586 autosomal CpGs and 953 newborns conceived naturally and for whom we had information on ultrasound-based GA (Supplementary Fig. 9). For the MoBa1 dataset, the sample available for the current analyses comprised 473,731 autosomal CpGs and 1062 newborns with information on ultrasound-based GA (Supplementary Fig. 9).

Principal component analysis (PCA) of estimated cell-type proportions was conducted using the `prcomp` R function. The R package `robustbase`⁸⁶ for MM-type robust regression was used to assess the relationship between cell-type composition and GA. Bonferroni correction was applied to the results from the conventional EWAS and cell-type-specific models to control for multiple testing. A Bonferroni p value (p_B) <0.05 was declared statistically significant.

Analyses in START. In the conventional EWAS model, we screened for associations between DNAm in cord blood and GA at birth by applying a linear mixed-effect model to each of the 770,586 CpG sites remaining after quality control. The β -values of the individual CpGs were used as the response (dependent) variables and GA was used as the explanatory (independent) variable, with adjustments made for newborn sex, maternal age, maternal smoking, cell-type proportions, and array plate in the regression model.

To assess interactions between cell-type specific DNAm and GA, we performed epigenome-wide analyses using the CellDMC framework as outlined in ref. 24 and the corresponding `CellDMC` function in the EpiDISH R package. Briefly, CellDMC runs a linear model similar to that used in our conventional EWAS, but it also includes an interaction term to inform the model whether there is a significant interaction between the exposure and the corresponding fraction of each specific cell type. Estimates of the regression coefficients and p values are calculated for each cell type using least squares. As with the conventional EWAS, newborn sex, maternal age, maternal smoking, and plate were also included as covariates in the CellDMC model. Bonferroni correction was applied to all the results from the conventional EWAS and CellDMC models to control for multiple testing. As before, a Bonferroni p value (p_B) <0.05 was declared statistically significant.

Besides CellDMC, we also applied the TCA framework developed by ref. 23 to detect cell-type specific DNAm–GA associations. In contrast to CellDMC, TCA is based on the concept of matrix factorization. Specifically, TCA uses the DNAm measurements from the mixed samples along with information on cell-type proportions (in our case, the ones that are estimated) for each individual and calculates a three-dimensional tensor of DNAm values for each cell type in each individual. The TCA framework further allows a search for statistical associations between cell-type specific signals and an outcome or exposure of interest. We used two different approaches for TCA based on the available functions in the TCA package²³. First, we applied a one-stage approach using the `tca` function, which fits a model for all cell types jointly and tests the effect of each cell type separately for statistical significance. We included the same covariates in the TCA model as in the CellDMC and conventional EWAS models (newborn sex, maternal age, maternal smoking, and array plate). Additionally, we applied a two-stage approach, where a tensor for each cell type is first inferred and then an EWAS of GA is conducted for each tensor. This was carried out by first using the `tca` function to fit a model including all covariates mentioned above except GA. The model resulting from the `tca` function was subsequently added as input for the `tensor` function, obtaining new DNAm tensors for each cell type. An EWAS of GA was then performed for each cell-type-specific tensor.

Analyses in MoBa1. To test whether array type had an impact on the findings obtained from the analysis of the EPIC-based START dataset, we re-ran the CellDMC analysis on the 450k-based MoBa1 dataset, testing all the 473,731 CpGs available in this dataset.

To compare the CellDMC results from MoBa1 with those from START, we applied the r value approach suggested by ref. 31, which allows a rigorous assessment of the replication of findings. In short, we tested each CpG for

association with GA in both datasets (MoBa1 and START) and computed an r value (the lowest FDR level at which the finding was replicated). We chose the r value approach over other approaches, such as those used in a standard meta-analysis or a two-step replication study, for the following reasons. First, a meta-analysis tests whether there is any signal across the two studies; however, it does not test whether the two studies show appropriate significance. Second, assessing replicability in a two-step replication study is not straightforward, as this requires adequate control of the type I error in both studies. This may involve a different number of tests, especially as we use two types of DNAm arrays (EPIC and 450k). Thus, the approach of ref. 31 provides a simpler solution for assessing replicability and for controlling the type I error.

Location of CpGs. Information on CpG location and regulatory regions was extracted from the respective Illumina Manifest Files (Infinium MethylationEPIC v1.0 B4 for START and HumanMethylation450 v1.2 for MoBa1). One-tailed hypergeometric tests were conducted to assess the relative enrichment of CpGs in specific regions of interest.

Gene annotation and enrichment analysis. CpGs were annotated using the online Genomic Regions Enrichment of Annotations Tool (GREAT³²) using the human genome build hg19 (GRCh37). GREAT was selected amongst other competing methods because it considers both proximal (5.0 kb upstream and 1.0 kb downstream) and distal (up to 1000 kb) regulatory regions. This is an advantage over other methods that only take proximal regions into account, because taking distal regulatory regions into account enables an assessment of the extra information gained from detecting DNAm on distal regulatory CpGs on the EPIC array. For gene enrichment analysis, GREAT performed a foreground/background hypergeometric test over genomic regions using the total number of CpGs surviving quality control as background (770,586 CpGs for the EPIC analyses and 473,731 CpGs for the 450k analyses). Finally, GREAT extracts information from Gene Ontology (GO) and other ontologies covering human and mouse phenotypes³².

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Access to the START and MoBa1 DNAm datasets can be obtained by applying to the Norwegian Institute of Public Health (NIPH). Restrictions apply regarding the availability of these data, which were originally used under specific approvals for the current study and are therefore not publicly available. Access can only be given after approval by the Norwegian Regional Committees for Medical and Health Research Ethics (REK) under the provision that the applications are consistent with the consent provided. An application form can be found on the NIPH website at <https://www.fhi.no/en/studies/moba/>. Specific questions regarding access to data in this study can also be directed to Dr. Siri E. Håberg (Siri.Haberg@fhi.no). The data generated in this study are provided as Supplementary Data.

Code availability

All statistical analyses were performed using R version 4.1.2⁸⁷. R scripts are available from the authors upon request.

Received: 3 July 2022; Accepted: 13 February 2023;

Published online: 27 February 2023

References

- Ghartey, K. et al. Neonatal respiratory morbidity in the early term delivery. *Am. J. Obstet. Gynecol.* **207**, 292.e291–294 (2012).
- Knight, A. K., Conneely, K. N. & Smith, A. K. Gestational age predicted by DNA methylation: potential clinical and research utility. *Epigenomics* <https://doi.org/10.2217/epi-2016-0157> (2017).
- Raby, B. A. et al. Low-normal gestational age as a predictor of asthma at 6 years of age. *Pediatrics* **114**, e327–e332 (2004).
- Yang, S., Bergvall, N., Cnattingius, S. & Kramer, M. S. Gestational age differences in health and development among young Swedish men born at term. *Int. J. Epidemiol.* **39**, 1240–1249 (2010).
- Merid, S. K. et al. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. *Genome Med.* **12**, 25 (2020).
- Guo, H. et al. The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).

7. Lim, Y. C. et al. A complex association between DNA methylation and gene expression in human placenta at first and third trimesters. *PLoS ONE* **12**, e0181155 (2017).
8. Sliker, R. C. et al. DNA methylation landscapes of human fetal development. *PLoS Genet.* **11**, e1005583 (2015).
9. Akhbari, L. et al. DNA methylation changes in cord blood and the developmental origins of health and disease - a systematic review and replication study. *BMC Genomics* **23**, 221 (2022).
10. Bohlin, J. et al. Prediction of gestational age based on genome-wide differentially methylated regions. *Genome Biol.* **17**, 207 (2016).
11. Knight, A. K. et al. An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biol.* **17**, 206 (2016).
12. York, T. P. et al. Replicated umbilical cord blood DNA methylation loci associated with gestational age at birth. *Epigenetics* **15**, 1243–1258 (2020).
13. Pidsley, R. et al. Critical evaluation of the illumina methylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
14. Sandoval, J. et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).
15. Reinius, L. E. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS ONE* **7**, e41361 (2012).
16. Teschendorff, A. E. & Zheng, S. C. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics* **9**, 757–768 (2017).
17. Titus, A. J., Gallimore, R. M., Salas, L. A. & Christensen, B. C. Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum. Mol. Genet.* **26**, R216–r224 (2017).
18. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinforma.* **13**, 86 (2012).
19. Gervin, K. et al. Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. *Clin. Epigenetics* **11**, 125 (2019).
20. Teschendorff, A. E., Zhu, T., Breeze, C. E. & Beck, S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol.* **21**, 221 (2020).
21. Li, Z., Wu, Z., Jin, P. & Wu, H. Dissecting differential signals in high-throughput data from complex tissues. *Bioinforma.* **35**, 3898–3905 (2019).
22. Luo, X., Yang, C. & Wei, Y. Detection of cell-type-specific risk-CpG sites in epigenome-wide association studies. *Nat. Commun.* **10**, 3113 (2019).
23. Rahmani, E. et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nat. Commun.* **10**, 3417 (2019).
24. Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat. Methods* **15**, 1059–1066 (2018).
25. Bauer, M. Cell-type-specific disturbance of DNA methylation pattern: a chance to get more benefit from and to minimize cohorts for epigenome-wide association studies. *Int. J. Epidemiol.* **47**, 917–927 (2018).
26. Glasser, L., Sutton, N., Schmeling, M. & Machan, J. T. A comprehensive study of umbilical cord blood cell developmental changes and reference ranges by gestation, gender and mode of delivery. *J. Perinatol.* **35**, 469–475 (2015).
27. Pérez, A., Gurbindo, M. D., Resino, S., Aguarón, A. & Muñoz-Fernández, M. A. NK cell increase in neonates from the preterm to the full-term period of gestation. *Neonatology* **92**, 158–163 (2007).
28. Magnus, P. et al. Cohort profile update: the Norwegian Mother and Child Cohort Study (MoBa). *Int. J. Epidemiol.* **45**, 382–388 (2016).
29. Häberg, S. E. et al. DNA methylation in newborns conceived by assisted reproductive technology. *Nat. Commun.* **13**, 1896 (2022).
30. Haftorn, K. L. et al. An EPIC predictor of gestational age and its application to newborns conceived by assisted reproductive technologies. *Clin. Epigenetics* **13**, 82 (2021).
31. Heller, R., Bogomolov, M. & Benjamini, Y. Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proc. Natl Acad. Sci. USA* **111**, 16262–16267 (2014).
32. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
33. You, C. et al. A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. *Nat. Commun.* **11**, 4779 (2020).
34. Lu, T. et al. Detecting cord blood cell type-specific epigenetic associations with gestational diabetes mellitus and early childhood growth. *Clin. Epigenetics* **13**, 131 (2021).
35. de Goede, O. M., Lavoie, P. M. & Robinson, W. P. Cord blood hematopoietic cells from preterm infants display altered DNA methylation patterns. *Clin. Epigenetics* **9**, 39 (2017).
36. Palis, J. & Segel, G. B. Developmental biology of erythropoiesis. *Blood Rev.* **12**, 106–114 (1998).
37. Hermansen, M. C. Nucleated red blood cells in the fetus and newborn. *Arch. Dis. Child Fetal Neonatal Ed.* **84**, F211–F215 (2001).
38. Sen, T., Chen, J. & Singbrant, S. Decreased PGC1 β expression results in disrupted human erythroid differentiation, impaired hemoglobinization and cell cycle exit. *Sci. Rep.* **11**, 17129 (2021).
39. Mei, Y., Liu, Y. & Ji, P. Understanding terminal erythropoiesis: An update on chromatin condensation, enucleation, and reticulocyte maturation. *Blood Rev.* **46**, 100740 (2021).
40. Menon, V. & Ghaffari, S. Erythroid enucleation: a gateway into a “bloody” world. *Exp. Hematol.* **95**, 13–22 (2021).
41. Chen, K. et al. Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis. *Proc. Natl Acad. Sci. USA* **106**, 17413–17418 (2009).
42. Liang, R. et al. A systems approach identifies essential FOXO3 functions at key steps of terminal erythropoiesis. *PLoS Genet.* **11**, e1005526 (2015).
43. Mathangasinghe, Y., Fauvet, B., Jane, S. M., Goloubinoff, P. & Nillegoda, N. B. The Hsp70 chaperone system: distinct roles in erythrocyte formation and maintenance. *Haematologica* **106**, 1519–1534 (2021).
44. Chasis, J. A. & Mohandas, N. Erythroblastic islands: niches for erythropoiesis. *Blood* **112**, 470–478 (2008).
45. Sankaran, V. G. & Orkin, S. H. The switch from fetal to adult hemoglobin. *Cold Spring Harb. Perspect. Med.* **3**, a011643 (2013).
46. Solano, M. E. & Arck, P. C. Steroids, pregnancy and fetal development. *Front. Immunol.* **10**, 3017 (2019).
47. Moisiadis, V. G. & Matthews, S. G. Glucocorticoids and fetal programming part 1: outcomes. *Nat. Rev. Endocrinol.* **10**, 391–402 (2014).
48. Tang, J. I., Seckl, J. R. & Nyirenda, M. J. Prenatal glucocorticoid overexposure causes permanent increases in renal erythropoietin expression and red blood cell mass in the rat offspring. *Endocrinology* **152**, 2716–2721 (2011).
49. Flygare, J., Rayon Estrada, V., Shin, C., Gupta, S. & Lodish, H. F. HIF1 α synergizes with glucocorticoids to promote BFU-E progenitor self-renewal. *Blood* **117**, 3435–3444 (2011).
50. von Lindern, M. et al. The glucocorticoid receptor cooperates with the erythropoietin receptor and c-Kit to enhance and sustain proliferation of erythroid progenitors in vitro. *Blood* **94**, 550–559 (1999).
51. Lee, H. Y. et al. PPAR- α and glucocorticoid receptor synergize to promote erythroid progenitor self-renewal. *Nature* **522**, 474–477 (2015).
52. Nicolaides, N. C., Galata, Z., Kino, T., Chrousos, G. P. & Charmandari, E. The human glucocorticoid receptor: molecular basis of biologic function. *Steroids* **75**, 1–12 (2010).
53. Stellacci, E. et al. Interaction between the glucocorticoid and erythropoietin receptors in human erythroid cells. *Exp. Hematol.* **37**, 559–572 (2009).
54. Bauer, A. et al. The glucocorticoid receptor is required for stress erythropoiesis. *Genes Dev.* **13**, 2996–3002 (1999).
55. Dolznig, H. et al. Erythroid progenitor renewal versus differentiation: genetic evidence for cell autonomous, essential functions of EpoR, Stat5 and the GR. *Oncogene* **25**, 2890–2900 (2006).
56. Leonard, M. O., Godson, C., Brady, H. R. & Taylor, C. T. Potentiation of glucocorticoid activity in hypoxia through induction of the glucocorticoid receptor. *J. Immunol.* **174**, 2250–2257 (2005).
57. Paulson, R. F., Hariharan, S. & Little, J. A. Stress erythropoiesis: definitions and models for its study. *Exp. Hematol.* **89**, 43–54.e42 (2020).
58. Porayette, P. & Paulson, R. F. BMP4/Smad5 dependent stress erythropoiesis is required for the expansion of erythroid progenitors during fetal development. *Dev. Biol.* **317**, 24–35 (2008).
59. Simpkin, A. J. et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum. Mol. Genet.* **24**, 3752–3763 (2015).
60. Paret, S. E. et al. Fetal DNA methylation associates with early spontaneous preterm birth and gestational age. *PLoS ONE* **8**, e67489 (2013).
61. Schroeder, J. W. et al. Neonatal DNA methylation patterns associate with gestational age. *Epigenetics* **6**, 1498–1504 (2011).
62. Lee, H. et al. DNA methylation shows genome-wide association of NFIX, RAPGEF2 and MSRB3 with gestational age at birth. *Int. J. Epidemiol.* **41**, 188–199 (2012).
63. Xu, J. et al. Corepressor-dependent silencing of fetal hemoglobin expression by BCL11A. *Proc. Natl Acad. Sci. USA* **110**, 6518–6523 (2013).
64. Varricchio, L. et al. Identification of NuRSERY, a new functional HDAC complex composed by HDAC5, GATA1, EKLF and pERK present in human erythroid cells. *Int. J. Biochem. Cell Biol.* **50**, 112–122 (2014).
65. Zermati, Y. et al. Caspase activation is required for terminal erythroid differentiation. *J. Exp. Med.* **193**, 247–254 (2001).
66. Satyanarayana, A. et al. RapGEF2 is essential for embryonic hematopoiesis but dispensable for adult hematopoiesis. *Blood* **116**, 2921–2931 (2010).
67. Schulz, V. P. et al. A unique epigenomic landscape defines human erythropoiesis. *Cell Rep.* **28**, 2996–3009.e2997 (2019).
68. Yu, Y. et al. High resolution methylome analysis reveals widespread functional hypomethylation during adult human erythropoiesis. *J. Biol. Chem.* **288**, 8805–8814 (2013).

69. Khouja, J. N. et al. Epigenetic gestational age acceleration: a prospective cohort study investigating associations with familial, sociodemographic and birth characteristics. *Clin. Epigenetics* **10**, 86 (2018).
70. Girchenko, P. et al. Associations between maternal risk factors of adverse pregnancy and birth outcomes and the offspring epigenetic clock of gestational age at birth. *Clin. Epigenetics* **9**, 49 (2017).
71. Davari-Tanha, F., Kaveh, M., Nemati, S., Javadian, P. & Salmanian, B. Nucleated red blood cells count in pregnancies with idiopathic intra-uterine growth restriction. *J. Fam. Reprod. Health* **8**, 77–81 (2014).
72. Gasparović, V. E., Ahmetasević, S. G. & Colić, A. Nucleated red blood cells count as first prognostic marker for adverse neonatal outcome in severe preeclamptic pregnancies. *Coll. Antropol.* **36**, 853–857 (2012).
73. Hebbbar, S., Misha, M. & Rai, L. Significance of maternal and cord blood nucleated red blood cell count in pregnancies complicated by preeclampsia. *J. Pregnancy* **2014**, 496416 (2014).
74. Morton, S. U., Brettin, K., Feldman, H. A. & Leeman, K. T. Association of nucleated red blood cell count with mortality among neonatal intensive care unit patients. *Pediatr. Neonatol.* **61**, 592–597 (2020).
75. Sokou, R. et al. Nucleated red blood cells: could they be indicator markers of illness severity for neonatal intensive care unit patients? *Children* <https://doi.org/10.3390/children7110197> (2020).
76. Byeon, Y., Ki, C. S. & Han, K. H. Isolation of nucleated red blood cells in maternal blood for Non-invasive prenatal diagnosis. *Biomed. Microdevices* **17**, 118 (2015).
77. Singh, R. et al. Fetal cells in maternal blood for prenatal diagnosis: a love story rekindled. *Biomark. Med.* **11**, 705–710 (2017).
78. Paltiel, L. et al. The biobank of the Norwegian Mother and Child Cohort Study – present status. *Norsk Epidemiologi* <https://doi.org/10.5324/nje.v24i1-2.1755> (2014).
79. Engel, S. M. et al. Neonatal genome-wide methylation patterns in relation to birth weight in the Norwegian Mother and Child Cohort. *Am. J. Epidemiol.* **179**, 834–842 (2014).
80. Müller, F. et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* **20**, 55 (2019).
81. McCartney, D. L. et al. Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationEPIC beadChip. *Genomics Data* **9**, 22–24 (2016).
82. Teschendorff, A. E. et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189–196 (2013).
83. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).
84. Pidsley, R. et al. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).
85. Salas, L. A. et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol.* **19**, 64 (2018).
86. Maechler, M. et al. robustbase: Basic Robust Statistics v. R package 0.93-6. <http://robustbase.r-forge.r-project.org/> (2020).
87. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2021).

Acknowledgements

We are grateful to the participants for their participation in the ongoing MoBa study. This work was funded by the Research Council of Norway, through its Centres of Excellence funding scheme, project number 262700, and by the National Institutes of Health (NIH) (grant R01 1HL134840-01).

Author contributions

K.L.H., J.B., and A.J. designed the research; K.L.H., W.R.P.D., Y.L., C.M.P., and J.R. conducted the analyses; K.L.H., W.R.P.D., Y.L., R.L., J.B., and A.J. interpreted the data; J.B. and A.J. supervised the study; K.L.H. and A.J. drafted the manuscript; P.M.M., S.E.H., and A.J. acquired funding, project administration, and resources. K.L.H., W.R.P.D., Y.L., C.M.P., J.R., R.L., Ø.E.N., D.K., P.M.M., S.E.H., J.B., and A.J. provided scientific input, revised the manuscript, and approved the final version.

Competing interests

All authors declare that they have no competing interests. In addition, the funding bodies did not play any role in the design of the study, collection, analysis, or interpretation of data, nor in writing the manuscript.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-023-04584-w>.

Correspondence and requests for materials should be addressed to Kristine L. Haftorn.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Manuel Breuer.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Paper 2

An EPIC predictor of gestational age and its application to newborns conceived by assisted reproductive technologies


2

RESEARCH

Open Access



An EPIC predictor of gestational age and its application to newborns conceived by assisted reproductive technologies

Kristine L. Haftorn^{1,2,3*} , Yunsung Lee^{1,2}, William R. P. Denault^{1,2,4}, Christian M. Page^{2,5}, Haakon E. Nustad^{2,6}, Robert Lyle^{2,7}, Håkon K. Gjessing^{2,4}, Anni Malmberg⁸, Maria C. Magnus^{2,9,10}, Øyvind Næss^{3,11}, Darina Czamara¹², Katri Räikkönen⁸, Jari Lahti⁸, Per Magnus², Siri E. Håberg², Astanand Jugessur^{1,2,4†} and Jon Bohlin^{2,13†}

Abstract

Background: Gestational age is a useful proxy for assessing developmental maturity, but correct estimation of gestational age is difficult using clinical measures. DNA methylation at birth has proven to be an accurate predictor of gestational age. Previous predictors of epigenetic gestational age were based on DNA methylation data from the Illumina HumanMethylation 27 K or 450 K array, which have subsequently been replaced by the Illumina MethylationEPIC 850 K array (EPIC). Our aims here were to build an epigenetic gestational age clock specific for the EPIC array and to evaluate its precision and accuracy using the embryo transfer date of newborns from the largest EPIC-derived dataset to date on assisted reproductive technologies (ART).

Methods: We built an epigenetic gestational age clock using Lasso regression trained on 755 randomly selected non-ART newborns from the Norwegian Study of Assisted Reproductive Technologies (START)—a substudy of the Norwegian Mother, Father, and Child Cohort Study (MoBa). For the ART-conceived newborns, the START dataset had detailed information on the embryo transfer date and the specific ART procedure used for conception. The predicted gestational age was compared to clinically estimated gestational age in 200 non-ART and 838 ART newborns using MM-type robust regression. The performance of the clock was compared to previously published gestational age clocks in an independent replication sample of 148 newborns from the Prediction and Prevention of Preeclampsia and Intrauterine Growth Restrictions (PREDO) study—a prospective pregnancy cohort of Finnish women.

Results: Our new epigenetic gestational age clock showed higher precision and accuracy in predicting gestational age than previous gestational age clocks ($R^2 = 0.724$, median absolute deviation (MAD) = 3.14 days). Restricting the analysis to CpGs shared between 450 K and EPIC did not reduce the precision of the clock. Furthermore, validating the clock on ART newborns with known embryo transfer date confirmed that DNA methylation is an accurate predictor of gestational age ($R^2 = 0.767$, MAD = 3.7 days).

Conclusions: We present the first EPIC-based predictor of gestational age and demonstrate its robustness and precision in ART and non-ART newborns. As more datasets are being generated on the EPIC platform, this clock will be valuable in studies using gestational age to assess neonatal development.

*Correspondence: KristineLokas.Haftorn@fhi.no

†Astanand Jugessur and Jon Bohlin: Joint senior authors

¹ Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Keywords: DNA methylation, Epigenetics, Gestational age, Illumina MethylationEPIC BeadChip, Assisted reproductive technologies, IVF, ICSI, MoBa, MBRN, PREDO

Background

Accurate determination of gestational age is important for assessing fetal development and maturity. This is necessary for investigating the impact of prenatal factors on pregnancy outcomes and any deviation from normal fetal development [1, 2]. Although gestational age at birth exhibits some normal variation, both preterm and post-term births are associated with an increased risk of adverse perinatal outcomes and health outcomes later in life [3–7]. The effects of gestational age at birth on health outcomes may be linked to epigenetic patterns established in utero or early in the postnatal period [8, 9]. Changes in these patterns may interfere with critical developmental processes [10–12] and trigger phenotypic changes that persist throughout life. This may be even more pertinent to children conceived by assisted reproductive technologies (ART), because ART procedures coincide with the extensive epigenetic reprogramming in the early embryo [13, 14].

DNA methylation (DNAm) is the most studied epigenetic mark in humans. It has, in recent years, been used to build gestational age clocks that can predict gestational age [15–18]. Earlier clocks were built using DNAm data from the Illumina HumanMethylation27 (27 K) or the Illumina HumanMethylation450 (450 K) BeadChip arrays, both of which have subsequently been replaced by the Illumina MethylationEPIC BeadChip (EPIC). EPIC has nearly twice (865,859 CpGs) as many CpGs as 450 K, and a stronger focus on regulatory elements [19]. Although EPIC includes over 90% of the probes on 450 K [19], six to eight of the CpGs included in existing gestational age clocks are not present on EPIC. This discrepancy may affect the precision of the published clocks in predicting gestational age when applied to DNAm data generated on EPIC [20]. Therefore, it is essential to develop a new gestational age clock that is updated and optimized for EPIC. Equally important is to elucidate whether the additional CpGs on EPIC enhance gestational age prediction.

A challenge in developing accurate gestational age clocks is the lack of information on the exact gestational age of the newborns. The standard approaches for estimating gestational age, based on ultrasound measurements or the last menstrual period (LMP), have thus far been used for training and testing epigenetic clocks. Ultrasound and LMP are widely used in clinical settings and have their individual advantages and limitations. While LMP can be informative, it suffers from large

variability, in part due to varying length of the follicular phase. Ultrasound is much more precise but still depends on the size of the fetus at the time of ultrasound [1, 21, 22]. On the other hand, for children conceived by ART, the exact time when the embryo is transferred back to the uterus is known. Although there may be some differences in the days before fertilization and embryo transfer, and the developmental speed may differ in the *in vitro* setting, the embryo transfer date (ETD) provides a more direct estimate of gestational age [23]. Therefore, DNAm data from ART births is particularly advantageous for developing and validating gestational age clocks. To our knowledge, no gestational age clock has yet been developed using ETD, although its use has been called for previously [16].

In addition to gestational age prediction, gestational age clocks can be used to estimate gestational age acceleration (GAA), which is defined as the discrepancy between gestational age predicted from DNAm data and gestational age derived from clinical measurements [16, 24]. Investigating GAA is important because of its reported association with several measures related to birth outcomes, such as the cerebroplacental ratio (a robust indicator of prenatal stress [25]), higher maternal body mass index, and larger birth size [26]. Although children conceived by ART have a higher risk of spontaneous preterm birth [27] and other adverse perinatal outcomes [28–30], only one small study has explored GAA in ART children [31].

To address these knowledge gaps, we developed a new gestational age clock based on EPIC-derived DNAm data from newborns in the Norwegian Study of Assisted Reproductive Technologies (START), which is a sub-study within the Norwegian Mother, Father and Child Cohort Study (MoBa) [32]. We validated this clock in test sets of ART and non-ART newborns in START, and also in an external dataset from the Finnish Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction (PREDO) study [33], which was used as a replication cohort. We also used the new EPIC-based clock to explore differences in GAA between ART and non-ART newborns.

Results

The EPIC gestational age clock

Table 1 and Fig. 1 provide overviews of the datasets used in this study. We fit a least absolute shrinkage and selection operator (Lasso) regression on DNAm data from 755

Table 1 Characteristics of the datasets used to evaluate the EPIC GA clock

Dataset	N	GA range (US, days)	Median GA (US, days)	GA range (ETD, days)	Median GA (ETD, days)	Sex ratio (% male)
START non-ART						
Training set	755	216–299	281.1	–	–	49
Test set	200	228–300	281.3	–	–	46
START ART						
Total	838	218–301	280.4	214–302	280.4	53
Training set	674	228–300	280.3	227–302	280.3	53
Test set	164	218–301	280.8	214–298	280.8	54
PREDO non-ART						
Test set	148	227–296	278.9	–	–	51

GA gestational age, US ultrasound, ETD embryo transfer date

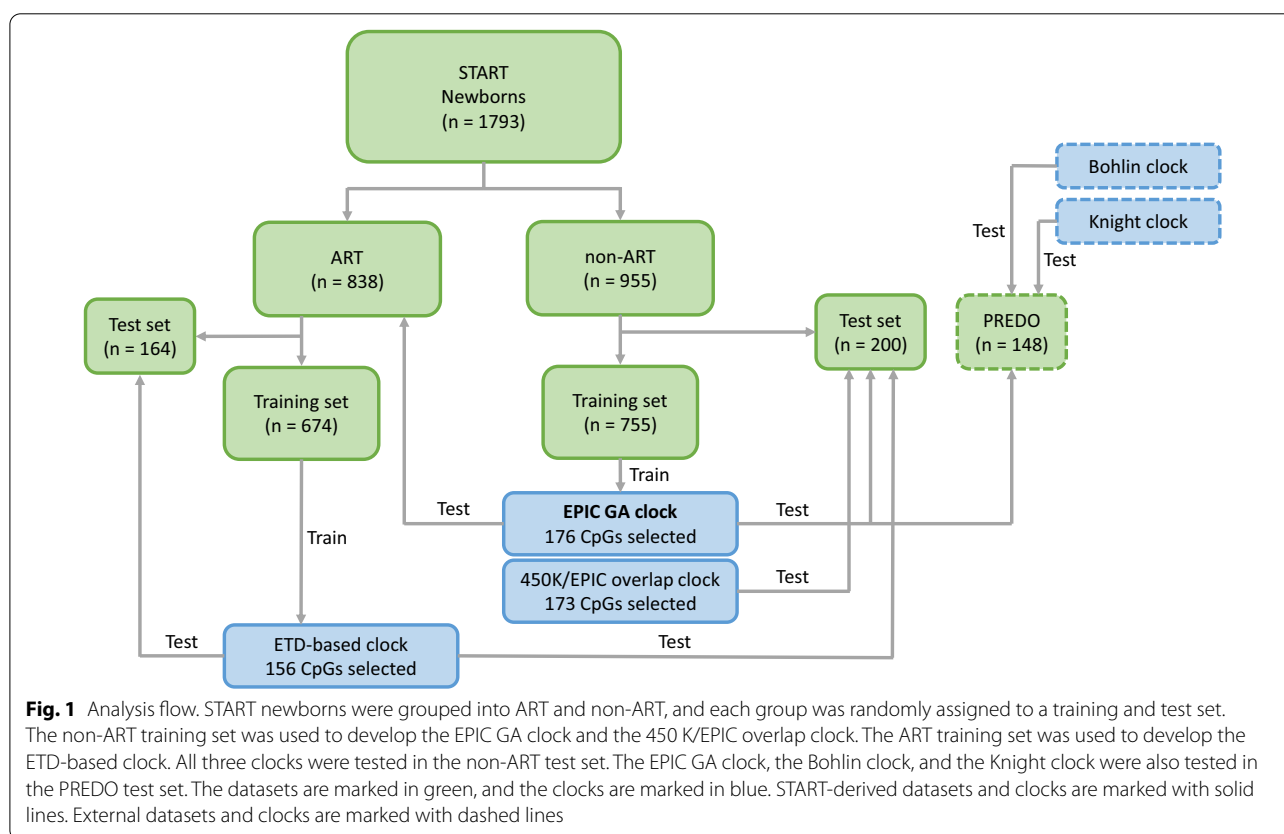


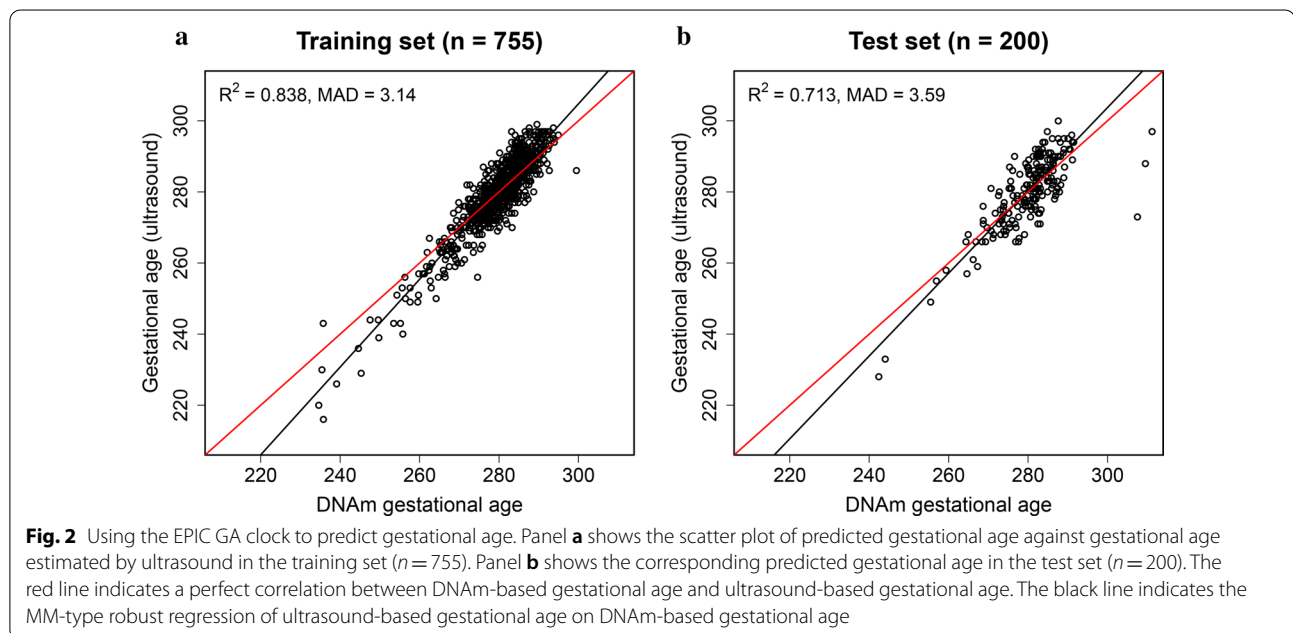
Fig. 1 Analysis flow. START newborns were grouped into ART and non-ART, and each group was randomly assigned to a training and test set. The non-ART training set was used to develop the EPIC GA clock and the 450 K/EPIC overlap clock. The ART training set was used to develop the ETD-based clock. All three clocks were tested in the non-ART test set. The EPIC GA clock, the Bohlin clock, and the Knight clock were also tested in the PREDO test set. The datasets are marked in green, and the clocks are marked in blue. START-derived datasets and clocks are marked with solid lines. External datasets and clocks are marked with dashed lines

non-ART newborns in START. 176 CpGs were selected for being predictive of gestational age. Individual CpG sites and their corresponding coefficients are provided in Additional file 4.

We validated the resulting predictor, referred to as “EPIC GA clock” hereafter, in a test set of 200 non-ART newborns from START. The EPIC GA clock showed an R^2 of 0.713 and a median absolute deviation (MAD) of 3.59 days (Fig. 2, Table 2).

Comparison with previously published gestational age clocks in an external replication cohort (PREDO)

Using an external dataset of EPIC-derived DNAm data on 148 non-ART newborns from the PREDO study [33], we compared the performance of our EPIC GA clock with two published epigenetic gestational age clocks that were built on DNAm data from the previous methylation arrays: the Bohlin clock [15], based on 450 K, and the Knight clock [16], based on 27 K and

**Table 2** Results of gestational age prediction in START and PREDO

Dataset* (count)	GA estimation method	Clock	R^2	SE	MAD
START non-ART ($n = 200$)	Ultrasound	EPIC GA clock	0.713	5.52	3.59
	Ultrasound	450 K/EPIC overlap clock	0.691	5.81	3.75
	Ultrasound	ETD-based clock	0.668	6.08	4.24
PREDO non-ART ($n = 148$)	Ultrasound	EPIC GA clock	0.724	5.08	3.42
	Ultrasound	Bohlin clock	0.610	6.06	6.69
	Ultrasound	Knight clock	0.406	6.99	4.55
START ART ($n = 838$)	Ultrasound	EPIC GA clock	0.767	5.32	3.80
	ETD	EPIC GA clock	0.767	5.30	3.70

*See also Table 1 and Fig. 1 for further details on these datasets

GA gestational age, SE standard error, MAD median absolute deviation, ETD embryo transfer date

450 K. Eight CpGs in the Bohlin clock and six CpGs in the Knight clock were absent from the PREDO dataset and were thus excluded from the analysis. Compared to the Bohlin and Knight clocks, our EPIC GA clock showed higher precision and accuracy in predicting gestational age (Fig. 3, Table 3). The difference in R^2 between the Bohlin clock and the EPIC GA clock was -0.062 (95% confidence interval (CI): $-0.117, -0.014$), and the difference in MAD was 3.27 days (95% CI: 1.87, 3.92). The corresponding statistics for the Knight clock versus our EPIC GA clock were -0.247 (95% CI: $-0.342, -0.161$) for R^2 and 1.13 days (95% CI: 0.196, 2.40) for MAD.

Assessing the impact of CpGs unique to EPIC on the prediction of gestational age

Of the 176 CpGs selected in the EPIC GA clock, 89 were found exclusively on EPIC. To assess whether the additional CpGs unique to EPIC affect the prediction parameters R^2 and MAD, we built a separate clock using the same training set but this time only including the 397,473 probes that are present on both 450 K and EPIC. We compared the performance of this new “450 K/EPIC overlap clock” (173 CpGs) to the EPIC GA clock (Fig. 4; Table 2) and found no significant difference in R^2 (-0.0001 ; 95% CI: $-0.021, 0.018$) or MAD (0.162; 95% CI: $-0.375, 0.794$) (Table 3). In terms of CpG overlap, 81

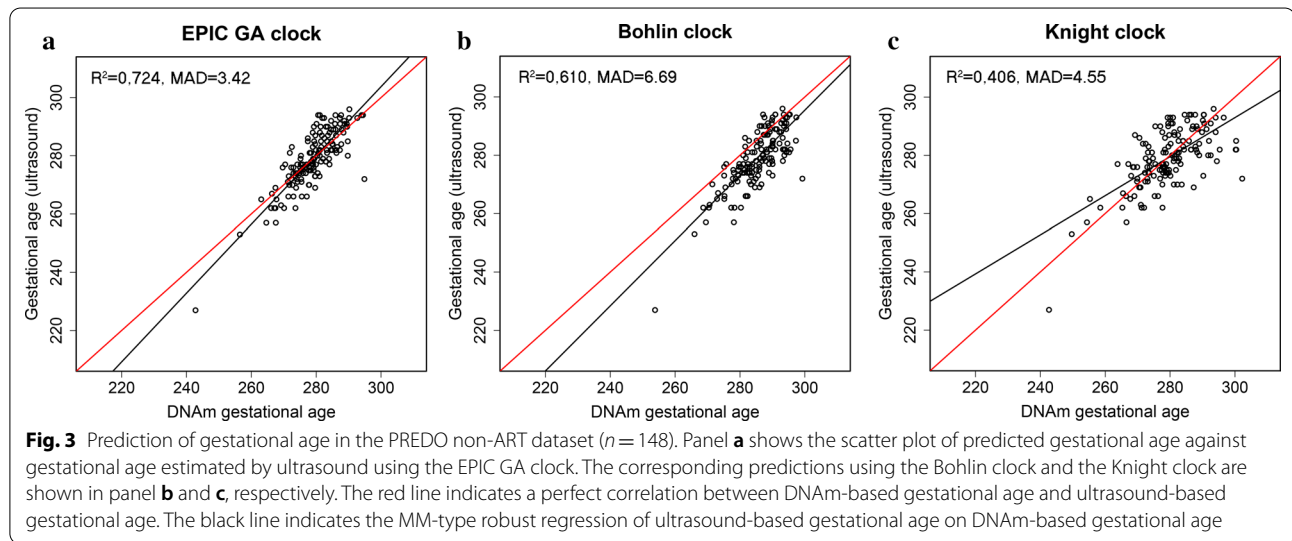
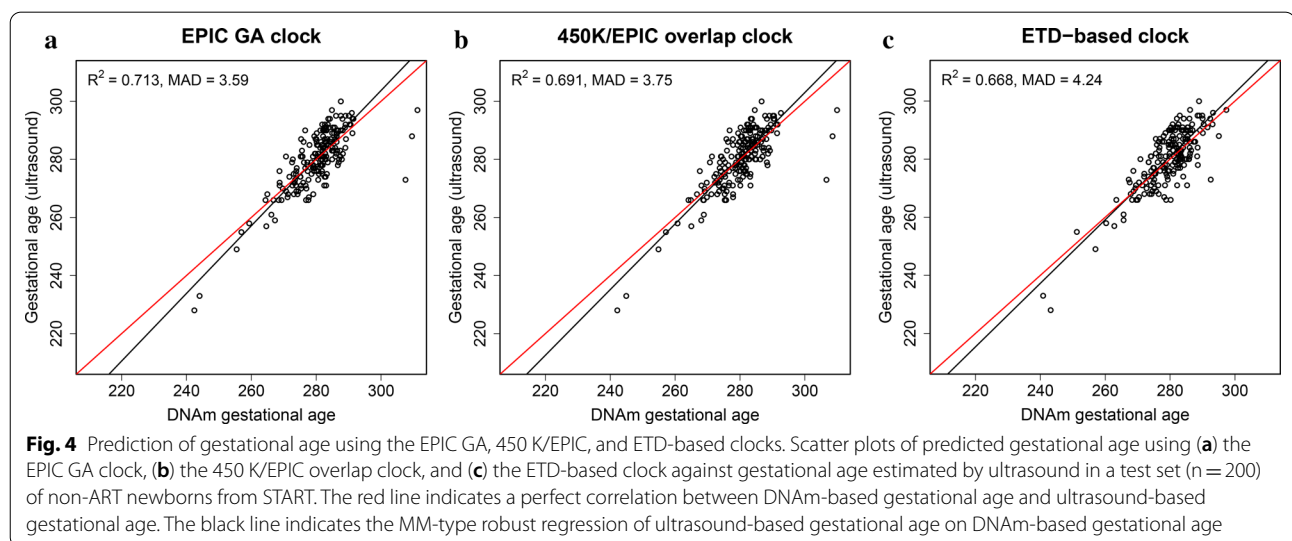


Table 3 Bootstrapped differences in R^2 , SE, and MAD between different clocks and GA estimation methods

Dataset * (count)	Comparison between clocks	R^2 (95% CI)	SE (95% CI)	MAD (95% CI)
START non-ART ($n = 200$)	450 K/EPIC overlap – EPIC GA	-0.0001 (-0.021, 0.018)	0.001 (-0.142, 0.175)	0.162 (-0.375, 0.794)
	ETD-based – EPIC GA	0.048 (-0.041, 0.123)	-0.409 (-1.00, 0.335)	0.645 (-0.181, 1.209)
	ETD-based – 450 K/EPIC overlap	0.048 (-0.039, 0.119)	-0.410 (-1.03, 0.308)	0.483 (-0.409, 0.984)
PREDO Non-ART ($n = 148$)	Bohlin – EPIC GA	-0.062 (-0.117, -0.014)	0.528 (0.095, 0.994)	3.27 (1.87, 3.92)
	Knight – EPIC GA	-0.247 (-0.342, -0.161)	1.89 (1.97, 2.69)	1.13 (0.196, 2.40)
	Knight – Bohlin	-0.185 (-0.273, -0.102)	1.36 (0.698, 1.97)	-2.15 (-3.11, -0.382)
Dataset * (count)	Comparison between GA estimation methods	R^2 (95% CI)	SE (95% CI)	MAD (95% CI)
START ART ($n = 838$)	ETD – ultrasound	0.015 (-0.003, 0.033)	-0.284 (-0.544, -0.037)	-0.102 (-0.465, 0.174)

*See Table 1 and Fig. 1 for further details on these datasets

GA gestational age, SE standard error, MAD median absolute deviation, ETD embryo transfer date



CpGs in the 450 K/EPIC overlap clock were also present in the EPIC GA clock.

Using the embryo transfer date (ETD) to predict gestational age

A great advantage of the ART dataset is that the ETD is known for the ART-conceived children. We thus developed a gestational age clock using the ETD of ART-conceived children to investigate whether it was possible to achieve a better predictor of gestational age. Six hundred and seventy-four ART newborns from START (Table 1, Fig. 1) were used to train the ETD-based clock. Additional file 1: Figure S1 shows the performance of the ETD-based clock for ultrasound- and ETD-estimated gestational age in the START ART training and test set, respectively. When compared to the EPIC GA clock in the non-ART test set from START, the ETD-based clock showed a similar performance, with an R^2 difference of 0.048 (95% CI: $-0.041, 0.123$) and a difference in MAD of 0.645 (95% CI: $-0.181, 1.209$) (Fig. 4; Table 3). The ETD-based GA clock contained 155 CpGs, and only 19 of them were in common with those of the EPIC GA clock.

Application of the EPIC GA clock to ART children

To assess the performance of the EPIC GA clock in ART-children, we applied the EPIC GA clock to the cord-blood DNAm data of 838 newborns conceived by ART (Table 1, Fig. 1). We compared predicted gestational age to gestational age estimated by ultrasound measurements and by ETD, respectively (Fig. 5). Gestational age estimated by ultrasound measurement and ETD was predicted with

similar precision (R^2 difference of 0.015 (95% CI: $-0.003, 0.033$); Fig. 5, Table 3) and accuracy (MAD difference of -0.102 (95% CI: $-0.465, 0.174$)).

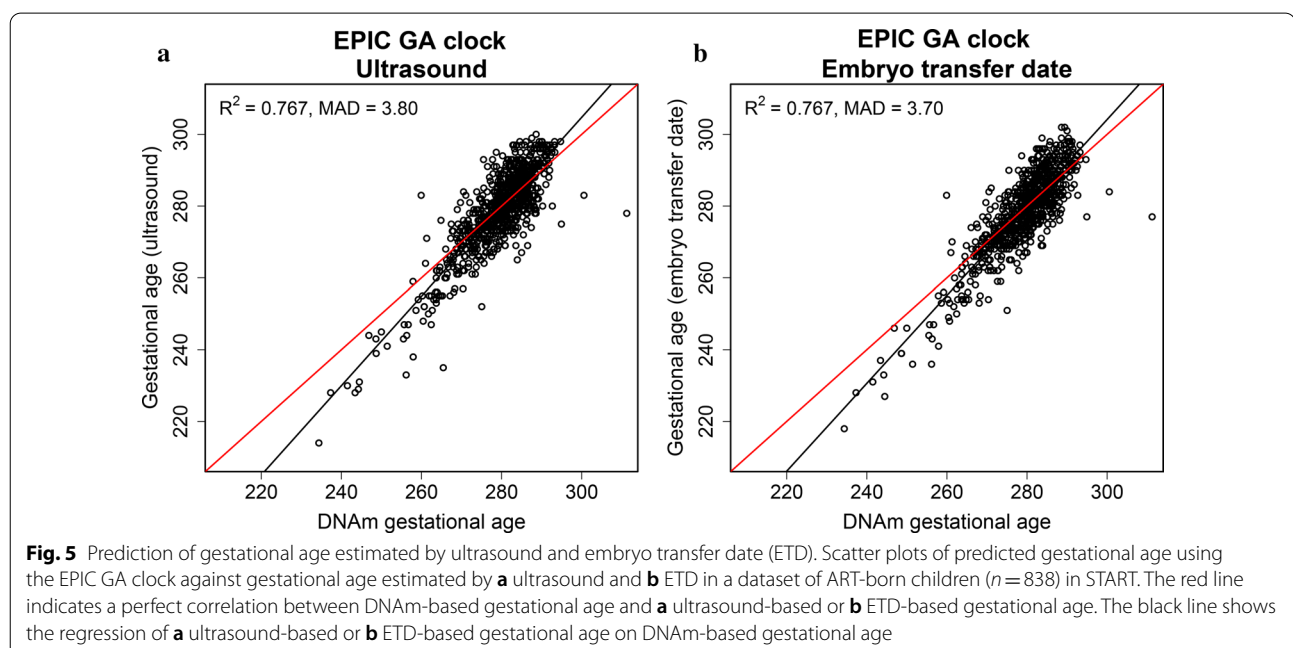
Gestational age acceleration in ART children

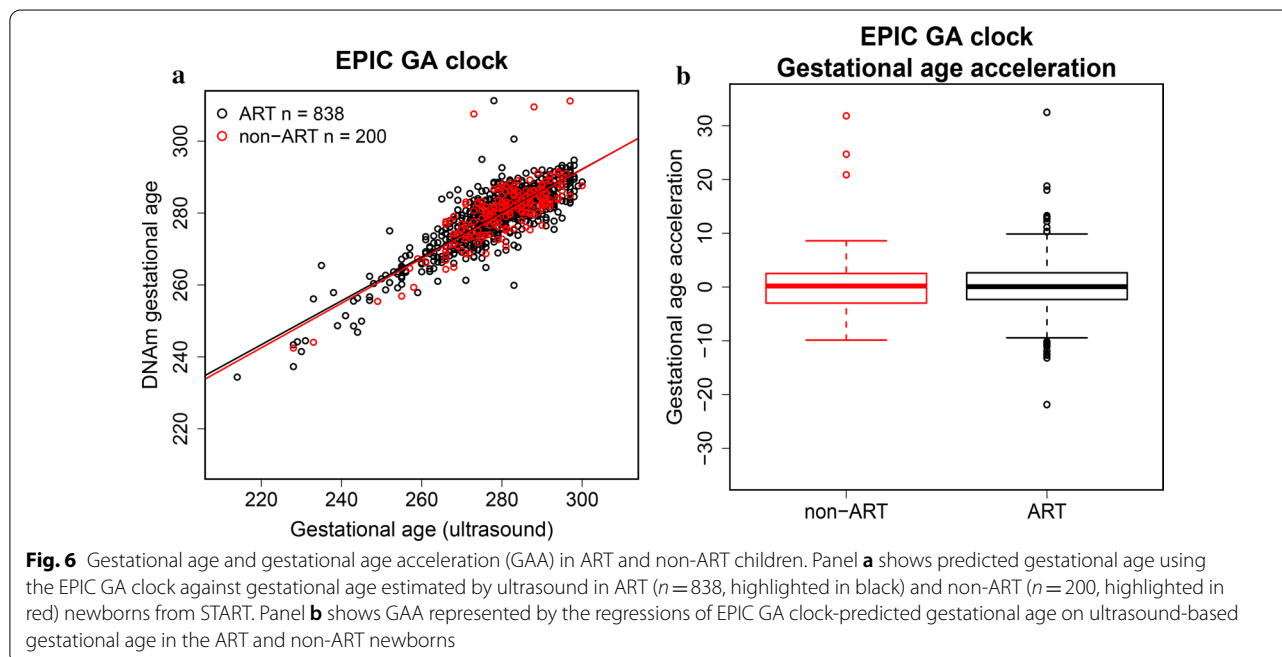
To assess whether GAA is associated with ART, we first regressed gestational age predicted by the EPIC GA clock on gestational age estimated by ultrasound in 200 non-ART and 838 ART newborns from START. GAA was calculated using the residuals from this regression. Next, we analyzed the relationship between GAA and ART by performing a logistic regression of ART on GAA. We found no significant difference in GAA between the ART ($n=838$) and non-ART ($n=200$) newborns ($p=0.388$, Fig. 6).

Aside from ETD, another major advantage of the START dataset is that the specific ART procedure used for conception was known, i.e., whether in vitro fertilization (IVF) was used alone or together with intracytoplasmic injection of sperm (ICSI), and whether the embryo was transferred fresh or after being frozen. We found no significant difference in GAA between newborns conceived by IVF alone ($n=470$) and those conceived by IVF in combination with ICSI ($n=338$) ($p=0.976$, Additional file 2: Figure S2). Furthermore, there was no significant difference between fresh ($n=693$) and frozen ($n=115$) embryo transfer ($p=0.274$, Additional file 3: Figure S3).

Gene-enrichment analysis

To explore the biological significance of the 176 CpGs selected in our EPIC GA clock, we performed





gene-enrichment analyses of the genes annotated for the selected CpGs. Using the annotation data provided in Illumina's Infinium MethylationEPIC v1.0 B4 Manifest file, we identified 154 unique gene names annotated for the 176 selected CpGs. A list of the 176 CpGs and their annotated genes is provided in Additional file 4. The software WebGestalt [34] was used to perform gene-enrichment analyses of the 154 genes [35]. WebGestalt identified 78 categories as being significantly enriched at a false discovery rate (FDR) < 0.01 . The category with the highest enrichment ratio was "regulation of platelet-derived growth factor receptor signaling pathway," containing *LRP1*, *HIP1R*, *HGS*, and *SRC* (enrichment ratio = 37; FDR = 0.003). Several of the significant hits were related to abnormal morphology of the eye, ear, nose, and other developmental categories, e.g., "plasma membrane-bounded cell projection organization" and "negative regulation of cellular biosynthetic process." The complete output of the WebGestalt analyses is provided in Additional file 5.

Discussion

We present the first EPIC-based predictor of gestational age and demonstrate its robustness and precision in ART versus non-ART newborns. This study benefited greatly from having the largest ART dataset to date, with detailed information on ETD and the specific procedure used for conception. Our EPIC GA clock, trained on the START dataset, outperformed previous cord blood-based

gestational age clocks when compared in an independent Finnish test set (PREDO).

Previous DNAm-based clocks were developed using the now outdated 27 K and 450 K. EPIC has almost twice as many CpGs as 450 K, and while 27 K and 450 K mostly cover areas around genes and CpG-islands, some of the additional probes on EPIC target distal regulatory elements and intergenic regions [36]. We, therefore, hypothesized that the additional CpGs unique to EPIC might have enhanced the performance of the EPIC GA clock. However, when we developed a separate clock featuring only those probes that are shared between 450 K and EPIC, we observed a similar performance to the EPIC GA clock, indicating that the additional CpGs on EPIC did not significantly enhance the prediction of gestational age. This observation is consistent with recent findings on age prediction by Lee et al. [37]. Another plausible explanation for the superior performance of our EPIC GA clock might be related to the fact that eight CpGs in the Bohlin clock and six CpGs in the Knight clock are absent from the EPIC array. This discrepancy might have reduced the prediction accuracy of the earlier clocks when applied to EPIC data.

A substantial advantage of the START dataset is its large sample size combined with detailed information on ETD for the ART-conceived newborns and the specific ART procedures used for conception. Using ETD provides a more direct estimate of gestational age than estimates based on ultrasound measurement or LMP [23]. We thus checked whether a clock trained on gestational

age estimated by ETD would lead to a further improvement in gestational age prediction. The results showed that the two clocks had similar performance, despite the low overlap in CpGs and genes. This suggests that using ETD-based gestational age estimates for training does not significantly enhance prediction compared to clocks trained on ultrasound-based estimates, further highlighting the precision of the EPIC GA clock.

A higher risk of spontaneous preterm birth and other adverse perinatal outcomes has been reported among ART-conceived children [28–30]. Given that the timing of ART procedures coincides with the extensive epigenetic remodeling in the gametes and early embryo, and, further that epigenetic alterations have been reported in ART embryos and children [38–40], we investigated whether the epigenetic gestational age of ART newborns differed significantly from that of non-ART newborns. When we applied the EPIC GA clock to ART newborns, the precision of the gestational age prediction remained similar to that of the non-ART newborns, indicating that the clock is also well suited for predicting gestational age in ART newborns. Furthermore, the EPIC GA clock predicted both ETD-based and ultrasound-based gestational age equally well, again underscoring the precision of the clock. Finally, we found no significant differences in GAA between ART and non-ART newborns.

ART is a collective term used to describe different procedures and categories that may have different impacts on fetal DNAm. It is therefore particularly important to investigate whether gestational age prediction differs according to the specific ART procedure used. For instance, embryos may be transferred to the uterus when they are fresh or after being frozen, and IVF may or may not involve ICSI. A previous study [31] examining GAA in ICSI newborns compared to non-ART newborns did not find any significant difference between the two groups. However, the authors detected a significant decrease in DNAm-predicted gestational age at birth among the ICSI newborns. To verify these findings in our dataset, we conducted another set of analyses to explore differences between IVF, ICSI, and non-ART newborns, as well as between fresh, frozen, and non-ART-conceived newborns. We found no significant differences in DNAm-predicted GA or GAA between any of the groups (Additional file 2: Figure S2 and Additional file 3: Figure S3), further strengthening the hypothesis that GAA is not associated with ART.

Although DNAm is strongly associated with gestational age, the mechanisms underlying this association are not well understood. A closer inspection of the specific CpGs selected for gestational age prediction and the overlap between different clocks may provide some answers. Of the 176 CpGs selected by the EPIC GA clock, only 11

were in common with the CpGs in the Bohlin clock, and none overlapped with the CpGs in the Knight clock. This could partly be explained by the 89 EPIC-specific CpGs. The lack of overlap in CpGs across different clocks has also been observed in age prediction models [41]. Our analyses showed little overlap between the EPIC GA clock and the ETD-based clock, even though both were trained on EPIC data. As Lasso regression and elastic net regression may select CpGs that are not associated with the outcome per se [42], dataset-specific CpGs could end up being included in the model. Furthermore, Lasso selects one CpG for each group of correlated (or neighboring) CpGs, whereas elastic net regression selects several CpGs, leading to a so-called “grouping effect” [43], which could lead to less overlap in CpGs between prediction models.

Unraveling the biological mechanisms underlying the gestational age clocks requires identifying the genes associated with the clock-specific CpGs and examining how they are related to gestational age. Our results revealed several genes in common across the different clocks. For example, 13 genes were shared between the EPIC GA clock and the Bohlin clock, while 15 genes were shared between the EPIC GA clock and the ETD-based clock. Some of the CpGs and genes in the EPIC GA clock appear to be stably associated with gestational age. For example, CpGs linked to Nuclear Receptor Corepressor 2 (*NCOR2*) and Insulin-Like Growth Factor 2 mRNA-binding protein 1 (*IGF2BP1*) were selected in both the EPIC GA clock and the Bohlin clock, and both of these genes have previously been identified in other studies of gestational age [44–47]. *NCOR2* is involved in vitamin A metabolism and lung function [48], and *IGF2BP1* plays an important role in embryogenesis and carcinogenesis [49]. The EPIC GA clock also identified CpGs related to Corticotropin-Releasing Factor-Binding Protein (*CRHBP*), consistent with previous studies of gestational age [8, 50]. *CRHBP* levels rise throughout pregnancy but drop markedly when approaching term [51]. Furthermore, Mastorakos and Ilias [52] showed that *CRHBP* might prevent aberrant pituitary-adrenal stimulation in pregnancy. In addition to the genes mentioned here, several other genes linked to the CpGs in our clock have previously been implicated in gestational age, including Muscleblind Like Splicing Regulator 1 (*MBNL1*), CD82 molecule (*CD82*), Integrin Subunit Beta 2 (*ITGB2*), and Rap Guanine Nucleotide Exchange Factor 3 (*RAPGEF3*) [47, 50]. Additional studies are needed to elucidate their roles in gestational age.

For a clock to be useful, it needs to be generalizable to other cohorts and populations. As with the Bohlin clock, our EPIC GA clock was trained on data from a relatively homogeneous cohort in terms of ethnicity,

socioeconomic status, and age [32, 53]. Our clock performed equally well in the independent Finnish PREDO cohort. However, while the use of a homogeneous training set may enhance the prediction model [42, 54], it can also result in a cohort-specific clock that is less generalizable to other populations.

Exploring associations between specific neonatal outcomes and DNAm-based gestational age is still in its nascent stages [26, 55], and there are many unanswered questions regarding neonatal development. The development of an EPIC-specific gestational age clock may offer additional insights into gestational age and neonatal development. As the 450 K array has been discontinued, we anticipate that future research on DNAm-based GA clocks will migrate to the more updated EPIC array. Research on GA-related topics and DNAm utilizing the 450 K array are expected to continue for some time, as many 450 K-based datasets are still in circulation and some are being used in consortia-led efforts. The clocks presented here may facilitate further research on DNAm-based clocks for both 450 K and EPIC-based arrays.

Conclusions

The new EPIC GA clock presented here predicted gestational age precisely in both ART and non-ART newborns and outperformed previous cord blood-based gestational age clocks when validated in an independent test set. The increased performance was not due to the higher coverage of CpGs on the EPIC array. Furthermore, the use of ETD-estimated gestational age for training did not improve the precision of gestational age prediction significantly compared with clocks trained on ultrasound-estimated gestational age. This is reassuring, as most datasets on newborns only have ultrasound- or LMP-based measures of gestational age. Finally, we did not find any significant association between GAA and ART. With a growing number of epigenetic datasets currently being generated on the EPIC platform, we expect our EPIC GA clock to become increasingly valuable in assessing developmental maturity in studies of neonatal development and disease.

Methods

Study population

MoBa is an ongoing, population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health (NIPH). Totally, 114,500 children, 95,200 mothers, and 75,200 fathers were recruited from all over Norway from 1999 through 2008 [32]. The MoBa mothers consented to participation in 41% of the pregnancies. Extensive details on the MoBa cohort have been provided elsewhere [32, 56]. START is a substudy of MoBa and

consists of 1,995 newborns and their parents. Blood samples from the newborns were obtained from the umbilical cord at birth [56].

PREDO is a prospective pregnancy cohort of Finnish women who gave birth to a singleton live child between 2006 and 2010 [33]. The cohort comprises 1079 pregnant women; 969 of these had one or more known risk factors for preeclampsia and intrauterine growth restriction, whereas the rest had no such risk factors. The women were enrolled in the study when they arrived for their first ultrasound screening at 12–14 gestational weeks in 10 study hospitals in Southern and Eastern Finland. Blood samples were obtained from the cord blood of 998 newborns [57]. To validate the gestational age clocks, we used cord blood-based DNAm data from 148 newborns (Fig. 1).

DNAm profiling and quality control

Cord blood samples taken by a midwife immediately after birth were frozen [56]. Five hundred nanograms of DNA extracted from the cord blood of START newborns were shipped to LIFE & BRAIN GmbH in Bonn, Germany, for measurement of DNAm on the Illumina MethylationEPIC array (Illumina, San Diego, USA). The raw iDAT files were imported and processed in four batches using the R-package *RnBeads* [58]. 44,210 cross-hybridizing probes [59] and approximately 10,000 probes with a high detection p-value (above 0.01) were removed. 16,117 probes with the last three bases overlapping with a single-nucleotide polymorphism (SNP) were also excluded. The remaining DNAm signal was processed using BMIQ [60] to normalize the type I and type II probe chemistries. Control probes output from *RnBeads* were visually inspected for all samples, and those with low overall signals were removed. The *GreedyCut* option [58] was used to remove outliers with markedly different DNAm signals than the rest of the samples. This resulted in the removal of 58 samples in total. For consistency, CpG sites removed from one batch, due to poor quality and detection p-value, were also removed from subsequent batches. After quality control, 770,586 autosomal CpGs and 1945 samples remained in the final dataset. 1793 subjects for whom we had information on ultrasound-based gestational age were used to develop and validate the gestational age clocks in this study.

For the PREDO samples, DNA was extracted according to standard procedures. Methylation analyses were performed at the Max Planck Institute of Psychiatry in Munich, Germany. DNA samples were bisulfite-converted using the EZ-96 DNA Methylation kit (Zymo Research, Irvine, CA) and assayed on the Illumina Infinium MethylationEPIC array (Illumina, San Diego, USA). Three samples were excluded for being outliers based

on their median intensity values. Another three samples showing discordant phenotypic and estimated sex were excluded. A further three samples were contaminated with maternal DNA and were also removed [61]. Methylation beta-values were normalized using the *funnorm* function [62] in the R-package *minfi* [63]. Three samples showed density artifacts after normalization and were removed from further analysis. We excluded probes on the sex chromosomes, probes containing SNPs, and cross-hybridizing probes according to previously published criteria [59, 64, 65]. Furthermore, CpGs with a detection p-value >0.01 in at least 25% of the samples were also excluded. Finally, one duplicate sample was removed after quality control. The final dataset contained 812,987 CpGs and 148 samples. After normalization, no significant batch effects were identified.

Variables

For the START dataset, information on gestational age, sex, and ART status was extracted from the Medical Birth Registry of Norway (MBRN). Gestational age at birth was estimated by ultrasound measurements in week 18 of pregnancy. For the ART children, we used the date of egg retrieval plus 14 days to obtain a second estimate of gestational age. When the date of egg retrieval was not known, the date of embryo insertion was used instead, minus two days. For embryos that were frozen, we used the date of embryo insertion plus 14 days, and the number of days between egg retrieval and freezing. These three estimations of gestational age were combined into a variable called embryo transfer date (ETD). IVF and ICSI were defined as ART treatments, whereas children conceived by intrauterine insemination were defined as non-ART births.

For the PREDO dataset, information on gestational age and sex was extracted from the Finnish Medical Birth Register. Gestational age at birth was estimated by ultrasound measurements between 12 and 14 weeks of pregnancy.

Gestational age prediction

Figure 1 shows a flowchart of the analyses performed. Children conceived without ART (non-ART) were randomly split into two groups: a training set (~80%) for developing the clock and a test set (~20%) for validating the clock. We used Lasso regression from the R-package *glmnet* [66] to develop DNAm-based predictors of gestational age. Clinically estimated gestational age was regressed on the 770,586 remaining CpGs after quality control in the START dataset. For the “450 K/EPIC overlap clock,” only the 397,473 CpGs that were in common between 450 K and EPIC were used. Missing probes were imputed using the median imputation procedure in the

R-package *Hmisc* [67]. Tuning parameters α and λ were selected after tenfold cross-validation in the training set. For the “EPIC GA clock,” Lasso regression selected 176 CpGs ($\alpha=1$, $\lambda=0.66$), while for the 450 K/EPIC overlap clock and the “ETD-based clock,” 173 CpGs ($\alpha=1$, $\lambda=0.63$) and 156 CpGs ($\alpha=1$, $\lambda=0.62$) were selected, respectively. Individual CpG sites and their corresponding coefficients are provided in Additional file 4.

The above clocks were used to estimate gestational age in (i) the START non-ART test set, (ii) the START ART newborns, and (iii) the non-ART newborns from PREDO (see Fig. 1 for more details). Predicted gestational age was regressed on clinically estimated gestational age using MM-type robust linear regression [68] from the R-package *robustbase* [69]. The precision of a given prediction model was defined as the proportion of variance explained by the model (i.e., by the R^2 value). Accuracy, on the other hand, was defined as the median absolute deviation (MAD) between observed and predicted gestational age.

Comparison of prediction parameters

To compare the performances of the different clocks and GA estimation methods, we calculated the differences in R^2 , SE, and MAD when computed by two different clocks or GA methods. To assess the size and significance of the differences, we computed bootstrap confidence intervals for each difference. Since all three performance measures can be calculated from observed and predicted GA values, each bootstrap sample selected individuals randomly and used the observed and predicted GA values already calculated for those individuals. The pairs of R^2 , SE, and MAD values were calculated from the same bootstrap sample to account for the same dataset being used in each comparison. Thus, we did not need to refit the full prediction model for each bootstrap sample.

The bootstrapping was performed using the R-package *boot* [70, 71]. 95% confidence intervals of the bootstrap differences were standard percentile intervals, reported as type “perc” by the *boot* package. A difference was considered statistically significant when the corresponding confidence intervals did not include the value 0.

Gestational age acceleration analysis

GAA was defined as the residuals from a linear regression of DNAm gestational age predicted by the EPIC GA clock on ultrasound-estimated gestational age [16]. We tested for association between GAA and ART by performing a logistic regression of ART on GAA.

Gene-enrichment analysis

The online functional enrichment software WebGeStalt [34] was used to search for enrichment within the

annotated genes of the EPIC GA clock. We identified 154 unique gene names annotated for the 176 CpGs selected in the EPIC GA clock using the annotation data from Illumina's Infinium MethylationEPIC v1.0 B4 Manifest file. We then performed an overrepresentation analysis on the 154 genes using Fisher's exact test [35], assigning a minimum of five genes per category, and using the genome as background. WebGestalt leverages data from the following databases for each category: gene ontology [72, 73] (Biological Process, Cellular Component, Molecular Function), pathway (KEGG [74], Panther [75], Reactome [76], WikiPathway [77]), network (Kinase target, Transcription Factor target, miRNA target), disease (DisGeNET [78], GLAD4U [79], OMIM [80]), drug (DrugBank [81]), phenotype (Human Phenotype Ontology [82]), and chromosomal location (Cytogenic Band). The Benjamini–Hochberg procedure was applied to the p-values, and categories with a false discovery rate below 0.01 were declared significantly enriched.

Abbreviations

ART: Assisted reproductive technologies; DNAm: DNA methylation; 27 K: Illumina HumanMethylation27; 450 K: Illumina HumanMethylation450; EPIC: Illumina MethylationEPIC Bead Chip; LMP: Last menstrual period; ETD: Embryo transfer date; GAA: Gestational age acceleration; START: Study of Assisted Reproductive Technologies; MoBa: Norwegian Mother, Father and Child Cohort Study; PREDO: Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction; Lasso: Least absolute shrinkage and selection operator; US: Ultrasound; MAD: Median absolute deviation; CI: Confidence interval; IVF: In vitro fertilization; ICSI: Intracytoplasmic sperm injection; FDR: False discovery rate; NIPH: Norwegian Institute of Public Health; MBRN: Medical Birth Registry of Norway; REK: The Regional Committees for Medical and Health Research Ethics.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13148-021-01055-z>.

Additional file 1: Figure S1. This figure shows the prediction of gestational age in ART newborns using the ETD-clock.

Additional file 2: Figure S2. This figure shows the subgroup analysis of GAA in ART newborns with or without ICSI.

Additional file 3: Figure S3. This figure shows the subgroup analysis of GAA in ART newborns with fresh or frozen embryo transfer.

Additional file 4. This file includes the CpGs selected by the (A) EPIC GA, (B) 450K/EPIC overlap and (C) ETD-based clocks, their corresponding coefficients and annotated genes.

Additional file 5. This file includes the results from the WebGestalt analysis.

Additional file 6. This file includes the intercept and coefficients of the EPIC GA clock.

Acknowledgements

MoBa is supported by the Norwegian Ministry of Health and Care Services and the Ministry of Education and Research. We are deeply indebted to all the participating families in Norway who participate in this ongoing cohort study. The Illumina EPIC arrays were processed at Life and Brain GmbH (www.lifeandbrain.com). This work was partly performed using the Services for Sensitive Data

facilities at the University of Oslo, Norway. The PREDO study would not have been possible without the dedicated contribution of the PREDO study group members: E Hamäläinen, E Kajantie, H Laivuori, PM Villa, and A-K Pesonen. We thank all the PREDO families for their enthusiastic participation. We also thank all the research nurses, research assistants, and laboratory personnel involved in the PREDO study.

Authors' contributions

KLH, AJ, and JB designed the research; HEN, CMP, and RL performed quality control on the START DNAm data; KLH and YL conducted analyses on the START data; DC performed quality control on the PREDO DNAm data; AM conducted analyses on the PREDO data; KLH, YL, AJ, JB, WRPD, HEN, HKG, KR, JL, MCM, and SEH interpreted the data; KLH drafted the manuscript; SEH, HKG and PM acquired funding, project administration, and resources for START; KR and JL acquired funding, project administration, and resources for PREDO. All authors provided scientific input, revised the manuscript, and approved the final version. All authors read and approved the final manuscript.

Funding

This work was supported in part by a grant from the National Institutes of Health (NIH) to AJ, PM, and HKG (Grant Number 1R01HL134840-01). The START project was funded by the Research Council of Norway through its Centres of Excellence funding scheme, project number 262700, and the NIPH. The PREDO Study was funded by the Academy of Finland, EVO (a special state subsidy for health science research), University of Helsinki Research Funds, the Signe and Ane Gyllenberg Foundation, the Emil Aaltonen Foundation, the Finnish Medical Foundation, the Jane and Aatos Erko Foundation, the Novo Nordisk Foundation, the Päivikki and Sakari Sohlberg Foundation, and the Sigrid Juselius Foundation granted to members of the PREDO Study Board. The funding bodies did not play any role in the design of the study, collection, analysis, or interpretation of data, nor in writing the manuscript.

Availability of data and materials

MoBa data can be accessed by applying directly to NIPH; <http://www.fhi.no/en/>. Due to ethical issues and written consent, the PREDO datasets analyzed in the current study are not publicly available. However, interested researchers can obtain a de-identified dataset after approval from the PREDO Study Board. Data requests may be subject to further review by the national register authority and by the ethical committees. Data can be obtained upon reasonable request from the PREDO Study Board (predo.study@helsinki.fi) or individual researchers. The intercept, CpG sites, and coefficients for the EPIC GA clock can be found in Additional file 6. The clock can be applied to DNAm data using the following procedure: (1) generate a matrix of beta values (n individuals by p CpG sites), (2) select the CpG sites for the EPIC GA clock (Additional file 6) out of the matrix of beta values, (3) calculate the linear combination of the beta values and coefficients of the selected CpGs sites, and 4) add the intercept (Additional file 6) to the linear combination.

Declarations

Ethics approval and consent to participate

The establishment of MoBa and the initial data collection were based on a license from the Norwegian Data Protection Agency and approval from The Regional Committees for Medical and Health Research Ethics (REK). The MoBa cohort is now based on regulations in the Norwegian Health Registry Act. The current study was approved by REK South-East C in Norway (Reference Number 21532). The study protocol of the PREDO cohort was approved by the Ethics Committees of the Helsinki and Uusimaa Hospital District and by the participating study hospitals.

Consent for publication

Written consents were obtained from the MoBa and PREDO participants.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway. ² Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway. ³ Institute of Health and Society, University

of Oslo, Oslo, Norway. ⁴ Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway. ⁵ Department of Mathematics, University of Oslo, Oslo, Norway. ⁶ Deepinsight, Karl Johans Gate 8, Oslo, Norway. ⁷ Department of Medical Genetics, Oslo University Hospital, Oslo, Norway. ⁸ Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland. ⁹ MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. ¹⁰ Bristol Medical School, Population Health Sciences, Bristol, UK. ¹¹ Division of Mental and Physical Health, Norwegian Institute of Public Health, Oslo, Norway. ¹² Department of Translational Research in Psychiatry, Max-Planck-Institute of Psychiatry, Munich, Germany. ¹³ Division for Infection Control and Environmental Health, Department of Infectious Disease Epidemiology and Modelling, Norwegian Institute of Public Health, Oslo, Norway.

Received: 7 December 2020 Accepted: 11 March 2021

Published online: 19 April 2021

References

- Knight AK, Conneely KN, Smith AK. Gestational age predicted by DNA methylation: potential clinical and research utility. *Epigenomics*. 2017.
- Kerstjens JM, de Winter AF, Bocca-Tjeertes IF, Bos AF, Reijneveld SA. Risk of developmental delay increases exponentially as gestational age of preterm infants decreases: a cohort study at age 4 years. *Dev Med Child Neurol*. 2012;54(12):1096–101.
- Boyle EM, Poulsen G, Field DJ, Kurinczuk JJ, Wolke D, Alfirevic Z, et al. Effects of gestational age at birth on health outcomes at 3 and 5 years of age: population based cohort study. *BMJ (Clinical research ed)*. 2012;344:e896.
- Yuan W, Basso O, Sorensen HT, Olsen J. Indicators of fetal growth and infectious disease in childhood—a birth cohort with hospitalization as outcome. *Eur J Epidemiol*. 2001;17(9):829–34.
- Kajantie E, Osmond C, Barker DJ, Eriksson JG. Preterm birth—a risk factor for type 2 diabetes? The Helsinki birth cohort study. *Diabetes Care*. 2010;33(12):2623–5.
- Bhutta AT, Cleves MA, Casey PH, Cradock MM, Anand KJ. Cognitive and behavioral outcomes of school-aged children who were born preterm: a meta-analysis. *JAMA*. 2002;288(6):728–37.
- El Marroun H, Zeegers M, Steegers EA, van der Ende J, Schenk JJ, Hofman A, et al. Post-term birth and the risk of behavioural and emotional problems in early childhood. *Int J Epidemiol*. 2012;41(3):773–81.
- Simpkin AJ, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, Ring SM, et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum Mol Genet*. 2015;24(13):3752–63.
- Hanson MA, Gluckman PD. Developmental origins of health and disease: new insights. *Basic Clin Pharmacol Toxicol*. 2008;102(2):90–3.
- Mani S, Ghosh J, Coutifaris C, Sapienza C, Mainigi M. Epigenetic changes and assisted reproductive technologies. *Epigenetics*. 2020;15(1–2):12–25.
- Morgan HD, Santos F, Green K, Dean W, Reik W. Epigenetic reprogramming in mammals. *Hum Mol Genet*. 2005;14 Spec No 1:R47–58.
- von Meyenn F, Reik W. Forget the parents: epigenetic reprogramming in human germ cells. *Cell*. 2015;161(6):1248–51.
- Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev*. 2014;28(8):812–28.
- Zhou F, Wang R, Yuan P, Ren Y, Mao Y, Li R, et al. Reconstituting the transcriptome and DNA methylome landscapes of human implantation. *Nature*. 2019;572(7771):660–4.
- Bohlin J, Haberg SE, Magnus P, Reese SE, Gjessing HK, Magnus MC, et al. Prediction of gestational age based on genome-wide differentially methylated regions. *Genome Biol*. 2016;17(1):207.
- Knight AK, Craig JM, Theda C, Baekvad-Hansen M, Bybjerg-Grauholm J, Hansen CS, et al. An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biol*. 2016;17(1):206.
- Mayne BT, Leemaqz SY, Smith AK, Breen J, Roberts CT, Bianco-Miotto T. Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation. *Epigenomics*. 2017;9(3):279–89.
- 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*. 2019;11(12):4238–53.
- 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 Biol*. 2016;17(1):208.
- Dhingra R, Kwee LC, Diaz-Sanchez D, Devlin RB, Cascio W, Hauser ER, et al. Evaluating DNA methylation age on the Illumina MethylationEPIC Bead Chip. *PLoS ONE*. 2019;14(4):e0207834.
- Skalkidou A, Kullinger M, Georgakis MK, Kieler H, Kesmodel US. Systematic misclassification of gestational age by ultrasound biometry: implications for clinical practice and research methodology in the Nordic countries. *Acta Obstet Gynecol Scand*. 2018;97(4):440–4.
- Gjessing HK, Grottnum P, Eik-Nes SH. A direct method for ultrasound prediction of day of delivery: a new, population-based approach. *Ultrasound Obstet Gynecol*. 2007;30(1):19–27.
- Delpachitra P, Palmer K, Onwude J, Meagher S, Rombauts L, Waalwyk K, et al. Ultrasound reference chart based on IVF dates to estimate gestational age at 6–9 weeks' gestation. *ISRN Obstet Gynecol*. 2012;2012:938583.
- Girchenko P, Lahti J, Czamara D, Knight AK, Jones MJ, Suarez A, et al. Associations between maternal risk factors of adverse pregnancy and birth outcomes and the offspring epigenetic clock of gestational age at birth. *Clin Epigenet*. 2017;9:49.
- Palma-Gudiel H, Eixarch E, Crispi F, Morán S, Zannas AS, Fañanás L. Prenatal adverse environment is associated with epigenetic age deceleration at birth and hypomethylation at the hypoxia-responsive EP300 gene. *Clin Epigenet*. 2019;11(1):73.
- Khouja JN, Simpkin AJ, O'Keeffe LM, Wade KH, Houtepen LC, Relton CL, et al. Epigenetic gestational age acceleration: a prospective cohort study investigating associations with familial, sociodemographic and birth characteristics. *Clin Epigenet*. 2018;10:86.
- Cavoretto P, Candiani M, Giorgione V, Inversetti A, Abu-Saba MM, Tiberio F, et al. Risk of spontaneous preterm birth in singleton pregnancies conceived after IVF/ICSI treatment: meta-analysis of cohort studies. *Ultrasound Obstet Gynecol*. 2018;51(1):43–53.
- Helmerhorst FM, Perquin DA, Donker D, Keirse MJ. Perinatal outcome of singletons and twins after assisted conception: a systematic review of controlled studies. *BMJ (Clinical research ed)*. 2004;328(7434):261.
- Kalra SK, Barnhart KT. In vitro fertilization and adverse childhood outcomes: what we know, where we are going, and how we will get there: a glimpse into what lies behind and beckons ahead. *Fertil Steril*. 2011;95(6):1887–9.
- Pandey S, Shetty A, Hamilton M, Bhattacharya S, Maheshwari A. Obstetric and perinatal outcomes in singleton pregnancies resulting from IVF/ICSI: a systematic review and meta-analysis. *Hum Reprod Update*. 2012;18(5):485–503.
- El Hajj N, Haertle L, Dittrich M, Denk S, Lehnen H, Hahn T, et al. DNA methylation signatures in cord blood of ICSI children. *Human Reprod (Oxford, England)*. 2017;32(8):1761–9.
- Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort profile update: the Norwegian mother and child cohort study (MoBa). *Int J Epidemiol*. 2016;45(2):382–8.
- Girchenko P, Lahti M, Tuovinen S, Savolainen K, Lahti J, Binder EB, et al. Cohort profile: prediction and prevention of preeclampsia and intrauterine growth restriction (PREDO) study. *Int J Epidemiol*. 2017;46(5):1380–1.
- Liao Y, Wang J, Jaehng EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47(W1):W199–205.
- Fisher RA. On the interpretation of χ_2 from contingency tables, and the calculation of P. *J R Stat Soc*. 1922;85(1):87–94.
- Heintzman ND, Ren B. Finding distal regulatory elements in the human genome. *Curr Opin Genet Dev*. 2009;19(6):541–9.
- Lee Y, Haftorn KL, Denault WRP, Nustad HE, Page CM, Lyle R, et al. Blood-based epigenetic estimators of chronological age in human adults using DNA methylation data from the Illumina MethylationEPIC array. *BMC Genom*. 2020;21(1):747.
- Melamed N, Choufani S, Wilkins-Haug LE, Koren G, Weksberg R. Comparison of genome-wide and gene-specific DNA methylation between ART and naturally conceived pregnancies. *Epigenetics*. 2015;10(6):474–83.
- Novakovic B, Lewis S, Halliday J, Kennedy J, Burgner DP, Czajko A, et al. Assisted reproductive technologies are associated with limited

- epigenetic variation at birth that largely resolves by adulthood. *Nat Commun.* 2019;10(1):3922.
40. White CR, Denomme MM, Tekpetey FR, Feyles V, Power SG, Mann MR. High frequency of imprinted methylation errors in human preimplantation embryos. *Sci Rep.* 2015;5:17311.
 41. Bergsma T, Rogaeva E. DNA methylation clocks and their predictive capacity for aging phenotypes and healthspan. *Neurosci Insights.* 2020;15:2633105520942221.
 42. Engebretsen S, Bohlin J. Statistical predictions with glmnet. *Clin Epigenet.* 2019;11(1):123.
 43. Zou H, Hastie T. Regularization and variable selection via the elastic net. 2005;6(2):301–20.
 44. Merid SK, Novoloaca A, Sharp GC, Küpers LK, Kho AT, Roy R, et al. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. *Genome Med.* 2020;12(1):25.
 45. Cruickshank MN, Oshlack A, Theda C, Davis PG, Martino D, Sheehan P, et al. Analysis of epigenetic changes in survivors of preterm birth reveals the effect of gestational age and evidence for a long term legacy. *Genome Med.* 2013;5(10):96.
 46. Fernando F, Keijser R, Henneman P, van der Kevie-Kersemaekers AM, Manens MM, van der Post JA, et al. The idiopathic preterm delivery methylation profile in umbilical cord blood DNA. *BMC Genom.* 2015;16:736.
 47. Wang XM, Tian FY, Fan LJ, Xie CB, Niu ZZ, Chen WQ. Comparison of DNA methylation profiles associated with spontaneous preterm birth in placenta and cord blood. *BMC Med Genom.* 2019;12(1):1.
 48. Minelli C, Dean CH, Hind M, Alves AC, Amaral AF, Siroux V, et al. Association of forced vital capacity with the developmental gene NCOR2. *PLoS ONE.* 2016;11(2):e0147388.
 49. Huang X, Zhang H, Guo X, Zhu Z, Cai H, Kong X. Insulin-like growth factor 2 mRNA-binding protein 1 (IGF2BP1) in cancer. *J Hematol Oncol.* 2018;11(1):88.
 50. Schroeder JW, Conneely KN, Cubells JC, Kilaru V, Newport DJ, Knight BT, et al. Neonatal DNA methylation patterns associate with gestational age. *Epigenetics.* 2011;6(12):1498–504.
 51. Perkins AV, Eben F, Wolfe CD, Schulte HM, Linton EA. Plasma measurements of corticotrophin-releasing hormone-binding protein in normal and abnormal human pregnancy. *J Endocrinol.* 1993;138(1):149–57.
 52. Mastorakos G, Ilias I. Maternal and fetal hypothalamic-pituitary-adrenal axes during pregnancy and postpartum. *Ann NY Acad Sci.* 2003;997:136–49.
 53. Nilsen RM, Vollset SE, Gjessing HK, Skjaerven R, Melve KK, Schreuder P, et al. Self-selection and bias in a large prospective pregnancy cohort in Norway. *Paediatr Perinat Epidemiol.* 2009;23(6):597–608.
 54. Simpkin AJ, Suderman M, Howe LD. Epigenetic clocks for gestational age: statistical and study design considerations. *Clin Epigenetics.* 2017;9:100.
 55. Knight AK, Smith AK, Conneely KN, Dalach P, Loke YJ, Cheong JL, et al. Relationship between epigenetic maturity and respiratory morbidity in preterm infants. *J Pediatr.* 2018;198:168–73.
 56. Paltiel L, Anita H, Skjerdet T, Harbak K, Bækken S, Nina Kristin S, et al. The biobank of the Norwegian Mother and Child Cohort Study—present status. *Norsk Epidemiologi.* 2014;24(1–2).
 57. Czamara D, Eraslan G, Page CM, Lahti J, Lahti-Pulkkinen M, Hämäläinen E, et al. Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. *Nat Commun.* 2019;10(1):2548.
 58. Muller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* 2019;20(1):55.
 59. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genomics data.* 2016;9:22–4.
 60. 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).* 2013;29(2):189–96.
 61. 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. *Clin Epigenet.* 2017;9:75.
 62. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* 2014;15(12):503.
 63. 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 (Oxford, England).* 2014;30(10):1363–9.
 64. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics.* 2013;8(2):203–9.
 65. 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. *Epigenet Chromatin.* 2013;6(1):4.
 66. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1–22.
 67. Harrell Jr. FE. Hmisc: Harrell Miscellaneous. R package 4.4–1 ed. <https://CRAN.R-project.org/package=Hmisc> 2020.
 68. Yohai V. High breakdown-point and high efficiency robust estimates for regression. *Ann Stat.* 1987;15.
 69. Maechler M RP, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Conceicao EL, Anna di Palma M. robustbase: Basic robust statistics. R package 0.93–6 ed. <http://robustbase.forge-r-project.org/>. 2020.
 70. Canty A. RBD. boot: Bootstrap R (S-Plus) Functions. R package version 1.3–25 ed. <https://CRAN.R-project.org/package=boot2020>.
 71. Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge: Cambridge University Press; 1997.
 72. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. *Gene Ontol Consortium Nat Genet.* 2000;25(1):25–9.
 73. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* 2019;47(D1):D330–d8.
 74. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2015;44(D1):D457–62.
 75. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 2003;13(9):2129–41.
 76. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* 2020;48(D1):D498–d503.
 77. Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, et al. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.* 2016;44(D1):D488–94.
 78. Piñero J, Ramírez-Anguaita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 2020;48(D1):D845–55.
 79. Jourquin J, Duncan D, Shi Z, Zhang B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genom.* 2012;13 Suppl 8(Suppl 8):S20.
 80. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD); [20.05.2020]. Available from: <https://omim.org/>.
 81. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 2006;34(Database issue):D668–72.
 82. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdiine JP, et al. Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* 2019;47(D1):D1018–27.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper 3

Stability selection enhances feature selection and enables accurate prediction of gestational age using only five DNA methylation sites

Stability selection enhances feature selection and enables accurate prediction of gestational age using only five DNA methylation sites

Kristine L. Haftorn^{1,2}, Julia Romanowska^{1,3}, Yunsung Lee¹, Christian M. Page^{1,4}, Per M. Magnus¹, Siri E. Håberg¹, Jon Bohlin^{1,5}, Astanand Jugessur^{1,3}†, and William R.P. Denault^{1,6}†

† Joint senior authors

Affiliations

¹ *Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway*

² *Institute of Health and Society, University of Oslo, Oslo, Norway*

³ *Department of Global Public Health and Primary Care, University of Bergen, 5020 Bergen, Norway*

⁴ *Department of Physical Health and Aging, Division for Mental and Physical Health, Norwegian Institute of Public Health, Oslo, Norway*

⁵ *Division for Infection Control and Environmental Health, Department of Infectious Disease Epidemiology and Modelling, Norwegian Institute of Public Health, Oslo, Norway*

⁶ *Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA*

Correspondence: Kristine L. Haftorn, KristineLokas.Haftorn@fhi.no

KEYWORDS: DNA methylation, epigenetics, gestational age, Illumina MethylationEPIC BeadChip, epigenetic clock, stability selection, feature selection, MoBa, MBRN, cord blood

ABSTRACT

Background: DNA methylation (DNAm) is robustly associated with chronological age in children and adults, and gestational age (GA) in newborns. This property has enabled the development of several epigenetic clocks that can accurately predict chronological age and GA. However, the lack of overlap in predictive CpGs across different epigenetic clocks remains elusive. Our main aim was therefore to identify and characterize CpGs that are stably predictive of GA.

Results: We applied a statistical approach called ‘stability selection’, which combines subsampling with variable selection, to DNAm data from 2,138 newborns in the Norwegian Mother, Father, and Child Cohort study. Twenty-four CpGs were identified as being stably predictive of GA. Intriguingly, only up to 10% of the CpGs in previous GA clocks were found to be stably selected. Based on these results, we used generalized additive model regression to develop a new GA clock consisting of only five CpGs that showed a similar predictive performance as previous GA clocks ($R^2 = 0.674$, median absolute deviation = 4.4 days). These CpGs were in or near genes implicated in immune responses, metabolism, and developmental processes. Furthermore, accounting for nonlinear associations improved prediction performance in preterm newborns.

Conclusion: We present a methodological framework for feature selection that is broadly applicable to any trait that can be predicted from DNAm data. We demonstrate its utility by identifying CpGs that are highly predictive of GA and present a new and highly performant GA clock based on only five CpGs that is more amenable to a clinical setting.

BACKGROUND

Epigenetic modifications are recognized for their prominent roles in aging and development (1, 2). DNA methylation (DNAm), one of the most studied epigenetic marks in humans (3), is strongly associated with gestational age (GA) in newborns and with chronological age in children and adults (4, 5, 6). This property of DNAm has enabled the development of several prediction models, commonly known as ‘epigenetic clocks’, that are highly predictive of age and GA (6, 7, 8, 9, 10, 11, 12). While it is now firmly established that epigenetic clocks perform exceptionally well in predicting chronological age and, in particular, GA, the reason for the lack of overlap in the selected DNAm sites (CpGs) across different epigenetic clocks has yet to be elucidated.

Current epigenetic clocks are based on variable selection methods such as penalized regression that suffer from two major drawbacks. First, they can be inconsistent in terms of variable selection when the covariates are measured with error and/or noise (13, 14). Second, if several correlated variables are predictive of the outcome, penalized regression methods tend to select only one among those variables (15). Given that DNAm is measured with noise (16, 17) and DNAm levels of neighboring CpGs often exhibit correlation (18, 19), the drawbacks of penalized regression methods may likely explain some of the inconsistency observed in the CpGs that are selected by different epigenetic clocks. To overcome these problems, we applied a statistical method called ‘stability selection’ (20) to identify CpGs that are repeatedly selected when predicting GA. In essence, stability selection combines subsampling with a chosen variable selection method, such as the ‘least absolute shrinkage and selection operator’ (lasso), to minimize the number of false discoveries in the set of selected variables.

Epigenetic clocks for GA have tremendous potential for epidemiological and clinical research as accurate predictors of GA and useful surrogates for assessing developmental maturity (21). However, current GA clocks comprise anywhere between a few dozen to several hundreds of CpGs (7, 8, 9, 12), which limit their utility. With current technology, quantifying such a large number of CpGs is too costly and not amenable to most clinical settings. One step towards broader applicability is to construct a more concise and cost-efficient epigenetic clock for GA using as few CpGs as possible without compromising too much on predictive performance. Specifically, this entails selecting the most biologically relevant CpGs while excluding those that mostly capture noise.

Our main aim here was to use stability selection to identify CpGs that are most likely to be stably predictive of GA across samples in an attempt to answer the following questions: i) Are there any CpGs that are stably predictive of GA, and, if yes, do these feature among those in existing GA clocks?; ii) Can the stably selected CpGs be used to build a GA clock consisting of fewer CpGs but that still shows a good performance compared to previously published GA clocks?; and iii) Can we obtain a biologically meaningful interpretation of how the predictive CpGs are linked to GA?

RESULTS

Study sample characteristics

The current analyses are based on DNAm data from 2,138 newborns from two random subsamples ($n = 956$ and $n = 1,182$) within the larger Norwegian Mother, Father, and Child Cohort (MoBa) study (22). DNAm data in both datasets were generated using the Illumina Infinium MethylationEPIC BeadChip (EPIC). The distributions of GA and sex were similar in the two datasets. GA ranged from 216 to 300 days (mean 279.8 days, SD 11.2 days) in the combined dataset (Table1).

Table 1: Characteristics of datasets used for selecting CpGs stably predictive of gestational age

Characteristic		Dataset 1 n = 956	Dataset 2 n = 1,182	Combined n = 2,138
GA in days	Mean (SD)	279.9 (10.8)	279.7 (11.6)	279.8 (11.2)
	Median	281	282	281
	Range	216-300	228-300	216-300
Sex (male), n (%)		470 (49%)	569 (48%)	1,039 (49%)

Abbreviations: GA, gestational age; SD, Standard deviation

Twenty-four CpGs were stably predictive of GA

To identify CpGs that are stably predictive of GA, we combined the stability selection methodology proposed by Meinshausen and Bühlmann (20) with lasso regression (23). We randomly selected 50% of the samples in our combined dataset and performed lasso regression on this subset. This process was repeated 1000 times. We then computed a selection probability for each CpG based on how many times it was selected as being predictive of GA. Finally, the formula derived by Meinshausen and Bühlmann (20) was used to choose a selection probability threshold above which CpGs were defined as being stably predictive of GA. The selection probability threshold depends on the maximum number of false discoveries we could allow on average in our set of stably selected CpGs. A more detailed explanation of the analytic pipeline is provided in the Methods section.

Figure 1 shows the 769,139 CpGs included in the analysis and their corresponding selection probabilities. When allowing for a maximum of two false discoveries, which corresponds to a selection probability of 0.73 and above (Supplementary Table 1), 24 CpGs were identified as stably predictive of GA (Table 2). The complete output of the stability selection analyses is provided in Supplementary Data 1.

Stability selection of CpGs predictive of GA

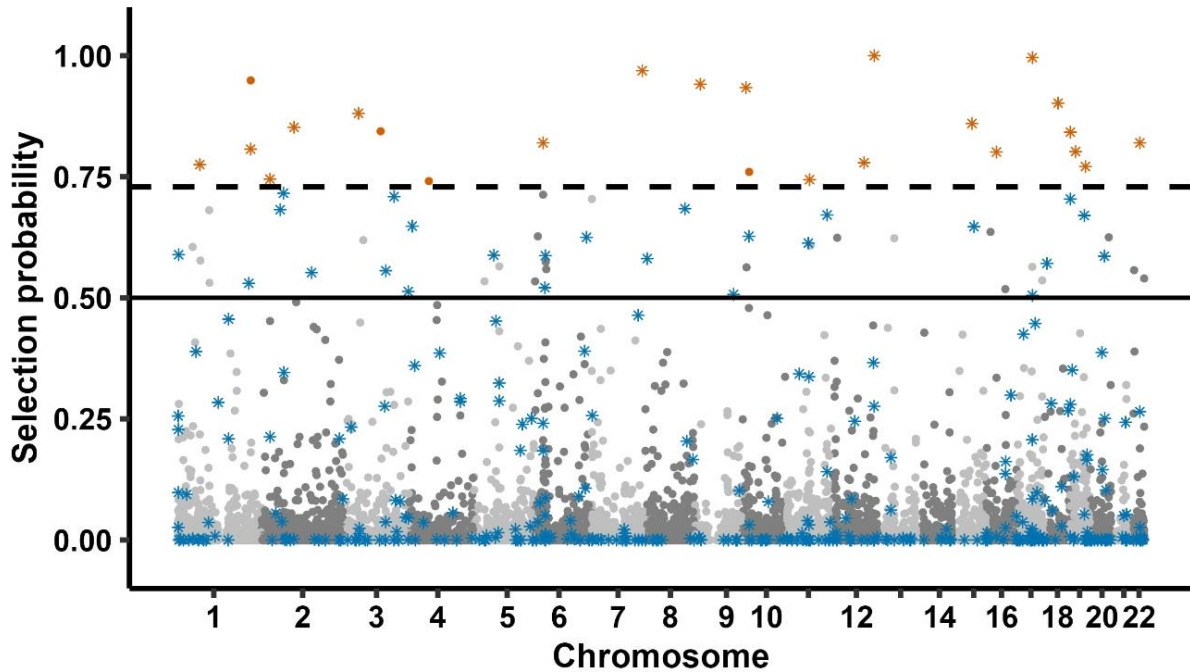


Figure 1. Selection probability of each CpG for the prediction of GA in cord-blood DNAm samples of newborns in MoBa ($n = 2,138$). Each point represents a single CpG ($n = 769,139$). The x-axis displays the CpGs according to their genomic coordinate, while the y-axis represents the selection probability calculated from the stability selection analysis. The solid horizontal line denotes a selection probability of 0.5, where a given CpG has an equal probability of being selected or excluded. The dashed black line denotes the selection probability threshold of 0.73. Asterisks signify CpGs that were selected in previously published GA clocks (specifically, the Haftorn clock (9), the Bohlin clock (7), or the Knight clock (8)). Orange signifies a CpG with a selection probability above the threshold of 0.73, and blue signifies a CpG from a previously published clock with a selection probability below that threshold.

Table 2: CpGs identified as being stably predictive of gestational age

CpG ID	Selection probability	Chr**	Genomic coordinates**	Relation to CpG Island**	Present on 450K **	Gene ID **
cg04347477	1.000	12	125,002,007	Island	yes	<i>NCOR2</i>
cg18183624	0.996	17	47,076,904	S_Shore	yes	<i>IGF2BP1</i>
cg25975961	0.969	7	150,600,818	Open sea	no	-
cg20320200	0.949	1	217,030,433	Open sea	yes	<i>ESRRG</i>
cg11387576	0.941	9	18,260,848	Open sea	no	-
cg11579708	0.934	10	13,142,679	S_Shore	no	<i>CCDC3; OPTN</i>
cg21180953	0.902	18	42,489,607	Open sea	no	<i>SETBP1</i>
cg09709426	0.881	3	45,911,521	Open sea	no	<i>LZTFL1</i>
cg07533333	0.860	15	59,793,834	Open sea	no	<i>FAM81A</i>
cg07749613	0.852	2	97,073,539	Open sea	yes	-
cg15393909	0.844	3	111,852,242	Open sea	no	<i>GCSAM</i>
cg10714639	0.842	19	1,075,104	S_Shore	yes	<i>HMHA1</i>
cg02567958	0.820	22	37,962,818	Island	yes	<i>CDC42EP1</i>
cg12681972	0.820	6	26,225,299	N_Shore	no	<i>HIST1H3E</i>
cg01833485	0.807	1	216,860,692	Open sea	yes	<i>ESRRG</i>
cg00840791	0.802	19	16,453,259	Open sea	no	-
cg16348385	0.801	16	30,106,822	N_Shore	yes	<i>YPEL3</i>
cg12999267	0.779	12	94,376,970	Open sea	yes	-
cg20301308	0.775	1	65,534,742	S_Shore	yes	-
cg12542255	0.771	19	45,976,195	Island	yes	<i>FOSB</i>
cg20734092	0.760	10	22,546,132	S_Shelf	no	<i>LOC100130992</i>
cg12434132	0.745	2	25,268,065	S_Shelf	no	<i>EFR3B</i>
cg11436362	0.744	11	67,053,929	S_Shore	yes	<i>ADRBK1</i>
cg03540917	0.741	4	57,686,587	N_Shore	no	<i>SPINK2</i>

Abbreviations: Chr, chromosome; S_Shore, south shore; N_Shore, north shore; S_Shelf, south shelf; 450K, Illumina HumanMethylation450 BeadChip

** Information extracted from the Illumina's Infinium MethylationEPIC v1.0 B4 manifest file. Genomic coordinates are according to the GRCh37 version of the human genome.

Most of the CpGs selected in GA clocks are not stably predictive of GA

To investigate the stability of CpGs selected for GA prediction in previously published GA clocks, we examined three different cord blood-based epigenetic GA clocks: (i) the 'Haftorn clock', based on EPIC samples (9), (ii) the 'Bohlin clock', based on 450K samples (7), and (iii) the 'Knight clock', based on 450K and 27K samples (8). In total, 389 unique CpGs in our analyses were previously selected in GA clocks; specifically, 176 in the Haftorn clock, 86 in the Bohlin clock, and 140 in the

Knight clock. Of these CpGs, two were in common between the Knight and the Bohlin clock, and 11 were in common between the Bohlin and the Haftorn clock. There were no shared CpGs between the Knight and the Haftorn clock. Eighteen (10.2%) of the Haftorn clock CpGs (Figure 2a) and eight (9.3%) of the Bohlin clock CpGs (Figure 2b) were found to be stably predictive of GA. By contrast, none of the Knight clock CpGs were found to be stably predictive of GA (Figure 2c). Interestingly, four of the CpGs identified as being stably predictive of GA, notably cg03540917, cg15393909, cg20320200 and cg20734092, were not selected by any of the above GA clocks.

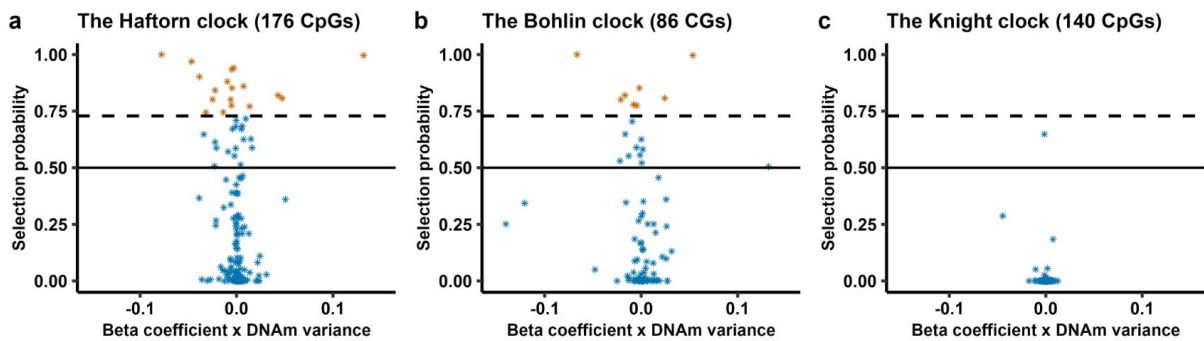


Figure 2. Selection probability of CpGs in our analyses that were selected for being predictive in three previously published GA clocks. Panel a shows the CpGs that were selected in the Haftorn clock ($n = 176$), panel b shows the CpGs that were selected in the Bohlin clock ($n = 86$), and panel c shows the CpGs that were selected in the Knight clock ($n = 140$). In each panel, the x-axis displays the beta coefficient for each CpG from the prediction model multiplied by the variance of DNAm in our samples, while the y-axis represents the selection probability calculated from the stability selection analysis. The solid horizontal line denotes a selection probability of 0.5 (i.e., a given CpG has an equal probability of being selected or excluded). The dashed black line denotes the selection probability threshold of 0.73. Orange signifies a selection probability above the threshold of 0.73, and blue signifies a clock-CpG with a selection probability below that threshold.

Five CpGs are enough to build a reliable GA clock

We investigated whether the CpGs identified as being stably predictive of GA could be used to build an independent epigenetic GA clock based on fewer CpGs but that still shows a similar performance as the previously published GA clocks. We randomly divided the total sample population into a training (80%, $n = 1,709$) and test set (20%, $n = 429$), and reran the stability selection analysis on the training set (Supplementary Data 2). When allowing for a maximum of two false discoveries, we identified 28 CpGs that were stably predictive of GA in this subset (selection probability threshold = 0.63). To further reduce the number of CpGs, we chose a stricter threshold by allowing a maximum of one false discovery (selection probability threshold = 0.76), which resulted in 15 stably selected CpGs (Figure 3).

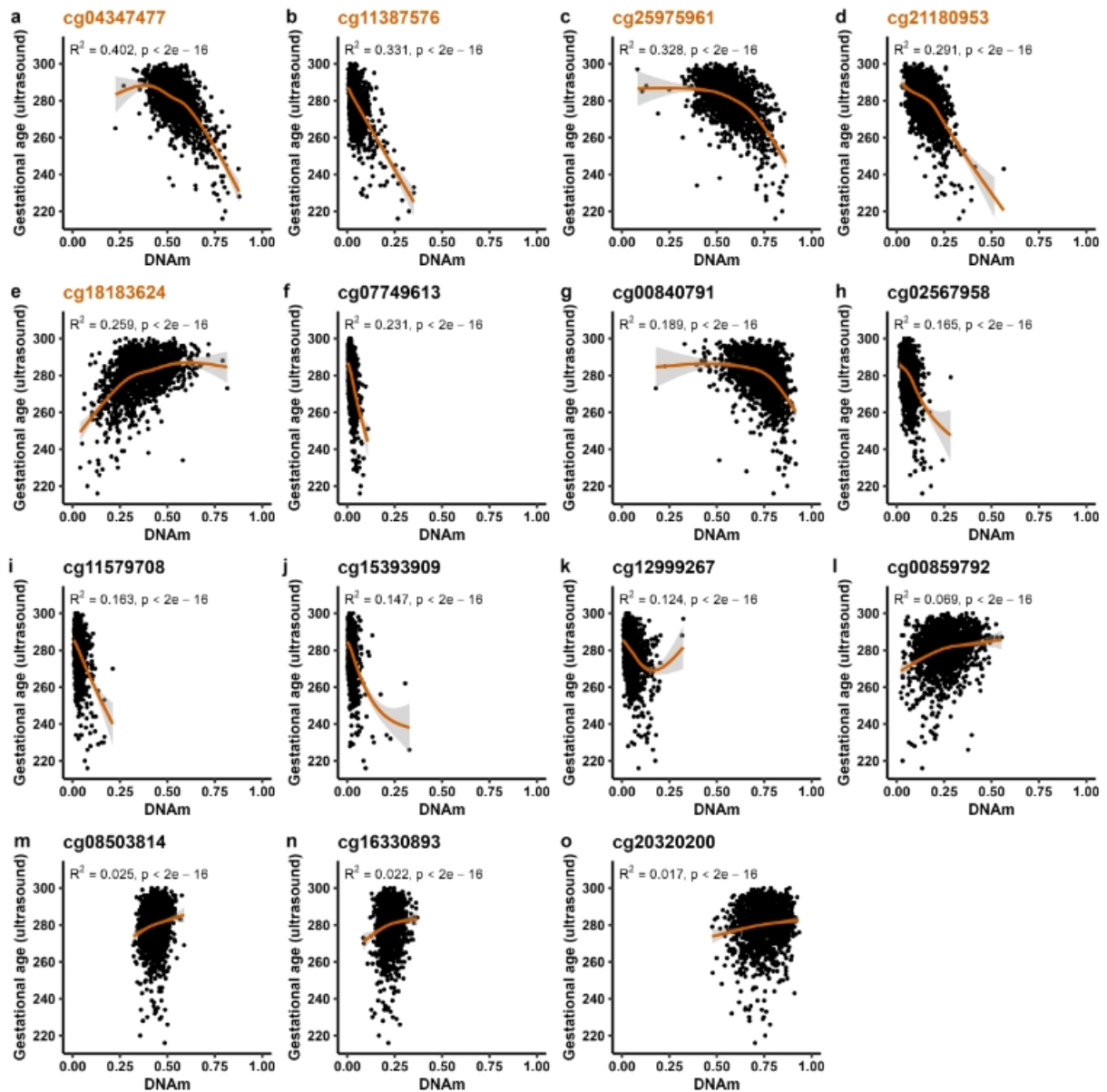


Figure 3. The relationship between DNAm level and GA for each of the 15 stably selected CpGs in the training set ($n = 1709$). In each of the panels (a-o), ultrasound-estimated GA (x axis) is plotted against the DNAm level (β -value) (y axis) for a given CpG. The orange line indicates the generalized additive model (GAM) regression of DNAm level on ultrasound-estimated GA. Orange CpG titles in panels a-e signify CpGs in the ‘5 stable CpG GA clock’.

To determine the number of CpGs needed to be included in a GA clock to achieve a similar predictive performance as that of previously published GA clocks, we first fitted generalized additive model (GAM) regressions of GA on DNAm levels in the training set for each of the 15 CpGs identified above and ordered them according to their R^2 value (Figure 3). The output of the regression on the CpG with the highest R^2 was used to predict GA in the test set ($n = 429$). This procedure was iterated by fitting a GAM regression of GA on DNAm levels of the two CpGs with the highest R^2 , then the

three CpGs with the highest R^2 , and so on and so forth, until we had constructed 15 different prediction models for GA. We then assessed predictive performance in the test set by comparing R^2 and median absolute deviation (MAD) for each of the 15 prediction models as well as one that was developed using a standard framework with lasso (Figure 4, Supplementary Table 2). When the predictive performance of the lasso model (with 233 CpGs) was compared to that of the rest of the clocks, it was evident that very few CpGs were needed to attain a sufficiently good prediction of GA. The top CpG (cg04347477) alone predicted GA with an R^2 of 0.52 and a MAD of 5.09 days. When including five CpGs (cg04347477, cg11387576, cg25975961, cg21180953 and cg18183624) in the ‘5 stable CpG GA clock’, we obtained an R^2 of 0.674 and a MAD of 4.4 days. These metrics are comparable to those of the Bohlin clock ($R^2 = 0.66$, standard error ± 12.5 days (95% prediction interval)) wherein 96 CpGs were needed for prediction (7). When using all 15 CpGs for prediction, R^2 increased only slightly, to 0.712 (MAD = 4.3) (Figure 5), suggesting that the five CpGs in the ‘5 stable CpG GA clock’ explain a remarkably high proportion of the variance in GA. Panels a-e in Figure 3 depict the relationship between GA and DNAm level of each of these five stably selected CpGs in the training set.

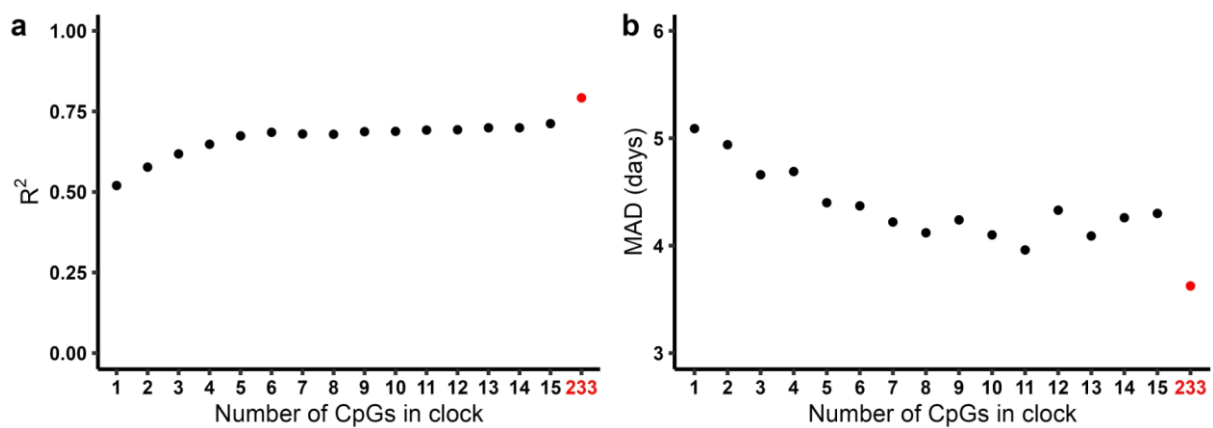


Figure 4. The relationship between the number of CpGs used for prediction and predictive performance in the test set (n = 429). Panel a shows the R^2 for each of the clocks and panel b shows the corresponding MAD in days. The red dot in each panel shows the predictive performance of a clock developed using the standard framework with lasso.

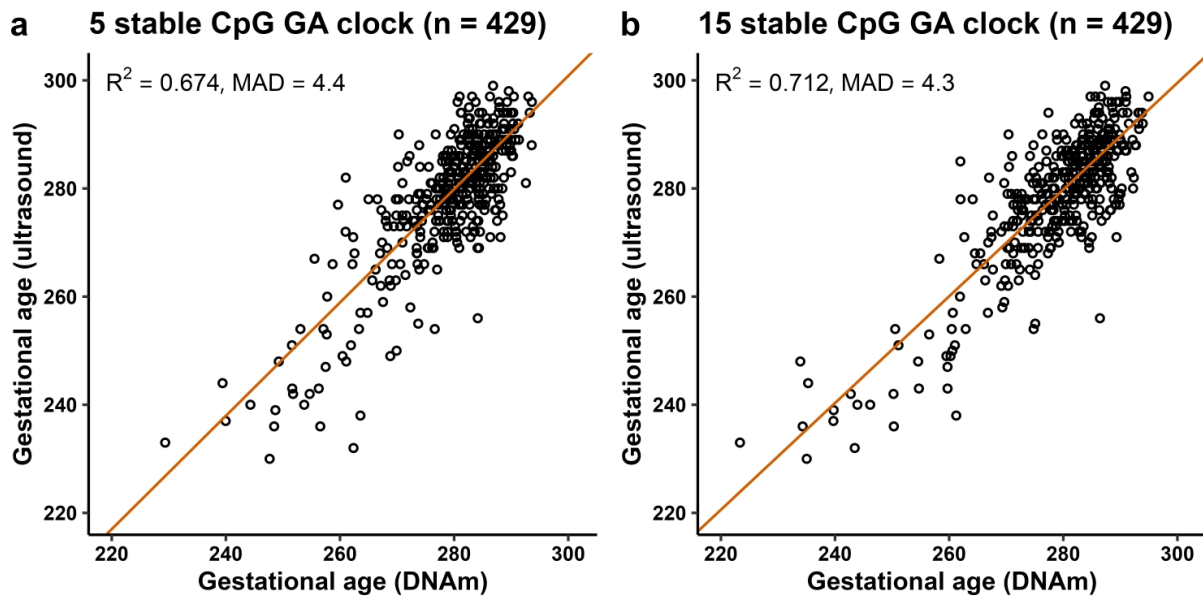


Figure 5. Prediction of GA in the test set (n = 429). Panel a shows the scatter plot of GA predicted by DNAm against GA estimated by ultrasound for the ‘5 stable CpG GA clock’. Panel b shows the corresponding predictions for the ‘15 stable CpG GA clock’. The orange diagonal line indicates the MM-type robust regression of ultrasound-estimated GA on DNAm-estimated GA.

Some of the predictive CpGs exhibit a nonlinear relationship with GA

When building clocks using stably predictive CpGs, GAM was used instead of regular linear regression to account for the observed nonlinearity in the relationship between DNAm and GA. The effective degrees of freedom (EDF) estimated from the GAM were used as a proxy for the degree of nonlinearity in the relationships between DNAm levels and GA (24). The EDF for the 15 CpGs ranged from 1 to 8.6, with 12 of the CpGs exhibiting a EDF higher than 1, indicating a nonlinear relationship (Supplementary Table 3). Only three of the CpGs had an EDF of 1, which is equivalent to a linear relationship. Moreover, the nonlinear relationships between DNAm and GA seem to have a larger effect on the precision of GA prediction in preterm compared to term newborns (Figure 6).

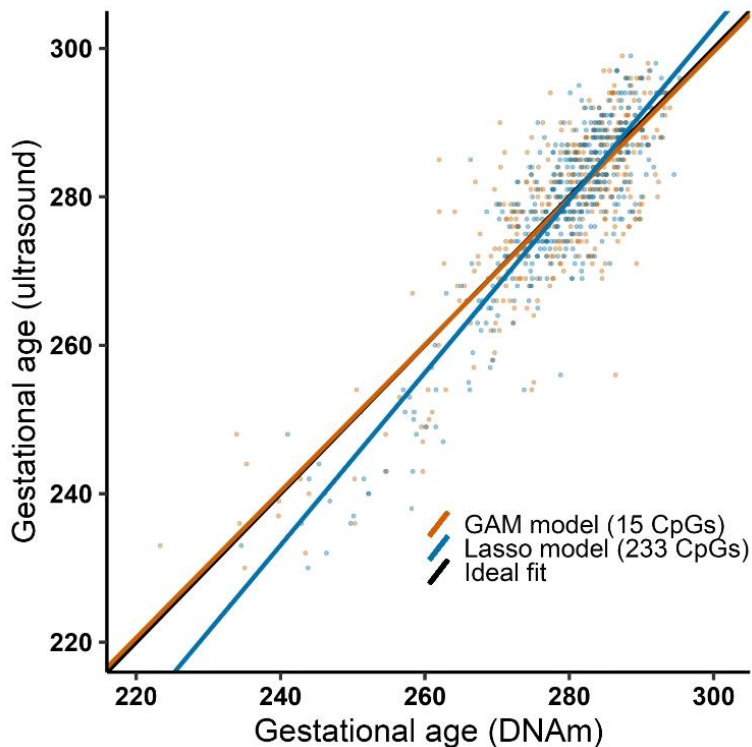


Figure 6. Prediction of gestational age using a GAM model versus a lasso model. Regression lines showing the relationship between ultrasound-estimated GA and predicted GA in the test set ($n = 29$) using a GAM model including 15 CpGs (orange line) and a lasso model including 233 CpGs (blue line). The black line indicates the ideal fit between ultrasound-estimated GA and DNAm-predicted GA.

Gene and regulatory region annotations of CpGs stably predictive of GA

We searched the *Ensembl* genome browser (25) to check whether the CpGs selected as being stably predictive of GA are located in or near genes or regulatory regions of known pathway annotations. Details on the regulatory region annotation of the remaining stably selected CpGs can be found in Supplementary Table 4 and in our GitHub repository. Almost half of the stably selected CpGs are located in promoter regions ($n = 11$, 46%). Table 3 presents a more detailed description of the gene and regulatory region annotations of the CpGs selected for the ‘5 stable CpG GA clock’. Three of the CpGs in this clock are located in or near specific genes: cg04347477 in *NCOR2*, cg21180953 in *SETBP1* and cg18183624 in *IGF2BP1*. Moreover, all five CpGs are linked to one or more regulatory regions. cg18183624, for example, is located in a region controlling a small cluster of different genes, several of which are implicated in prenatal development (*IGF2BP1* (26), *KAT7* (27), *HOXB13* and *HOXB5* (28)) immune responses (*TAC4* (29), *CALCOCO2* (30)), in addition to multiple regions encoding long non-coding RNAs (lncRNAs) (*ENSG00000250838*, *ENSG00000262837*, *NFE2L1-DT*, *ENSG00000251461*) (see Table 3 and Figure 7).

Table 3: Gene and regulatory region annotation of CpGs in the ‘5 stable CpG GA clock’

CpG ID	Gene (Ensembl annotation)	Gene Ensembl ID	Regulatory region type	Regulatory region Ensembl ID	Genes controlled by regulatory region
cg04347477	<i>NCOR2</i>	ENSG00000196498	Promoter	ENSR00001046350	-
cg11387576	-	-	Enhancer	ENSR00001448127	<i>SAXO1, PSMC3P1, HSALNG0070247, RF00017-7032, ADAMTSL1, HSALNG0070244</i>
cg25975961	-	-	Promoter flanking region	ENSR00001734862	-
			CTCF binding site	ENSR00000414350	-
cg21180953	<i>SETBP1</i>	ENSG00000152217	Promoter flanking region	ENSR00001902774	<i>Lnc-EPG5-10, 5MWI_A-078, SETBP1, SLC14A2</i>
cg18183624	<i>IGF2BP1</i>	ENSG00000159217	Promoter	ENSR00000095417	<i>IGF2BP1, ENSG00000250838, ENSG00000262837; UBE2Z; ENSG00000204584, FAM117A; LOC124904116, KAT7, PRAC1, PRAC2, HOXB13, TAC4, CALCOCO2, HOXB5, NXPH3, NFE2L1-DT, ENSG00000251461, ATP5MC1, LOC124904020, B4GALNT2</i>

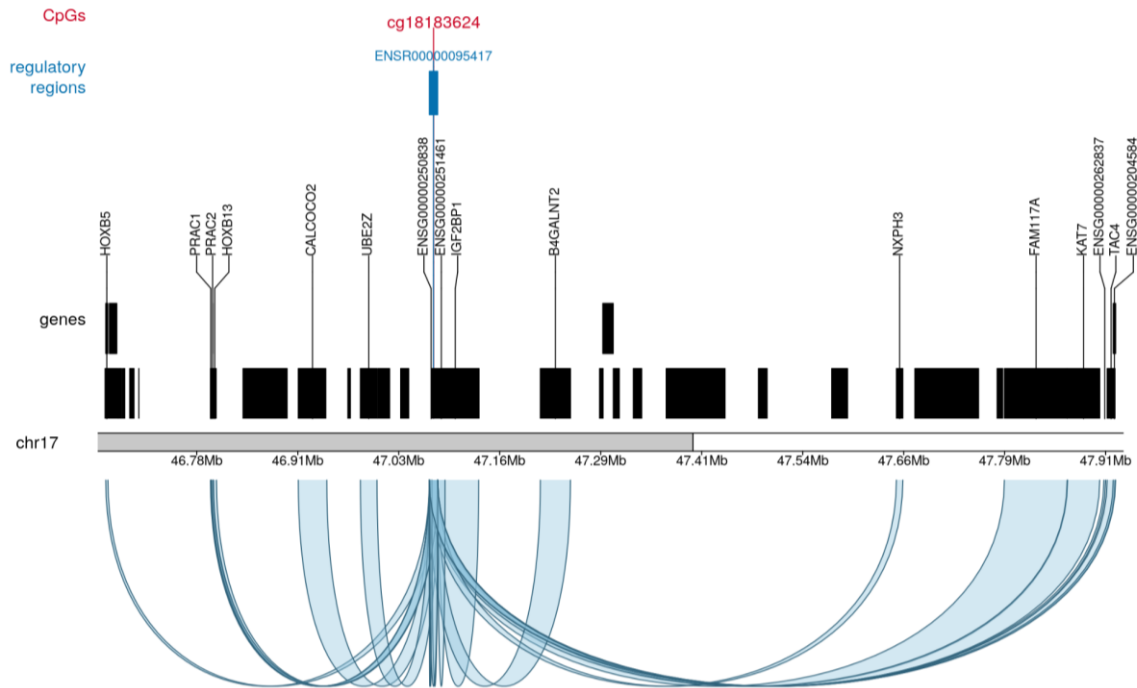


Figure 7. An illustrative example of the regulation map for cg18183624 on chromosome 17. The CpG, shown in red, is encompassed by the regulatory region ENSR00000095417 (blue-colored vertical bar). Below the regulatory region, all the genes are marked as black rectangles and those controlled by ENSR00000095417 are labeled by their gene symbols. The curves underneath the ideogram represent regulatory relationships between ENSR00000095417 and the genes, as predicted by GeneHancer.

Further, we searched for all the 24 stably predictive CpGs in the EWAS catalog (31) and the EWAS atlas (32). Many of the CpGs were found in previous studies of GA and preterm birth, of aging in early childhood, and of various pregnancy-related phenotypes like gestational diabetes and prenatal smoke exposure. The whole output from this analysis can be found in our GitHub repository.

DISCUSSION

We found 24 CpGs to be stably predictive of GA after applying a statistical framework that restricts the number of false discoveries in a set of predictive CpGs selected by penalized regression. The results also suggested that most of the CpGs included in previously published epigenetic GA clocks are dispensable. Furthermore, we showed that the stably selected CpGs can be used to construct new GA clocks based on a substantially smaller number of CpGs than previous GA clocks. Importantly, the new GA clocks retained a similar predictive performance to already established GA clocks. These

findings underscore the relevance of feature selection, not only in building more efficient epigenetic clocks for GA as here but also for other outcomes and epigenetic clocks.

Epigenome-wide association studies (EWAS) of GA have unraveled thousands of CpGs across the genome that are associated with GA (4, 7, 8, 33, 34). However, previous studies have shown that most CpGs exhibit a modest effect size (35). In theory, the presence of many predictive CpGs, where each explains approximately the same amount of variance, is likely to exacerbate the issue of different GA clocks selecting different CpGs. However, our identification of CpGs that were selected up to 100% of the time in different subsamples and that were also highly predictive of GA strongly indicate that only a handful of selected CpGs are needed to explain a remarkably large proportion of the DNAm variance related to GA.

When we compared our stably selected CpGs to those selected by three previously developed GA clocks, namely the Haftorn (9), Bohlin (7) and Knight (8) clocks, only about 10% of CpGs selected in the Bohlin and Haftorn clocks were stably predictive of GA. Moreover, none of the CpGs in the Knight clock were stably predictive of GA. It is important to note that the Bohlin and Haftorn clocks were both developed using samples from the MoBa study, whereas the Knight clock was trained on a combination of datasets from different cohorts. Additionally, the training set used to develop the Knight clock also differs from the Haftorn and Bohlin clocks with respect to several other important parameters, such as the range of GA, the sample size, and type of DNAm array (36). A particularly interesting observation in our study is that, even though the Haftorn clock was developed using a subset ($n = 755$) of the samples used in the current analyses and was validated in an external replication cohort, 90% of the CpGs in that clock were not considered stably predictive by the current statistical framework. This implies that most of the CpGs selected in epigenetic clocks developed using conventional penalized regression methods are either a selection of many CpGs that have varying degrees of association with GA individually, or that they are simply false positives (i.e., CpGs that are not directly associated with GA but merely tag along other CpGs that are associated with GA (15)). However, it is important to note that, with the stability selection approach, we may fail to detect CpGs that are highly correlated with each other or are part of larger genetic and/or epigenetic networks. Such CpGs may be selected less frequently individually and, therefore, would not be stably selected, although they might still be predictive of GA.

Epigenetic clocks for GA have substantial clinical potential since they can be used for the accurate prediction of GA and as useful surrogates for assessing developmental maturity (21). One of the main reasons why existing epigenetic GA clocks have had limited clinical utility thus far is the large number of CpGs needed to be assayed to achieve accurate prediction and the costly infrastructure needed to obtain DNA methylation data from cord blood DNA. The new epigenetic GA clock presented here, based on only five stably selected CpGs, is a significant methodological advance because it affords a

similar precision and accuracy as previous GA clocks while substantially curbing the number of CpGs needed to be tested.

Previously published GA clocks tended to overestimate the GA of preterm newborns (7, 8, 9). A similar tendency was also observed in the standard lasso-based clock developed in this study. One possible reason for this overestimation is the typically lower proportion of preterm compared to term newborns in the training sets. However, the Knight clock, which included a larger proportion of preterm newborns in the training set, also tended to overestimate the GA of preterm newborns (8). A key advantage of the stability selection framework over lasso and elastic net regression is that it separates the *feature selection* step from the *prediction* step. This enables taking nonlinear relationships into account by using methods such as GAM when building the prediction model (24). When using GAM to build the clock, the GA predictions for preterm newborns were improved compared to the scenario where only the lasso approach was used. Furthermore, for 12 of the 15 CpGs used to develop stable CpG clocks, the calculated EDF indicated a nonlinear relationship between DNAm and GA. These results suggest that at least some of the predictive CpGs exhibit a nonlinear relationship with GA, and that this may be important to account for, especially when applying epigenetic GA clocks to preterm newborns.

Several of the stably selected CpGs are in or near genes that have previously been linked to GA. One example is cg04347477 which had a 100% selection probability in our analysis. This CpG alone predicted GA with an R^2 of 0.52 and a MAD of 5.09 days in our test set. It is located in the promoter region of the nuclear corepressor 2 gene (*NCOR2*, formerly known as *SMRT*). CpGs in this gene have been identified in multiple EWASs of GA as well as in several GA clocks (4, 7, 9, 34, 37, 38). *NCOR2* encodes a nuclear receptor corepressor that facilitates transcriptional repression by recruiting histone deacetylase complexes (HDACs) and chromatin-remodeling factors (39, 40, 41). The role of *NCOR2* in GA is not clear, but the protein encoded by this gene is essential for a range of biological processes related to mammalian development (42, 43), regulation of inflammation (44, 45), and metabolic homeostasis and aging (46, 47, 48).

CpGs linked to the insulin-like growth factor 2 mRNA-binding protein 1 gene (*IGF2BP1*) have also been consistently associated with GA (4, 7, 9, 34, 37, 38). cg18183624, located within the promoter region of *IGF2BP1*, was assigned a selection probability of 0.996 in our stability selection analyses. *IGF2BP1* regulates the translation of specific genes by binding to their mRNAs and contributing to their stability and storage under both normal and stressful conditions (49). One of the genes regulated by *IGF2BP1* is *IGF2*, which is highly expressed *in utero* and is essential for fetal and placental growth (50). In addition, *IGF2BP1* is pivotal for the switch between fetal to adult hemoglobin, a process that occurs around birth (26, 51, 52).

Two of the CpGs found to be stably predictive of GA in our study, with a selection probability of 0.949 (cg20320200) and 0.807 (cg01833485), are linked to the estrogen-related receptor gamma gene (*ESRRG*). Like *NCOR2* and *IGF2BP1*, CpGs in or near *ESRRG* have also been identified in several other studies of GA (4, 7, 9, 37, 38). Estrogens are a group of steroid-based sex hormones that are involved in several important developmental and physiological processes, including cartilage proliferation and growth (53), skeletal muscle development and glucose homeostasis (54), and the development of both male and female reproductive tracts (55). *ESRRG* also plays a critical role in cardiac developmental maturation, particularly in directing and maintaining the metabolic switch from a predominant dependence on carbohydrates during prenatal life to a greater dependence on oxidative metabolism after birth (56, 57).

Furthermore, we recently showed that the association between DNAm and GA is highly cell-type specific and that most of the GA-associated CpGs were restricted to nucleated red blood cells (nRBCs) (38). However, when we searched for any overlap between the set of stably selected CpGs and the cell-type specific associations between DNAm and GA, most of the stably selected CpGs do not map to any specific cell type (Supplementary Table 5). The stably selected CpGs that were also found to be cell-type specific were either in nRBCs, granulocytes, or both, indicating that biological processes in these cell types may be particularly important for the relationship between DNAm and GA.

CONCLUSIONS

In summary, we identified 24 CpGs that were stably predictive of GA using a statistical framework for variable selection that combines subsampling with penalized regression. These CpGs were located in or near genes and regulatory regions that are relevant for immune responses, metabolism and developmental processes, including changes in hemoglobin expression and metabolic processes that occur in the transition from pre- to postnatal life. We showed that most CpGs in existing GA clocks are not stably selected and are not necessary for accurate prediction of GA. Furthermore, the use of GAM regression for GA prediction revealed that some of the predictive CpGs exhibit a nonlinear relationship with GA. Finally, we used the stably selected CpGs to construct a more parsimonious GA clock based on only five CpGs that showed a similar predictive performance as previous GA clocks, creating new opportunities for a more efficient use of DNAm-based GA estimations in research and clinical settings.

METHODS

Study population

Participants in this study are from the Norwegian Mother, Father, and Child Cohort Study (MoBa), an ongoing population-based pregnancy cohort study conducted by the Norwegian Institute of Public

Health (NIPH) (22). In total, approximately 114,500 children, 95,200 mothers, and 75,200 fathers were recruited from all over Norway from 1999 through 2008. The MoBa mothers consented to participation in 41% of the pregnancies. Extensive details on the MoBa cohort have been provided elsewhere (22, 58).

For this study, we used two subsamples of newborns for whom information on ultrasound-estimated GA was available: (i) dataset 1 ($n = 956$) and (ii) dataset 2 ($n = 1,186$). Both datasets are based on randomly selected cord-blood samples from the same source population (MoBa). As four individuals were included in both datasets, they were removed from one of the datasets (dataset 2) prior to analysis. The two datasets were then merged into a single dataset comprising a total of 2,138 newborns. Figure 8 provides an overview of the sample selection scheme and analysis flow. Detailed characteristics of the study participants and eligibility criteria for dataset 1 have been provided in our recent work (59). Dataset 2 was sampled in a similar way to make the datasets as compatible as possible.

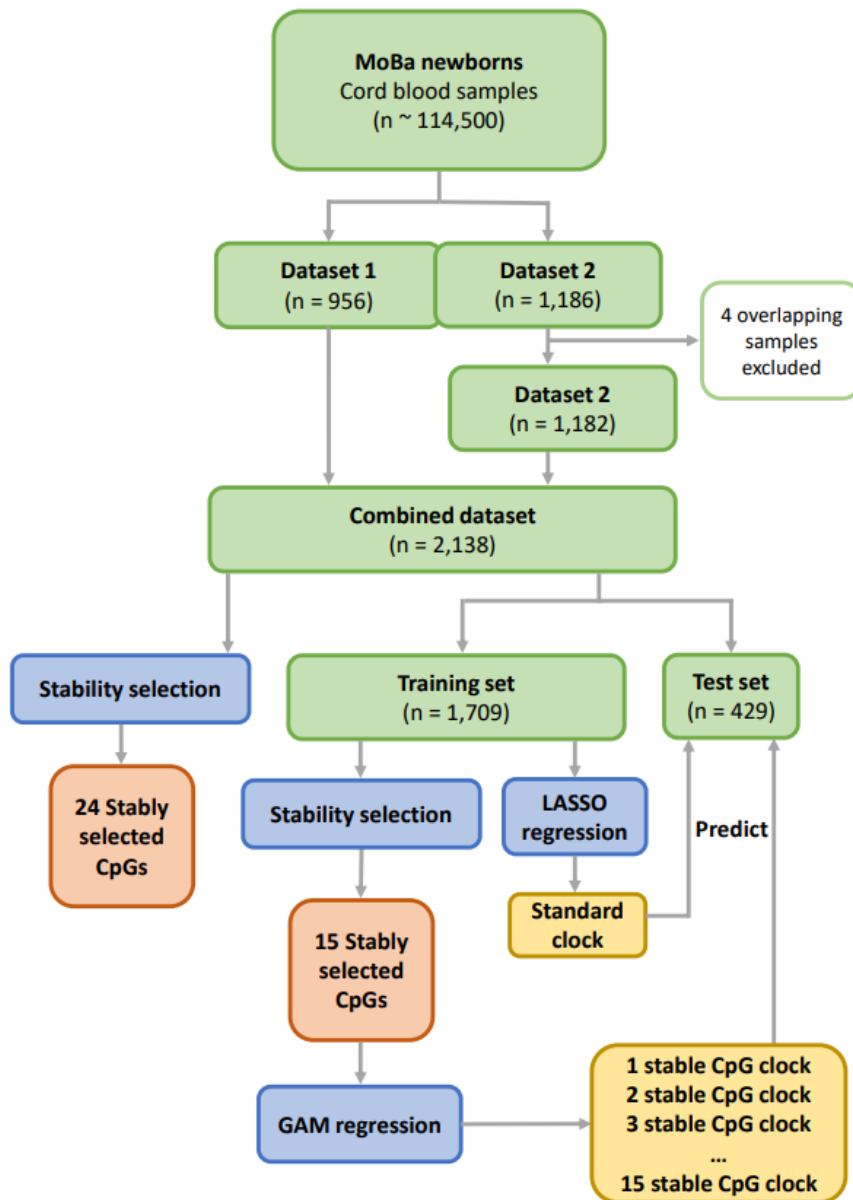


Figure 8. Overview of sample selection and analysis flow. Datasets are highlighted in green, methods in blue, analysis output in orange, and epigenetic clocks in yellow. Two randomly sampled subsets from MoBa (dataset 1 and dataset 2) were included in the current study. Data from four individuals that were present in both datasets were excluded from dataset 2. The two datasets were then merged into a single dataset ('combined dataset'). The samples from the combined dataset were randomly assigned to a training and test set. Stability selection was performed both on the combined dataset and the training set. Generalized additive model (GAM) regression was used to model the effect of the stably selected CpGs on gestational age (GA) to build clocks based on the stably selected CpGs. In parallel, lasso regression was performed directly on the training set to build a standard GA clock. The standard GA clock and the clocks based on the stably selected CpGs were used to predict GA in the test set.

DNAm profiling and quality control

Cord-blood samples were taken immediately after birth and kept frozen (58). The quality control procedures for dataset 1 have been extensively detailed in our previous work (59). Dataset 2 was processed using the same pipeline to make sure that the two datasets were as compatible as possible. Briefly, DNAm was measured at 885,000 CpG sites using the Illumina Infinium MethylationEPIC BeadChip version 1 (Illumina, San Diego, USA). The raw iDAT files were processed in four batches. Cross-hybridizing probes and probes that had a detection p value greater than 0.01 were excluded. Probes in which the last three bases overlapped with a single-nucleotide polymorphism (SNP) were also removed. BMIQ (60) was used to normalize type I and type II probe chemistries. Samples with low overall signals in control probes were removed after visual inspection, and samples with markedly different DNAm signals than the rest of the samples were also excluded. For consistency, CpG sites excluded from one batch due to poor quality and low detection p value were also removed from all subsequent batches. After quality control, 770,586 CpGs remained in dataset 1 and 795,171 CpGs in dataset 2. 769,139 CpGs were available for analysis in the combined dataset.

Variables

Information on GA and sex was extracted from the Medical Birth Registry of Norway (MBRN). GA at birth was estimated from ultrasound measurements around week 18 of pregnancy.

Penalized regression

We used lasso regression from the `glmnet` R package (61) to select CpGs that are predictive of GA in our samples. Ultrasound-based GA was regressed on the 769,139 CpGs in the combined dataset. Tuning parameter α was set to 1, while λ was selected after 10-fold cross-validation.

Stability selection of CpGs predictive of GA

We combined the stability selection framework proposed by Meinshausen and Bühlmann (20) with lasso regression to identify CpGs that were stably predictive of GA in our total sample of 2,138 newborns. By resampling the dataset multiple times, stability selection seeks to identify variables that are repeatedly chosen as predictors while simultaneously controlling the number of selected variables due to noise. We fitted a lasso model ($\lambda = 0.386$) as described above on a random subsample of $n/2$ ($n = 2,138$) and repeated this process 1000 times. We performed 1000 repetitions, 10 times more than the recommended number (20), because a higher number of repetitions increases the precision of the method. For each CpG, we computed the proportion of runs in which it was selected, which is referred to as the ‘selection probability’. Finally, we used the following formula (Theorem 1 from Meinshausen and Bühlmann (20)) to choose a threshold that determines the appropriate selection probability threshold for declaring a CpG as stably predictive of GA:

$$E(V) \leq \frac{q^2}{(2\pi_{thr} - 1)p}$$

$E(V)$ is the expected number of false discoveries in the stably selected set, q is the average number of variables (CpGs) selected by the variable selection method (here, lasso), and p is the total number of variables included in the analyses (here $n_{CpGs} = 769,139$).

The average number of selected CpGs (q) was found by repeating the stability selection procedure with permuted GA values and calculating the average number of CpGs selected ($q = 593.8$). We decided to allow up to two false discoveries on average, resulting in a probability threshold of 0.729. The above approach was repeated on a random subsample of 80% ($n = 1,709$) of our original sample of 2,138 newborns. This truncated dataset is referred to as the training set. The selected λ for the training set was 0.475 and the chosen probability threshold was 0.764 when allowing up to one false discovery on average ($q = 450.5$).

Predicting GA from DNAm

The CpGs that were declared stably predictive of GA in the above training set were subsequently used to create prediction models for GA. We used the `gam` function from the `mgcv` R package (62) to fit GAM models with GA as the response variable and the stably selected CpGs as the explanatory variables. The effect of each of the CpGs was modeled using a smooth spline. Supplementary Figure 1 shows the plots for the smooth functions for each of the 15 CpGs.

The output of the GAM regression was used to predict GA in the remaining 20% of our samples – the test set ($n = 429$). Predicted GA was then regressed on ultrasound-estimated GA using MM-type robust linear regression (63) from the R package `robustbase` (64). MM-type robust linear regression was used because it is less influenced by outliers than, for example, the ordinary least squares (OLS) regression method (65). The precision of a given prediction model was defined as the proportion of variance explained by the model (i.e., its R^2 value), while accuracy was defined as the median absolute deviation (MAD, in days) between ultrasound-based and predicted GA.

Downstream bioinformatics analyses of the selected CpGs

The R package `biomaRt` (66) was used to fetch annotations for each CpG from the *Ensembl* server (www.ensembl.org) (25), according to the GRCh37 version of the human genome. The *ensembl* regulatory IDs of the regulatory regions identified were then used to manually query the GeneHancer database (<https://www.genecards.org/>) (67). The genes predicted to be affected by these regulatory regions were then visually presented using the R package `karyoploteR` (68). In addition, we downloaded data from the EWAS catalog (31) and EWAS atlas (32) databases (as of Feb 16th, 2023) and searched for studies involving the stably selected CpGs identified in the current study.

Code availability

All statistical analyses were performed using R version 4.1.2 and 4.2.2 (69). The code used to perform the analyses, as well as R objects containing the stable CpG clocks developed in this study are available on GitHub at github.com/KristineLH/stabsel-clock.

DECLARATIONS

Ethics approval and consent to participate

The establishment of MoBa and the initial data collection were based on a license from the Norwegian Data Protection Agency and an approval from the Regional Committees for Medical and Health Research Ethics ('REK'). MoBa is regulated by the Norwegian Health Registry Act. The current study was approved by the REK Southeast (committee C, reference number: 21532).

Consent for publication

Written informed consents were obtained from the MoBa participants.

Availability of data and materials

Access to the DNAm datasets can be obtained by applying to the Norwegian Institute of Public Health (NIPH). Restrictions apply regarding the availability of these data, which were originally used under specific approvals for the current study and are therefore not publicly available. Access can only be given after approval by REK under the provision that the applications are consistent with the consent provided. An application form can be found on the NIPH website at <https://www.fhi.no/en/studies/moba/>. Specific questions regarding access to data in this study can also be directed to Dr. Siri E. Håberg (Siri.Haberg@fhi.no). The data generated in this study are provided in the Supplementary Information.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Research Council of Norway (RCN) through its Centre of Excellence funding scheme (project number 262700). The funding body did not play any role in the design of the study, collection, analysis, or interpretation of data, nor in writing the manuscript.

Authors' contributions

KLH and WRPD designed the research; CMP performed the quality control on the DNAm data; KLH and JR conducted the analyses; WRPD, YL and CMP supervised the statistical analyses; KLH, JR, YL, AJ, JB and WRPD interpreted the data; KLH, AJ and WRPD drafted the first version of the manuscript; AJ, PM, and SEH acquired funding, project administration, and resources. KLH, JR, YL,

CMP, PM, SEH, JB, AJ and WRPD provided scientific input, revised the manuscript, and approved the final version.

Acknowledgments

The Norwegian Mother, Father, and Child Cohort Study (MoBa) is supported by the Norwegian Ministry of Health and Care Services and the Ministry of Education and Research. We are deeply indebted to all the families in Norway who participate in this ongoing cohort study. The DNA samples were processed and subjected to DNA methylation measurements on Illumina Infinium EPIC arrays by Life and Brain GmbH in Bonn, Germany (dataset 1) and by Human Genomics Facility at Erasmus MC in Rotterdam, Netherlands (dataset 2). This work was partly performed using the Services for Sensitive Data (TSD) facilities at the University of Oslo, Norway.

REFERENCES

1. Wang K, Liu H, Hu Q, Wang L, Liu J, Zheng Z, et al. Epigenetic regulation of aging: implications for interventions of aging and diseases. *Signal Transduct Target Ther.* 2022;7(1):374.
2. John RM, Rougeulle C. Developmental Epigenetics: Phenotype and the Flexible Epigenome. *Front Cell Dev Biol.* 2018;6:130.
3. Villicaña S, Bell JT. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol.* 2021;22(1):127.
4. Merid SK, Novoloaca A, Sharp GC, Küpers LK, Kho AT, Roy R, et al. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. *Genome medicine.* 2020;12(1):25.
5. Day K, Waite LL, Thalacker-Mercer A, West A, Bamman MM, Brooks JD, et al. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol.* 2013;14(9):R102.
6. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature reviews Genetics.* 2018;19(6):371-84.
7. Bohlin J, Håberg SE, Magnus P, Reese SE, Gjessing HK, Magnus MC, et al. Prediction of gestational age based on genome-wide differentially methylated regions. *Genome Biol.* 2016;17(1):207.
8. Knight AK, Craig JM, Theda C, Bækvad-Hansen M, Bybjerg-Grauholm J, Hansen CS, et al. An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biol.* 2016;17(1):206.
9. Haftorn KL, Lee Y, Denault WRP, Page CM, Nustad HE, Lyle R, et al. An EPIC predictor of gestational age and its application to newborns conceived by assisted reproductive technologies. *Clin Epigenetics.* 2021;13(1):82.
10. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 2013;14(10):R115.
11. McEwen LM, O'Donnell KJ, McGill MG, Edgar RD, Jones MJ, MacIsaac JL, et al. The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proc Natl Acad Sci U S A.* 2020;117(38):23329-35.
12. 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.* 2019;11(12):4238-53.
13. Sørensen Ø, Hellton KH, Frigessi A, Thoresen M. Covariate Selection in High-Dimensional Generalized Linear Models With Measurement Error. *Journal of Computational and Graphical Statistics.* 2018;27(4):739-49.
14. Sørensen Ø, Frigessi A, Thoresen M. MEASUREMENT ERROR IN LASSO: IMPACT AND LIKELIHOOD BIAS CORRECTION. *Statistica Sinica.* 2015;25(2):809-29.
15. Engebretsen S, Bohlin J. Statistical predictions with glmnet. *Clin Epigenetics.* 2019;11(1):123.
16. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nature reviews Genetics.* 2011;12(8):529-41.
17. Dugué PA, English DR, MacInnis RJ, Joo JE, Jung CH, Milne RL. The repeatability of DNA methylation measures may also affect the power of epigenome-wide association studies. *International journal of epidemiology.* 2015;44(4):1460-1.

18. Nustad HE, Steinsland I, Ollikainen M, Cazaly E, Kaprio J, Benjamini Y, et al. Modeling dependency structures in 450k DNA methylation data. *Bioinformatics* (Oxford, England). 2021;38(4):885-91.
19. Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Res.* 2016;44(11):5123-32.
20. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010;72(4):417-73.
21. Knight AK, Conneely KN, Smith AK. Gestational age predicted by DNA methylation: potential clinical and research utility. *Epigenomics*. 2017.
22. Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *International journal of epidemiology*. 2016;45(2):382-8.
23. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996;58(1):267-88.
24. Wood SN. *Generalized Additive Models: An Introduction with R*. 2nd ed: Chapman and Hall/CRC; 2017.
25. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. *Nucleic Acids Research*. 2020;49(D1):D884-D91.
26. de Vasconcellos JF, Tumburu L, Byrnes C, Lee YT, Xu PC, Li M, et al. IGF2BP1 overexpression causes fetal-like hemoglobin expression patterns in cultured human adult erythroblasts. *Proc Natl Acad Sci U S A*. 2017;114(28):E5664-e72.
27. Kueh AJ, Dixon MP, Voss AK, Thomas T. HBO1 is required for H3K14 acetylation and normal transcriptional activity during embryonic development. *Mol Cell Biol*. 2011;31(4):845-60.
28. Cerdá-Esteban N, Spagnoli FM. Glimpse into Hox and tale regulation of cell differentiation and reprogramming. *Dev Dyn*. 2014;243(1):76-87.
29. Tran AH, Berger A, Wu GE, Paige CJ. Regulatory mechanisms in the differential expression of Hemokinin-1. *Neuropeptides*. 2009;43(1):1-12.
30. Xu Y, Shen J, Ran Z. Emerging views of mitophagy in immunity and autoimmune diseases. *Autophagy*. 2020;16(1):3-17.
31. Battram T, Yousefi P, Crawford G, Prince C, Sheikhalil Babaei M, Sharp G, et al. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res*. 2022;7:41.
32. Li M, Zou D, Li Z, Gao R, Sang J, Zhang Y, et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res*. 2019;47(D1):D983-d8.
33. Schroeder JW, Conneely KN, Cubells JC, Kilaru V, Newport DJ, Knight BT, et al. Neonatal DNA methylation patterns associate with gestational age. *Epigenetics*. 2011;6(12):1498-504.
34. Simpkin AJ, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, Ring SM, et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum Mol Genet*. 2015;24(13):3752-63.
35. Battram T, Gaunt TR, Relton CL, Timpson NJ, Hemani G. A comparison of the genes and genesets identified by GWAS and EWAS of fifteen complex traits. *Nature communications*. 2022;13(1):7816.
36. Simpkin AJ, Suderman M, Howe LD. Epigenetic clocks for gestational age: statistical and study design considerations. *Clin Epigenetics*. 2017;9:100.
37. Parets SE, Conneely KN, Kilaru V, Fortunato SJ, Syed TA, Saade G, et al. Fetal DNA Methylation Associates with Early Spontaneous Preterm Birth and Gestational Age. *PloS one*. 2013;8(6):e67489.
38. Haftorn KL, Denault WRP, Lee Y, Page CM, Romanowska J, Lyle R, et al. Nucleated red blood cells explain most of the association between DNA methylation and gestational age. *Communications Biology*. 2023;6(1):224.
39. Jepsen K, Rosenfeld MG. Biological roles and mechanistic actions of co-repressor complexes. *J Cell Sci*. 2002;115(Pt 4):689-98.
40. Perissi V, Jepsen K, Glass CK, Rosenfeld MG. Deconstructing repression: evolving models of co-repressor action. *Nature reviews Genetics*. 2010;11(2):109-23.
41. Jones PL, Shi YB. N-CoR-HDAC corepressor complexes: roles in transcriptional regulation by nuclear hormone receptors. *Curr Top Microbiol Immunol*. 2003;274:237-68.
42. Jepsen K, Solum D, Zhou T, McEvelly RJ, Kim HJ, Glass CK, et al. SMRT-mediated repression of an H3K27 demethylase in progression from neural stem cell to neuron. *Nature*. 2007;450(7168):415-9.
43. Jepsen K, Gleiberman AS, Shi C, Simon DI, Rosenfeld MG. Cooperative regulation in development by SMRT and FOXP1. *Genes & development*. 2008;22(6):740-5.
44. Ghisletti S, Huang W, Jepsen K, Benner C, Hardiman G, Rosenfeld MG, et al. Cooperative NCoR/SMRT interactions establish a corepressor-based strategy for integration of inflammatory and anti-inflammatory signaling pathways. *Genes & development*. 2009;23(6):681-93.
45. Barish GD, Yu RT, Karunasiri MS, Becerra D, Kim J, Tseng TW, et al. The Bcl6-SMRT/NCoR cistrome represses inflammation to attenuate atherosclerosis. *Cell Metab*. 2012;15(4):554-62.

46. Pei L, Leblanc M, Barish G, Atkins A, Nofsinger R, Whyte J, et al. Thyroid hormone receptor repression is linked to type I pneumocyte-associated respiratory distress syndrome. *Nat Med.* 2011;17(11):1466-72.
47. Nofsinger RR, Li P, Hong SH, Jonker JW, Barish GD, Ying H, et al. SMRT repression of nuclear receptors controls the adipogenic set point and metabolic homeostasis. *Proc Natl Acad Sci U S A.* 2008;105(50):20021-6.
48. Reilly SM, Bhargava P, Liu S, Gangl MR, Gorgun C, Nofsinger RR, et al. Nuclear receptor corepressor SMRT regulates mitochondrial oxidative metabolism and mediates aging-related metabolic deterioration. *Cell Metab.* 2010;12(6):643-53.
49. Huang H, Weng H, Sun W, Qin X, Shi H, Wu H, et al. Recognition of RNA N(6)-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nat Cell Biol.* 2018;20(3):285-95.
50. Sandovici I, Georgopoulou A, Pérez-García V, Hufnagel A, López-Tello J, Lam BYH, et al. The imprinted Igf2-Igf2r axis is critical for matching placental microvasculature expansion to fetal growth. *Developmental Cell.* 2022;57(1):63-79.e8.
51. Chambers CB, Gross J, Pratt K, Guo X, Byrnes C, Lee YT, et al. The mRNA-Binding Protein IGF2BP1 Restores Fetal Hemoglobin in Cultured Erythroid Cells from Patients with β -Hemoglobin Disorders. *Mol Ther Methods Clin Dev.* 2020;17:429-40.
52. Tumburu L, Byrnes C, Lee Y, de Vasconcellos J, Rabel A, Miller J. IGF2BP1 Reverses Hemoglobin Switching in Adult Erythroblasts 2015. 639- p.
53. Ahi EP. Signalling pathways in trophic skeletal development and morphogenesis: Insights from studies on teleost fish. *Dev Biol.* 2016;420(1):11-31.
54. Kim SO, Albrecht ED, Pepe GJ. Estrogen promotes fetal skeletal muscle myofiber development important for insulin sensitivity in offspring. *Endocrine.* 2022;78(1):32-41.
55. Cunha GR, Li Y, Mei C, Derpinghaus A, Baskin LS. Ontogeny of estrogen receptors in human male and female fetal reproductive tracts. *Differentiation.* 2021;118:107-31.
56. Sakamoto T, Matsuura TR, Wan S, Ryba DM, Kim JU, Won KJ, et al. A Critical Role for Estrogen-Related Receptor Signaling in Cardiac Maturation. *Circ Res.* 2020;126(12):1685-702.
57. Alaynick WA, Kondo RP, Xie W, He W, Dufour CR, Downes M, et al. ERR γ directs and maintains the transition to oxidative metabolism in the postnatal heart. *Cell Metab.* 2007;6(1):13-24.
58. Paltiel L, Anita H, Skjerden T, Harbak K, Bækken S, Nina Kristin S, et al. The biobank of the Norwegian Mother and Child Cohort Study – present status. *Norsk Epidemiologi.* 2014;24(1-2).
59. Håberg SE, Page CM, Lee Y, Nustad HE, Magnus MC, Haftorn KL, et al. DNA methylation in newborns conceived by assisted reproductive technology. *Nature communications.* 2022;13(1):1896.
60. 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).* 2013;29(2):189-96.
61. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software.* 2010;33(1):1 - 22.
62. Wood SN. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association.* 2004;99(467):673-86.
63. Yohai V. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics.* 1987;15.
64. Maechler M RP, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Conceicao EL, Anna di Palma M. robustbase: Basic Robust Statistics. R package 0.93-6 ed. <http://robustbase.r-forge.r-project.org/>. 2020.
65. Varin S, Panagiotakos DB. A review of robust regression in biomedical science research. *Arch Med Sci.* 2020;16(5):1267-9.
66. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc.* 2009;4(8):1184-91.
67. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database.* 2017;2017.
68. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics (Oxford, England).* 2017;33(19):3088-90.
69. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.