Epigenetic scores indicate differences in the proteome of preterm infants

Katie Mckinnon¹, Eleanor L.S. Conole², Kadi Vaher¹, Robert F. Hillary³, Danni A. Gadd³, Justyna Binkowska¹, Gemma Sullivan¹, Anna J. Stevenson³, Amy Corrigan¹, Lee Murphy⁵, Heather C. Whalley⁴, Hilary Richardson⁶, Riccardo E. Marioni³, Simon R. Cox², James P. Boardman¹,⁴

1. Centre for Reproductive Health, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh, UK.
2. Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, UK.
3. Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK.
4. Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK.
5. Edinburgh Clinical Research Facility, University of Edinburgh, Edinburgh, UK.
6. School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, UK.

Corresponding author: Professor James P Boardman
Institute for Regenerative Medicine, 4-5 Little France Drive, Edinburgh, EH16 4UU
James.Boardman@ed.ac.uk

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Background: Epigenetic scores (EpiScores), reflecting DNA methylation (DNAm)-based surrogates for complex traits, have been developed for multiple circulating proteins. EpiScores for pro-inflammatory proteins, such as C-reactive protein (DNAm CRP), are associated with brain health and cognition in adults and with inflammatory comorbidities of preterm birth in neonates. Social disadvantage can become embedded in child development through inflammation, and deprivation is over-represented in preterm infants. We tested the hypotheses that preterm birth and socioeconomic status (SES) are associated with alterations in a set of EpiScores enriched for inflammation-associated proteins.

Results: 104 protein EpiScores were derived from saliva samples of 332 neonates born at gestational age (GA) 22.14 to 42.14 weeks. Saliva sampling was between 36.57 and 47.14 weeks. Forty-three (41%) EpiScores were associated with low GA at birth (standardised estimates |0.14 to 0.88|, Bonferroni-adjusted p-value <8.3x10^{-3}). These included EpiScores for chemokines, growth factors, proteins involved in neurogenesis and vascular development, cell membrane proteins and receptors, and other immune proteins. Three EpiScores were associated with SES, or the interaction between birth GA and SES: afamin, intercellular adhesion molecule 5 and hepatocyte growth factor-like protein (standardised estimates |0.06 to 0.13|, Bonferroni-adjusted p-value <8.3x10^{-3}). In a preterm sub-group (n=217, median [range] GA 29.29 weeks [22.14 to 33.0 weeks]), SES-EpiScore associations did not remain statistically significant after adjustment for sepsis, bronchopulmonary dysplasia, necrotising enterocolitis, and histological chorioamnionitis.

Conclusions: Low birth GA is substantially associated with a set of EpiScores. The set was enriched for inflammatory proteins, providing new insights into immune dysregulation in preterm infants. SES had fewer associations with EpiScores; these tended to have small effect sizes and were not statistically significant after adjusting for inflammatory comorbidities. This suggests that inflammation is unlikely to be the primary axis through which SES becomes embedded in the development of preterm infants in the neonatal period.

Keywords: Epigenetic, Inflammation, Neonatal, Preterm birth, Socioeconomic status.
1.0 Background

Preterm birth (delivery at less than 37 weeks of gestation) affects around 10% of births worldwide and is closely associated with increased likelihood of cerebral palsy, neurocognitive impairment, behavioural, social and communication difficulties, and mental and cardiometabolic health diagnoses across the life course [1–5]. These adverse outcomes can be explained, in part, by deleterious effects of early exposure to extrauterine life on brain and cardiac development, and they are often accompanied by changes in blood proteins, including those reflecting the perinatal innate and adaptive immune response [6–8].

Socioeconomic status (SES) is also associated with the adverse neurodevelopmental and health outcomes listed above [9–12], and social deprivation is consistently over-represented among preterm children and their families [13,14]. In a meta-analysis of 43 studies (n=111,156 individuals), low SES associated with increased inflammatory markers of disease risk (C-reactive protein [CRP] and interleukin-6 [IL6]), which suggests that pro-inflammatory pathways may be important mechanisms for translating social inequalities into health disparities [15]. However, only four studies included participants under 10 years of age, leaving uncertainty about SES-inflammation correlations in early life [16–19].

Although protein levels are commonly used as biomarkers of exposure and disease risk, they are limited because they are often phasic in the systemic circulation, rely on venepuncture, and may not capture baseline status or chronicity. For example, inflammation is often measured using acute-phase inflammatory proteins such as CRP [20,21], but it is not always reliable [22], particularly in neonates, and a single-time-point measure may not reflect baseline inflammation or capture chronic inflammation [23]. These challenges have been addressed by the development of DNA methylation (DNAm) markers of protein expression (EpiScores), which are derived from a linear weighted sum of DNAm sites that are correlated with protein levels. Several EpiScores are associated with magnetic resonance imaging (MRI) measures of brain health, cognition, child mental health, stroke, ischaemic heart disease Alzheimer's disease, lung cancer [24–31]. In neonates, DNAm CRP is associated with birth gestational age (GA), perinatal inflammatory processes, and MRI features of encephalopathy of prematurity [32]. Childhood SES is associated with differential DNAm in inflammation-related genes.
[33,34] and at CpG sites that correlate with an inflammation index [35]. Adult SES and social mobility are associated with variations in DNA methylation (DNAm) in inflammation-related genes [33,36]. Importantly, SES-related DNA methylation variations are associated with differences in gene expression so may have functional consequences [33,36].

We investigated relationships between preterm birth, SES, and 104 EpiScores enriched for inflammation-related proteins [26–28,31,37]. We tested the following hypotheses: first, low GA is associated with differences in EpiScores; second, SES is correlated with EpiScores, and interacts with birth GA, but the relationship is attenuated by inflammatory disease burden in preterm infants.

2.0 Methods

2.1 Participants

Participants were preterm infants (born ≤33 weeks’ gestation) and term-born infants born at the Royal Infirmary of Edinburgh, UK. These infants were recruited to a longitudinal cohort study designed to investigate the effect of preterm birth on brain development and outcomes with multimodal data collection [38]. Infants were recruited between February 2012 and December 2021.

Exclusion criteria were congenital malformation, chromosomal abnormality, congenital infection, cystic periventricular leukomalacia, haemorrhagic parenchymal infarction, and post-haemorrhagic ventricular dilatation. These criteria mean the cohort is representative of the majority of survivors of modern intensive care practices [38].

Final participants included were 217 preterm infants (born ≤33 weeks’ GA) and 115 term-born infants, with median birth GA of 29.29 weeks and 39.71 weeks respectively. Their demographic characteristics are shown in Table 1. The three SES measures (Scottish Index of Multiple Deprivation (SIMD 2016) [39], maternal education and maternal occupation) differed between the preterm and term groups (Cohen’s d effect sizes 0.52–0.68). Ethnicity did not differ between groups and is representative of the Edinburgh area [40].

[Table 1 near here]
2.2 DNA methylation

Saliva samples for DNAm were collected at term-equivalent age using Oragene OG-575 Assisted Collection kits (DNA Genotek, ON, Canada) and DNA was extracted using prepIT.L2P reagent (DNA Genotek, ON, Canada). Saliva sampling was used due to accessibility and the non-invasiveness of the method; DNAm patterns measured via saliva samples correlate with brain and other tissue DNAm patterns [41,42]. We chose to sample at the term-corrected gestation timepoint to include the allostatic load of both prenatal and early postnatal exposures.

DNA was bisulphite converted and methylation levels were measured using Illumina HumanMethylationEPIC BeadChip (Illumina, San Diego, CA, USA) at the Edinburgh Clinical Research Facility (Edinburgh, UK). The arrays were imaged on the Illumina iScan or HiScan platform, and genotypes were called automatically using GenomeStudio Analysis software version 2011.1 (Illumina). DNAm was processed in four batches.

Raw intensity (.idat) files were read into the R environment using minfi. wateRmelon and minfi were used for preprocessing, quality control and normalisation [43]. The pfilter function in wateRmelon was used to exclude samples with 1% of sites with a detection $p$-value >0.05, sites with beadcount <3 in 5% of samples, and sites with 1% of samples with detection $p$-value >0.05. Cross hybridising probes, probes targeting single nucleotide polymorphisms with overall minor allele frequency $\geq 0.05$, and control probes were also removed. Samples were removed if there was a mismatch between predicted sex (minfi) and recorded sex ($n=3$), or if samples did not meet preprocessing quality control criteria ($n=29$). Data were datanet normalised, which includes background correction and dye bias correction [43]. Saliva contains different cells types, including buccal epithelial cells. Epithelial cell proportions were estimated with epigenetic dissection of intra-sample heterogeneity with the reduced partial correlation method implemented in the R package EpiDISH [44]. Probes located on sex chromosomes were removed before analysis. The cohort includes twins ($n=32$); these were randomly removed leaving one participant per twin pair. This left a final sample size of $n=332$.

2.3 EpiScore calculation
The 104 protein EpiScores included 100 EpiScores from Gadd et al [26], a study enriched for inflammatory-related proteins, excluding those where the required CpGs were not available, owing to differences in assay platform CpG coverage relative to Gadd et al [26]. For duplicate proteins, those developed using Olink platform-identified proteins (antibody-based assays) were prioritised over those from SOMAscan platforms (aptamer-based assays), due to specificity and reproducibility, and the variable correlation between the two methods [45–48] (for details see Supplementary eMethods, Additional File 1). In addition, we included EpiScores for growth and differentiation factor 15 (GDF15) and N-terminal-pro B-type natriuretic peptide (NTproBNP) [37], and CRP. The CRP EpiScore used was Barker et al’s seven-CpG variation of Ligthart et al’s CRP EpiScore [27,31], as this is known to correlate with birth GA, perinatal pro-inflammatory exposures, and neonatal brain development [32].

For each individual, EpiScores were obtained by multiplying the methylation proportion at a given CpG by the effect size from previous studies. This was performed using the MethylDetectR platform [49] for those inflammatory proteins currently included and using R for those not currently included (CRP, GDF15, IL6, NTproBNP). All CpG sites and coefficients required to calculate the 104 EpiScores are in Supplementary Table 1 (Additional File 2).

2.4 Statistics

The predictor variables were SES and birth GA. SES was operationalised in three ways: neighbourhood-level SES using the Scottish Index of Multiple Deprivation (SIMD) [39], and two measures of family-level SES, which were maternal education (highest educational qualification) and maternal occupation (current or most recent occupation). For further details, see Supplementary eMethods, Additional File 1. Birth GA was a continuous variable to maximise statistical power [50,51]. We adjusted for GA at saliva sampling, DNAm batch, infant sex, and birthweight z-score.

All statistical analyses were performed in R (version 4.3.1) and were preregistered [52].

Principal component analysis (PCA) was used to determine the significance threshold for controlling type 1 error in analyses of multiple EpiScores [53]. We began with a correlation analysis, which showed correlation coefficients between EpiScores of |0.01 to 0.93| (Figure 1A). To determine the
number of statistical “families” among the 104 EpiScores, PCA was performed. This yielded two
principal components with eigenvalues >1, our pre-specified threshold, which explained 59.5% and
17.2% of variance, respectively (Figure 1B). Standardised component loadings are provided in
Supplementary Table 3 (Additional File 1). In all subsequent analyses, we corrected for multiple
comparisons across EpiScores and SES measures using a Bonferroni-adjusted $p$-value threshold of
8.3x10$^{-3}$. This is 0.05/(2 x 3), with two reflecting the two principal components for Episcores and three
reflecting the number of SES measures used.

We constructed generalised linear regression models for each EpiScore as outcome measure to
assess associations between GA, each of the three SES measures (separate models for each of
SIMD, maternal education, and maternal occupation), and the product interaction term SES*birth GA
(removing the term if not significant), and adjusting for GA at sampling, sex, and batch.

For the preterm sub-group, we then additionally adjusted for perinatal inflammatory exposures known
to be associated with the CRP EpiScore as, to our knowledge, this is the only DNAm proxy of an
inflammatory protein that has been studied in this context [32]. These were histological
chorioamnionitis (HCA), sepsis, bronchopulmonary dysplasia (BPD) and necrotising enterocolitis
(NEC). For definitions see Supplementary eMethods (Additional File 1) and for frequencies see
Supplementary Table 2 (Additional File 1).

3.0 Results

3.1 Associations between gestational age and EpiScores
Gestational age associated with 43 of the 104 EpiScores after adjustment for SIMD, maternal
education, or maternal occupation (Figure 2A-C). The proteins represented by the 43 EpiScores are
listed in Table 2 categorised by functional annotation adapted from the STRING database[54], and
their broader roles in immune processes and inflammation, and the pathogenesis of neonatal
diseases, where known, are described in Supplementary Table 4 (Additional File 1).
29 (67%) EpiScores negatively associated with birth GA (standardised estimates |0.14-0.76|, adjusted 
$p$-value <8.3x10$^{-3}$), and 14 (33%) EpiScores positively associated with birth GA (standardised 
estimates 0.14-0.88, adjusted $p$-value <8.3x10$^{-3}$).

33 EpiScores associated with low GA irrespective of SES measure used in the model. The results for 
all 104 EpiScores are provided in Supplementary Figures 1-9 (Additional File 1).

3.2 Associations between SES, EpiScores, and the effect of inflammatory comorbidities of preterm 
birth.

Three out of 104 EpiScores associated with SES measures or the interaction between SES and GA 
(Figure 3). There was a small effect size association between higher afamin EpiScore and higher 
maternal occupation (standardised $\beta$=0.06, 95% confidence interval (CI) 0.02-0.11, $p$=0.0082), and 
DNAm afamin associated with the birth GA*maternal education interaction term such that afamin 
positively correlated with birth GA among babies with mothers without university education, and 
negatively correlated with birth GA among babies with mothers with university education 
(undergraduate or postgraduate) (standardised $\beta$=-0.12, 95% CI -0.20 to -0.04, $p$=0.0041, 
Supplementary Figure 10, Additional File 1).

Higher intercellular adhesion molecule 5 (ICAM5) EpiScore associated with higher SIMD 
(standardised $\beta$=0.13, 95% CI 0.03-0.23, $p$=0.0079). Hepatocyte growth factor-like protein (HGFI) 
associated with the birth GA*maternal occupation interaction term (standardised $\beta$=-0.10, 95% CI -
0.16- -0.04, $p$=0.0021, Supplementary Figure 11, Additional File 1), such that HGFI EpiScore 
positively correlated with birth GA among babies with mothers who were unemployed, homemakers, 
in full time education, or in unskilled or manual occupations, but negatively correlated with birth GA 
among babies with mothers in partly skilled, non-manual skilled or professional occupations.

In the planned analysis of preterm-born babies only, when controlling for inflammatory exposures 
(sepsis, HCA, NEC, and BPD), a greater proportion of $R^2$ was explained ($R^2$=0.053-0.28 in unadjusted
models, $R^2=0.029-0.437$ in adjusted models), but the EpiScores no longer met our statistical threshold ($p$-values $>0.045$ with adjusted $p$-value threshold $<8.3 \times 10^{-3}$, see Supplementary Table 5, Additional File 1).

[Figure 3 near here]

4.0 Discussion

In this study, we identified several associations between a set of EpiScores enriched for inflammatory proteins and low GA at birth. Few EpiScores associated with SES within the whole sample and these associations were partially attenuated in preterm infants who experienced inflammatory comorbidities. This is the first study to assess the impact of preterm birth and social status using epigenetic signatures designed to reflect the circulating proteome.

4.1 Associations between birth GA and EpiScores

43 EpiScores associated with preterm birth when sampled at term equivalent age. The EpiScores reflect chemokines, growth factors, proteins required for neurogenesis and vascular development, cell membrane proteins and receptors, and immune response proteins (Table 2). As well as having specific immunoregulatory roles, in the neonatal setting or relevant animal models, these proteins are associated with several comorbidities and developmental consequences of preterm birth. These include lung development and disease such as BPD [55–67], in utero and postnatal growth failure [68–71], HCA [7,72,73], patent ductus arteriosus [74–77], retinopathy of prematurity [78–82], NEC [83–85], hyperglycaemia [86], sepsis [85,87–90], brain injury [32,91–94], and neurodevelopmental outcomes [95–99].

4.2 Associations between SES measures and EpiScores

SES appears to play a much smaller role in the patterning of EpiScores compared to birth GA. SES measures, or interactions between birth GA and SES measures, were associated with only three of the 104 EpiScores studied: afamin, ICAM5 and HGFI. Afamin and ICAM5 positively associated with maternal occupation and SIMD, respectively. Afamin and HGFI both associated with the interaction term between SES and birth GA. Afamin is involved in vitamin E transport [100], ICAM5 has a role in
microglial regulation [101], and HGFI is a macrophage-stimulating protein [102]. The relationships between these proteins and SES have not previously been investigated, although afamin is associated with the development of metabolic syndrome [103], which varies with SES [104].

Among preterm infants, no SES-EpiScore associations survived adjustment for inflammatory exposures, which suggests that the weak effects of SES on the neonatal proteome that we observed in a small number of EpiScores are at least partially accounted for by inflammatory pathologies in early life. Taken together, the results suggest immune dysregulation, proxied by EpiScores, may not be the primary axis through which SES becomes embedded in the development of preterm infants during neonatal intensive care.

SES has been consistently associated with inflammation in adulthood, including in relation to childhood deprivation [15,105], but less is known about the relationship between SES and inflammation in the neonatal period. A longitudinal study by Leviton et al [106], with five sampling timepoints during the first month of life after preterm birth, showed that maternal eligibility for Medicaid associated with levels of 14 inflammatory proteins (IL6R, TNFR1, TNFR2, IL8, ICAM1, VCAM1, TSH, EPO, bFGF, IGFI, VEGF, PIGF, Ang-1, Ang-2). However, only three were significant at more than one timepoint during the month after preterm birth (TSH, bFGF, Ang-1), and none associated at all five measurement timepoints. These studies, taken together with our results, suggest that the impact of SES on immune regulation is relatively modest and inconsistent in the newborn period but accrues through to adulthood. Further research is required to understand how and when SES becomes embedded in child development and whether early life events such as preterm birth modify that process; EpiScores could be a powerful tool for investigating the temporal dynamics of social determinants of child health.

4.3 Strengths and limitations

Strengths of this study include the large sample of term and preterm neonates; to the best of our knowledge, this is the first examination of multiple DNA methylation-based estimators of circulating proteins in a neonatal sample. We derived EpiScores from minimally-invasive sampling (buccal cells from saliva) which overcomes the ethical challenge of venepuncture for research in children.
EpiScores, serving as proxies of inflammatory proteins and sampled at term-equivalent age in preterm infants, were selected for their potential to capture chronic, cumulative inflammation associated with preterm birth and neonatal intensive care exposures [23–25]. We adjusted for variables associated with DNAm, and additionally for inflammatory exposures to increase the clinical validity of our results.

The study has some limitations. The EpiScores used were developed in adult cohorts [26–28,37] and have not been validated in neonatal populations with neonatal protein levels, although we have previously established that neonatal DNAmCRP scores correlate with cumulative inflammatory exposures [32]. Further studies that evaluate serial blood protein levels with EpiScores could be informative but would need to rely on small volume samples taken during time of venepuncture for clinical reasons, given the practical and ethical challenges of serial venepuncture for research purposes in neonates. The EpiScores were also trained using blood samples [26–28,37], whereas we have projected these scores into saliva samples. However, previous studies have successfully used similar cross-tissue techniques [32,107,108], and in neonates saliva provides a non-invasive and accessible sample method. Not all inflammatory-related proteins are represented, as we were limited by available EpiScores, so we may have underestimated the full complexity of the relationship between birth GA, SES, and inflammation. Longitudinal investigations are imperative for elucidating whether the DNA methylation signatures associated with gestational age identified in this study exert a causal influence on the inflammation-associated mechanisms in preterm birth. It remains crucial to discern whether these signatures represent a direct downstream consequence of GA itself or are induced by specific factors correlated with GA, yet not necessarily driven by chronic inflammation. Mendelian randomization studies, integrating genomic and epigenomic determinants, are a promising methodological approach to disentangle the directionality of these intricate relationships. The study population is comparable to other neonatal populations in high-income, majority-white settings, but these results may not generalise to settings with different socioeconomic or ethnicity profiles. We studied several measures of SES but were not able to include all that could be relevant, such as household income.

5.0 Conclusion
We identified 43 EpiScores enriched for inflammatory proteins that associated with low birth GA. These 43 proteins offer novel insights into the physiological response to preterm birth and warrant further study to explore their role in the relationship between preterm birth, inflammation, and longer-term outcomes. We found only three EpiScores associated with SES in the neonatal period, none of which survived adjustment for perinatal pro-inflammatory exposures, suggesting that inflammation is unlikely to be the primary axis through which SES becomes embedded in the development of preterm infants in the neonatal period.

6.0 Declarations

6.1 Ethics approval and consent to participate
Ethical approval for the study was obtained from the National Research Ethics Service, South-East Scotland Research Ethics Committee (REC 11/55/0061, 13/SS/0143, 16/SS/0154), and NHS Lothian Research and Development (2016/0255). Written informed consent was obtained from parents.

6.2 Consent for publication
Not applicable.

6.3 Availability of data and materials
DNA methylation data are available to researchers subject to the terms of the Data Access Policy: https://www.ed.ac.uk/centre-reproductive-health/tebc/about-tebc/for-researchers/data-access-collaboration.

6.4 Competing interests
LM has received speaker and consultancy fees from Illumina. REM is a scientific advisor to the Epigenetic Clock Development Foundation and to Optima Partners. All other authors declare that they have no competing interests.

6.5 Funding
This work was supported by Theirworld (www.theirworld.org) and a UKRI MRC Programme Grant MR/X003434/1. KM receives salary from NHS Scotland. GS is funded by an MRC Clinician Scientist...
Fellowship MR/X019535/1. SRC is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (221890/Z/20/Z).

6.6 Authors’ contributions
KM: Conceptualisation, methodology, formal analysis, writing (original draft), writing (review/editing), visualization. ELSC: Formal analysis, writing (review/editing). KV: Formal analysis, writing (review/editing), visualization. RH: Methodology, formal analysis, writing (review/editing). DG: Methodology, formal analysis, writing (review/editing). JB: Formal analysis, writing (review/editing). GS: Investigation, writing (review/editing). AJS: Methodology, writing (review/editing). AC: Investigation, resources, data curation, writing (review/editing). LM: Investigation, resources, writing (review/editing). HCW: Methodology, writing (review/editing). HR: Conceptualisation, writing (original draft), writing (review/editing), supervision. REM: Conceptualisation, methodology, writing (original draft), writing (review/editing), supervision. SRC: Conceptualisation, methodology, writing (original draft), writing (review/editing), supervision. JPB: Conceptualisation, methodology, writing (original draft), writing (review/editing), supervision, project administration, funding acquisition.

6.7 Acknowledgements
The authors are grateful to the families who consented to take part in the study.
Figure legends

Figure 1. Determining significance threshold
Principal component analysis was used to determine the adjusted statistical significance threshold, given multiple statistical comparisons. Figure 1A shows a correlation matrix of 104 EpiScores, showing correlation coefficient as red for positive and blue negative associations when significant (p<0.05). Figure 1B shows a scree plot of principal components, with the eigenvalues for each component. Standardised component loadings for principal components one and two are provided in Supplementary Table 3 (Additional File 1).

Figure 2. EpiScores associated with gestational age in regression models adjusted for socioeconomic status
EpiScores associated with gestational age in regression models with (A) Scottish Index of Multiple Deprivation, (B) maternal education, and (C) maternal occupation. Figure 2A (n=331) shows 39 associations, Figure 2B (n=323) shows 35 associations, and Figure 2C shows (n=328) shows 39 associations. Points and bars represent standardised beta and 95% confidence intervals, with red indicating positive and blue negative associations. Covariates included in all models: age at sample, birthweight z-score, sex, and methylation processing batch. Bonferroni-adjusted p-value <8.3x10⁻³.

protein 5, VCAM1 – Vascular cell adhesion protein 1, VEGFA – Vascular endothelial growth factor A,
WFIKKN2 – WAP Kazal immunoglobulin Kunitz and NTR domain-containing protein 2.

Figure 3. EpiScores associated with socioeconomic status or an interaction between socioeconomic
status and birth gestational age
EpiScores associated with socioeconomic status (Scottish Index of Multiple Deprivation, maternal
education, or maternal occupation), or with an interaction between socioeconomic status and birth
gestational age. Regression models with gestational age at birth, gestational age at sample,
birthweight z-score, sex, and methylation processing batch. Sample sizes: for the Scottish Index of
Multiple Deprivation n=331, for maternal education n=323, and for maternal occupation n=328. 3/104
EpiScores were significant (Bonferroni-adjusted $p$-value <8.3x10^-3). Points and bars represent
standardised beta and 95% confidence intervals, with red indicating positive and blue negative
associations.
CI – confidence interval, GA – gestational age, HGFI – Hepatocyte growth factor-like protein alpha
chain, ICAM5 – Intercellular adhesion molecule 5, SIMD – Scottish Index of Multiple Deprivation.
## Table 1. Participant characteristics

<table>
<thead>
<tr>
<th>Demographic measure</th>
<th>Preterm (n=217)</th>
<th>Term (n=115)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>114/217 (52.5%)</td>
<td>64/115 (55.7%)</td>
</tr>
<tr>
<td>Female</td>
<td>103/217 (47.5%)</td>
<td>51/115 (44.3%)</td>
</tr>
<tr>
<td><strong>Birth GA (weeks) - median (range)</strong></td>
<td>29.29 (22.14-33.0)</td>
<td>39.71 (37.0-42.14)</td>
</tr>
<tr>
<td><strong>Birthweight (g) - median (range)</strong></td>
<td>1200 (370-2510)</td>
<td>3450 (2346-4670)</td>
</tr>
<tr>
<td><strong>Birthweight z-score - median (range)</strong></td>
<td>0.10 (-3.13-2.07)</td>
<td>0.43 (-2.3-2.96)</td>
</tr>
<tr>
<td><strong>SIMD rank - median (range)</strong></td>
<td>3720 (6-6966)</td>
<td>5344 (267-6967)</td>
</tr>
<tr>
<td><strong>Maternal ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African</td>
<td>1/217 (0.5%)</td>
<td>0/115 (0%)</td>
</tr>
<tr>
<td>Bangladeshi</td>
<td>0/217 (0%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>Caribbean</td>
<td>0/217 (0%)</td>
<td>0/115 (0%)</td>
</tr>
<tr>
<td>Chinese</td>
<td>0/217 (0%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>Indian</td>
<td>3/217 (2.4%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>Pakistani</td>
<td>4/214 (1.8%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>White</td>
<td>195/217 (89.9%)</td>
<td>106/115 (92.2%)</td>
</tr>
<tr>
<td>White/Asian</td>
<td>2/217 (0.9%)</td>
<td>0/115 (0%)</td>
</tr>
<tr>
<td>White/Black African</td>
<td>0/217 (0%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>White/Black Caribbean</td>
<td>1/217 (0.5%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>Other Asian</td>
<td>1/217 (0.5%)</td>
<td>2/115 (1.7%)</td>
</tr>
<tr>
<td>Other ethnic group</td>
<td>6/217 (2.8%)</td>
<td>0/115 (0%)</td>
</tr>
<tr>
<td>Other mixed ethnic background</td>
<td>4/217 (1.8%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td><strong>Maternal education</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>7/208 (3.4%)</td>
<td>0/115 (0%)</td>
</tr>
<tr>
<td>Basic high school qualification (1-4)</td>
<td>5/208 (2.4%)</td>
<td>2/115 (1.7%)</td>
</tr>
<tr>
<td>Basic high school qualification (&gt;4)=5</td>
<td>8/208 (3.8%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>Advanced high school</td>
<td>32/208 (15.4%)</td>
<td>3/115 (2.6%)</td>
</tr>
<tr>
<td>Qualification</td>
<td>Batch 1</td>
<td>Batch 2</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>College qualification</td>
<td>46/208 (22.1%)</td>
<td>8/115 (7.0%)</td>
</tr>
<tr>
<td>University undergraduate</td>
<td>61/208 (29.3%)</td>
<td>50/115 (43.5%)</td>
</tr>
<tr>
<td>University postgraduate</td>
<td>49/208 (23.6%)</td>
<td>51/115 (44.3%)</td>
</tr>
<tr>
<td>Maternal occupation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployed</td>
<td>12/213 (5.6%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>Homemaker</td>
<td>10/213 (4.7%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>Still in full time education</td>
<td>9/213 (4.2%)</td>
<td>1/115 (0.9%)</td>
</tr>
<tr>
<td>Sheltered employment</td>
<td>1/213 (0.5%)</td>
<td>0/115 (0%)</td>
</tr>
<tr>
<td>Unskilled</td>
<td>11/213 (5.2%)</td>
<td>3/115 (2.6%)</td>
</tr>
<tr>
<td>Partly skilled</td>
<td>7/213 (3.3%)</td>
<td>2/115 (1.7%)</td>
</tr>
<tr>
<td>Manual skilled</td>
<td>25/213 (11.7%)</td>
<td>9/115 (7.8%)</td>
</tr>
<tr>
<td>Non-manual skilled</td>
<td>44/213 (20.7%)</td>
<td>11/115 (9.6%)</td>
</tr>
<tr>
<td>Professional</td>
<td>94/213 (44.1%)</td>
<td>87/115 (75.7%)</td>
</tr>
<tr>
<td>Sample GA (weeks) - median (range)</td>
<td>40.57 (36.57-45.86)</td>
<td>42.0 (39.86-47.14)</td>
</tr>
</tbody>
</table>

Notes:
1. Preterm n=216, term n=115
2. “Other Asian” includes Sri Lankan (n=1), Malay (n=1)
3. “Other ethnic group” includes Arab (n=1), Iraqi (n=1), white Bulgarian (n=1), Fijian (n=1), Japanese (n=1), Hong Kong (n=1).
4. “Other mixed ethnic background” includes Pakistani/Scottish (n=2), British/Arab (n=1), Sri Lankan/Black (n=1), Sri Lankan/Indian (n=1).
5. GA = gestational age, SIMD = Scottish index of multiple deprivation.
<table>
<thead>
<tr>
<th>Chemokines</th>
<th>Growth factors</th>
<th>Neurogenesis</th>
<th>Vascular development</th>
<th>Cell membrane proteins / receptors</th>
<th>Other immune response</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCL11</td>
<td>FGF21</td>
<td>NCAM1</td>
<td>SKR3</td>
<td>Afamin</td>
<td>Complement C9</td>
</tr>
<tr>
<td>CCL18</td>
<td>GDF15</td>
<td>Semaphorin 3E</td>
<td>VCAM1</td>
<td>CD5L</td>
<td>CRP</td>
</tr>
<tr>
<td>CCL21</td>
<td>GHR</td>
<td>SLITRK5</td>
<td>VEGFA</td>
<td>CD6</td>
<td>CRTAM</td>
</tr>
<tr>
<td>CCL22</td>
<td>HGF</td>
<td></td>
<td></td>
<td>CD48</td>
<td>FAP</td>
</tr>
<tr>
<td>CCL25</td>
<td>IGFBP4</td>
<td></td>
<td></td>
<td>CD163</td>
<td>HClII</td>
</tr>
<tr>
<td>CXCL9</td>
<td>SCGF alpha</td>
<td></td>
<td></td>
<td>CD209</td>
<td>L-selectin</td>
</tr>
<tr>
<td>CXCL10</td>
<td>SCGF beta</td>
<td></td>
<td></td>
<td>Contactin-4</td>
<td>Lactotransferrin</td>
</tr>
<tr>
<td></td>
<td>WFIKKN2</td>
<td></td>
<td></td>
<td>FcRL2</td>
<td>MMP9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FcGR3B</td>
<td>S100A9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PIGR</td>
<td>Sialoadhesin</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Thrombopoietin receptor</td>
<td>Trypsin-2</td>
</tr>
</tbody>
</table>

4 EpiScores associated with birth gestational age in regression models adjusted for socioeconomic status. Roles adapted from the STRING database [54].

See Supplementary Table 4, Additional File 1, for further details of the functional roles of each protein, including roles in immunity and inflammation, and in preterm infants specifically.

CCL11 – C-C chemokine 11, CCL18 – C-C chemokine 18, CCL21 – C-C chemokine 21, CCL22 – C-C chemokine 22, CCL25 – C-C chemokine 25, CD5L – CD5 antigen-like protein, CD6 – T-cell differentiation antigen, CD163 – scavenger receptor cysteine-rich type 1 protein M130, CRP – C-reactive protein, CRTAM – Cytotoxic and regulatory T-cell molecule, CXCL9 – C-X-C motif chemokine 9, CXCL10 – C-X-C motif chemokine 10, FAP – Fibroblast Activation Protein alpha, FcGR3B – Low affinity immunoglobulin gamma Fc region receptor III-B, FcRL2 – Fc receptor-like protein 2, FGF21 – Fibroblast growth factor...
References


22


81. The Royal College of Ophthalmologists. Treating Retinopathy of Prematurity in the UK. The Royal College of Ophthalmologists; 2022.


