BMJ Open

Prospective longitudinal study of the pregnancy DNA methylome: the US Pregnancy, Race, Environment, Genes (PREG) study

Dana M Lapato,[1] Sara Moyer,[2] Emily Olivares,[3] Ananda B Amstadter,[4] Patricia A Kinser,[5] Shawn J Latendresse,[6] Colleen Jackson-Cook,[1,7] Roxann Roberson-Nay,[4] Jerome F Strauss,[2] Timothy P York[1,2]

To cite: Lapato DM, Moyer S, Olivares E, *et al*. Prospective longitudinal study of the pregnancy DNA methylome: the US Pregnancy, Race, Environment, Genes (PREG) study. *BMJ Open* 2018;**8**:e019721. doi:10.1136/bmjopen-2017-019721

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-019721).

Received 20 September 2017
Revised 15 January 2018
Accepted 16 February 2018

For numbered affiliations see end of article.

Correspondence to
Professor Timothy P York;
timothy.york@vcuhealth.org

ABSTRACT

Purpose The goal of the Pregnancy, Race, Environment, Genes study was to understand how social and environmental determinants of health (SEDH), pregnancy-specific environments (PSE) and biological processes influence the timing of birth and account for the racial disparity in preterm birth. The study followed a racially diverse longitudinal cohort throughout pregnancy and included repeated measures of PSE and DNA methylation (DNAm) over the course of gestation and up to 1 year into the postpartum period.

Participants All women were between 18 and 40 years of age with singleton pregnancies and no diagnosis of diabetes or indication of assisted reproductive technology. Both mother and father had to self-identify as either African-American (AA) or European-American (EA). Maternal peripheral blood samples along with self-report questionnaires measuring SEDH and PSE factors were collected at four pregnancy visits, and umbilical cord blood was obtained at birth. A subset of participants returned for two additional postpartum visits, during which additional questionnaires and maternal blood samples were collected. The pregnancy and postpartum extension included n=240 (AA=126; EA=114) and n=104 (AA=50; EA=54), respectively.

Findings to date One hundred seventy-seven women (AA=89, EA=88) met full inclusion criteria out of a total of 240 who were initially enrolled. Of the 63 participants who met exclusion criteria after enrolment, 44 (69.8%) were associated with a medical reason. Mean gestational age at birth was significantly shorter for the AA participants by 5.1 days (M=272.5 (SD=10.5) days vs M=277.6 (SD=8.3)).

Future plans Future studies will focus on identifying key environmental factors that influence DNAm change across pregnancy and account for racial differences in preterm birth.

Strengths and limitations of this study

- Pregnancy, Race, Environment, Genes encompassed approximately 1100 person time points across the gestational and postpartum periods, making it the one of the largest longitudinal studies of preterm birth to incorporate genetic, epigenetic and environmental measures.
- The experimental design provides traction for the testing of causal hypotheses on the contribution of risk to preterm birth and major depression in the peripartum.
- Both the environmental exposure and biological data are multidimensional: environmental exposure data incorporates both objective and self-report exposure measures, and the biological data include genome-wide DNA methylation (DNAm) measurements, gene expression profiles, telomere length and micronuclei frequency.
- Due to the resources allocated to increase the depth of phenotyping, the number of individual women followed is modest; however, even large cross-sectional studies are not amenable to testing causal mechanisms.
- DNAm measurement with the Illumina Infinium HumanMethylation 450 BeadChip interrogates a relatively small fraction of all CpG sites in the genome; however, the probes target regions of known functional importance, and the 450 k is one of the most frequently used platforms, which will facilitate standardised comparisons with other studies.

INTRODUCTION

Background

Preterm birth (PTB; <37 completed weeks of gestation) represents one of the most significant concerns for perinatal health.[1] PTB is the leading cause of infant mortality and has been associated with a large number of negative consequences, including higher rates of cerebral palsy, respiratory illness, feeding difficulties, neurological disabilities, vision problems and learning difficulties.[1] PTB also represents a persistent health disparity with African-American (AA) women in the USA being at a significantly higher risk to experience PTB. A large number of social and environmental determinants of health (SEDH) have been suspected as risk factors for PTB, but socioeconomic models have failed to account for the differences in PTB rates. One

possible reason that efforts to account for racial differences in PTB rates using non-genetic measures have been unsuccessful[2] could be that research on SEDH and biological processes has been disconnected in the PTB literature. Genetic studies should be directed to understand racial differences in the sociocultural sources of environmental heterogeneity that exist within and between races.[3]

Substantial evidence from twin and family studies points to both genetic and environmental factors contributing to the risk liability of PTB, but the contribution of each depends on maternal self-reported race.[3] This difference is not surprising given that heritability estimates assess the relative importance of genes and environments, are population-specific and are not necessarily constant across populations, especially if there are known disparities in environmental exposures. For example, although height is a highly heritable trait, observed differences in mean height between populations are largely attributable to environmental differences such as diet and quality of health care.[4] Population differences in terms of genetic and environmental contributions can reveal factors responsible for racial differences in health outcomes. Recently, biometrical models that simultaneously account for both genetic and environmental factors have shown environmental sources contributed 3.1 times more to the risk liability of PTB in AA compared with European-Americans (EA).[3]

In conjunction with these results, multiple lines of evidence support the primacy of social inequities for racial health disparities, including PTB rates.[5] The PTB rate in 2003 for the US-born AA was 18.2% but 13.9% for foreign-born AA, which is a rate similar to that of EA.[6] Because the US-born and foreign-born AA women are expected to share similar genetic ancestry, these results support, in part, a sociodemographic explanation for the disparate PTB rates between foreign-born and the US-born births. Broad categories of non-genetic factors thought to contribute to racial disparities include social factors, such as maternal education; marital status; stressful life events, such as maternal exposure to financial, partner-associated or traumatic stress; racism and environmental factors including pollution, water quality, tobacco exposure and diet.[7–9]

Despite insight into risk factors that could be contributing to PTB liability, past research has failed to elucidate causal mechanisms integral to PTB pathophysiology, in part due to incomplete phenotyping and a lack of longitudinal sample collection.[10] Many of the risk factors are inherently entangled and occur at different frequencies over the life course. Moreover, the timing and frequency of risk factors can affect how much they impact health outcomes.[11] Differences in risk factor prevalence are expected to drive the environmental heterogeneity that contributes to racial health disparities, but attributing causality to environmental variables is difficult for a number of reasons, not least because many risk factors are correlated with self-identified race. For these reasons, precise, repeated environmental measurements across the entire gestational period are critical to providing insight into causal factors that contribute to perinatal outcomes.

The Pregnancy, Race, Environment, Genes (PREG) study was designed to address the complexity of estimating the factors that contribute to racial health disparities in perinatal outcomes in several ways. First, the PREG study used repeated sampling of biological and environmental measures over the course of pregnancy to test aetiological models of causal relationships between environmental and biological measures. The study design was guided by the presence of environmental heterogeneity between races, environmentally influenced changes in gene expression (GE), contribution of both fetal and maternal genetic factors to gestational age at birth and the appreciation that individual differences in complex traits are best understood through gene-environment interactions. Second, both environmental and biological measures obtained from multiple sources were collected over the course of pregnancy (eg, self-report, objective measures, medical records, blood from infant and mother), which allows thorough phenotyping of environmental and biological factors, and leverage to investigate their relationships over time. Third, the PREG study included two postpartum assessments to test similar mechanistic hypotheses regarding major depression in the peripartum (MDP), defined as an episode of major depressive disorder that onsets either during pregnancy or within 4 weeks postpartum.[12]

Conceptual model overview

Current literature supports a 'complex, multifactorial causal framework' describing racial disparity in birth outcomes.[13] Figure 1 illustrates a theory-driven, developmental model of potential environmental and biological contributions to PTB. Arrows from each aetiologic factor correspond to established or theoretically possible causal pathways. This model is based on empirical evidence demonstrating: (1) epidemiological support for SEDH contributing to poor pregnancy outcomes[14]; (2) the effect of pregnancy-specific environments (PSE) on birth outcomes[15]; (3) changes in DNAm following either differential GE or environmental exposures in both human[16] and animal models[17]; (4) changes in GE following DNAm changes and/or environmental exposures[18]; (5) association of DNAm and GE profiles with PTB[19 20] and birth outcomes; (6) influence of sequence variation on methylation (mQTL) and GE (eQTL) levels[21] and (7) the consistent and pervasive association of race with environmental risk factors and poor pregnancy outcomes.[22] The model allows for tests of mediation by DNAm and GE on the association between environmental exposures and PTB and the moderating effects of DNA sequence on DNAm and GE.

The overall hypothesis to be tested is that social and stressful environments exert their biological effects on physiological and pathological functioning by regulation

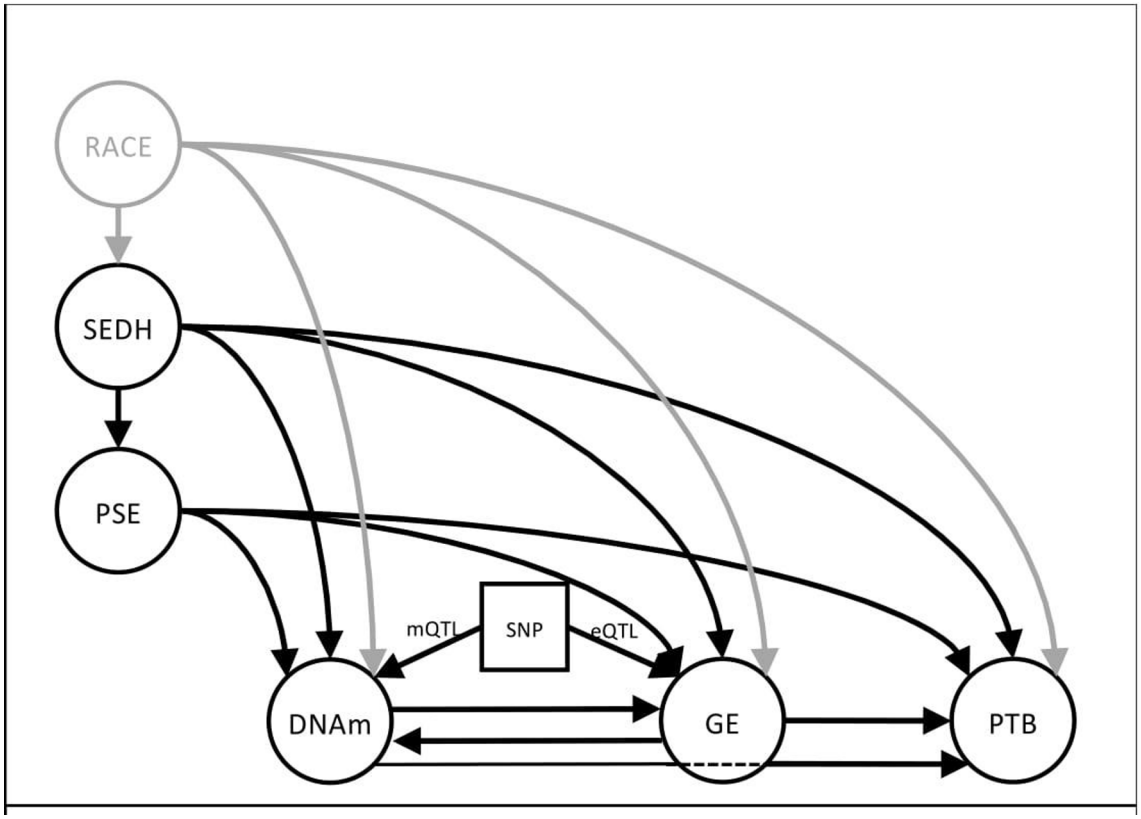

Figure 1 This model illustrates the relationships between environmental factors (left) and how each may affect biological processes (bottom) important to the timing of birth either directly (eg, PSE→DNAm) or indirectly (PSE→DNAm→GE). Chronic stressors are represented by SEDH. Many of those factors are correlated with race and will influence the type of environment a woman experiences during pregnancy. Single headed arrows represent possible causal pathways based on empirical evidence that links SEDH to poor birth outcomes. This framework allows for tests of mediation via DNAm and GE and of moderation effects of DNA sequence (eQTLs/mQTLs). DNAm, DNA methylation; GE, gene expression; PSE, pregnancy-specific environment; PTB, preterm birth; SEDH, social and environmental determinants of health; SNP, single nucleotide polymorphism moderating DNA methylation (mQTL) or GE (eQTL).

of GE in key biological response networks. Environmental risk relevant to pregnancy outcomes can be partitioned into two general groups: (1) those that are established before pregnancy and result in sources of chronic stress (SEDH) and (2) environmental factors that are initiated and can change during pregnancy (PSE). SEDH factors are expected to contribute and correlate with PSE factors. For instance, chronic stress manifested by living in an unsafe neighbourhood before pregnancy is a SEDH, while witnessing a neighbourhood crime during pregnancy would be a PSE factor. In the current model, the SEDH construct of chronic stress is consistent with the 'weathering' hypothesis in which cumulative impact of lifelong social and environmental adversity correlates with deteriorating reproductive health.[14] There can be a 'direct' effect of SEDH on PTB as indicated by the causal arrow in figure 1 (SEDH→PTB), an 'indirect' effect through mediating pathways (SEDH→DNAm→GE→PTB) or both. Indirectly, SEDH could influence PTB risk through the PSE. For instance, socioeconomic status or maternal education are best viewed as SEDH constructs that could directly influence a pregnant mothers' access to prenatal care or a healthy diet. Mechanistic insight into how different SEDH constructs influence PTB can be elucidated through indirect paths to PTB. Lack of social support, ineffective coping strategies and high levels of perceived stress could be associated with epigenetic and GE pathways involved in neuroendocrine deregulation while poor health-related behaviours and neighbourhood environments[23] might elicit gene networks regulating host-pathogen immune response. Longitudinal assessments of PSE and DNAm allow for the assessment of change in these measures and causal relationships between constructs during pregnancy. The conceptual model as presented is specific to racial disparities in gestational age outcomes, but is equally applicable to investigations into other perinatal outcomes like MDP.

COHORT DESCRIPTION

Participant eligibility and recruitment

Eligible women were aged 18–40 years with singleton pregnancies and no diagnosis of diabetes or indication of assisted reproductive technology. Women over the age of 40 years were excluded because they were more likely to have age-related pregnancy complications or be referred

to a high-risk clinic.[24] In addition, both mother and father had to self-identify as either both EA or both AA and be absent of Hispanic or Middle Eastern ancestry. Exclusion criteria at birth included any congenital abnormality, polyhydramnios/oligohydramnios, pre-eclampsia/pregnancy-induced hypertension (PIH)/haemolysis, elevated liver enzymes, low platelet count (HELLP), Rh sensitisation, abruptio placentae, placenta previa, cervical cerclage, medically necessitated preterm delivery and drug abuse. Women participating in fewer than three study time point assessments (including birth) were excluded. Starting in 2014, women meeting the above criteria were eligible to continue into the postpartum extension (cord blood collection was not required).

PREG recruitment began in 2013 and ended by Spring of 2016. The PREG study employed a research nurse to manage and implement recruitment activities at VCU Medical Center Nelson Clinic in downtown Richmond, Virginia and at VCU Medical Center at Stony Point Clinic. The research nurse reviewed appointment records to identify women attending for routine well-baby prenatal visits. Potential research participants were approached in the clinic waiting rooms and, if interested, were provided with a brief summary of the research project opportunity and a participant volunteer brochure. In addition, brochures and flyers describing the study were placed in and around the clinics. All study flyers and brochures included contact information for women interested in participating in PREG.

Study design

Women who were in early pregnancy (<24 weeks gestation) were eligible to enrol in the PREG study if they met all inclusion criteria. Peripheral blood and a detailed inventory of SEDH were assessed via questionnaires to establish baseline. Follow-up questionnaires designed to measure PSE factors were obtained at three follow-up visits along with maternal peripheral blood. Umbilical cord blood was collected at birth (table 1; see online supplementary figure S1 for study flow diagram). Additional information for mother and child was obtained by the study research nurse through medical records abstraction. The goal was to conclude the study with n=200 women meeting all inclusion and no exclusion criteria. Thus, an additional 40 participants (20%) were consented to account for attrition and late pregnancy-related study exclusions. While the final target sample size is modest compared with other epidemiological investigations, the depth of the phenotyping and the longitudinal, repeated nature of the data provides traction for characterising the pathways that mediate effects of the environment on PTB that might be too small to be detected individually.

Items of small monetary value were provided as gifts to participants at each study visit (eg, ultrasound

Table 1 Sample collection schedule for PREG and the MDP extension

	PREG time points (weeks gestation)					MDP time points	
Measure	**0–15**	**10–25**	**20–40**	**37–42**	**Birth**	**6 weeks**	**6–12 months**
Maternal data*							
SEDH survey	232	–	–	–	–	–	–
PSE survey	232	185	192	147	–	105	57†
DNAm	234	190	183	161	–	107	55†
Gene expression	–	–	–	159	–	–	–
GWAS‡	234	–	–	–	–	–	–
Telomere length	234	–	–	–	–	–	–
Micronuclei	234	–	–	–	–	–	–
NIfETy§	139	–	–	–	–	–	–
Child data (newborn)¶							
DNAm	–	–	–	–	136	–	–
Gene expression	–	–	–	–	87	–	–
Telomere length	–	–	–	–	136	–	–

The numbers shown in the table represent the number of samples collected. Only samples from participants who met all birth inclusion criteria will be processed.
*DNAm, gene expression, GWAS, telomere length and micronuclei measurements using peripheral blood.
†Approximately half of the participants who continued to the MDP extension aged out before the second postpartum visit because funding was not obtained until 2014.
‡Intended to identify methylation quantitative trait loci.
§NIfETy ratings were completed either during pregnancy or within 1 year of birth.
¶DNAm, gene expression and telomere length measured in umbilical cord blood.
DNAm, DNA methylation; GWAS, genome-wide association study; MDP, major depression in the peripartum; NIfETy, Neighbourhood Inventory for Environmental Typology; PSE, pregnancy-specific environment; SEDH, social and environmental determinants of health.

picture magnet frames, onesies, swaddling blankets). Women also were compensated financially at each visit for their study participation. In 2014, additional funding was obtained to follow PREG participants for up to 1 year postpartum. The sample size goal for the postpartum extension was n=100 because by the time the additional funding was received, around half of the PREG sample had been recruited already. Research participants completing all required PREG study visits were contacted and offered the opportunity to participate in the MDP extension and enrolled if they met all PREG birth inclusion criteria. The MDP study extension included two additional visits, during which additional blood draws and questionnaires were collected along with medical records abstraction. The first blood draw occurred 6–8 weeks postdelivery and the second within a year postdelivery.

IRB approval, privacy and informed consent

The research nurse conducted the informed consent with all research participants at recruitment. The study design, rationale and aims were described to participants, and ample time was provided to answer questions. Informed consent was obtained separately for eight different aspects of the study, including medical data abstraction, cord blood collection and long-term sample storage for future reproductive research studies. Participants could opt out of any part of the study. In 2014–2016, additional grant funding was obtained to collect two additional MDP study visits after delivery to assess onset and indicators of perinatal depression.

Data collection and handling procedure

Questionnaires

Self-report questionnaires were collected at four pregnancy visits (table 1) to measure SEDH and PSE exposures (see online supplementary table S1). Surveys were completed using the Research Electronic Data Capture (REDCap) software on tablet computers. REDCap is a web-based application that can create, distribute and store data securely from questionnaires. The research nurse and trained student assistants were present to help participants with any questions regarding the software or to clarify questionnaire items. If anyone needed more time or preferred to complete the questionnaire at home, they were emailed a link.

Blood samples

Maternal peripheral blood was drawn during each of the four pregnancy visits by the research nurse. Every attempt was made by the research nurse to coordinate study blood draws with clinic blood draws as part of usual care to minimise extra needle sticks. Cord blood was collected immediately following birth by either the research nurse or trained members of the Labor and Delivery Department, depending on availability. In the event that a cord blood sample was missed at birth, the research nurse retrieved umbilical cord specimens that were refrigerated for later testing. These samples were suitable for DNAm analysis but not GE.

Medical records abstraction

Pregnancy and birth outcome data from patient records was accessed by the research nurse using the CERNER medical records service and entered into a REDCap form. After each visit, the research nurse reviewed prenatal visit notes and verified inclusion criteria were still being met.

Neighbourhood Inventory for Environmental Typology

Neighbourhood Inventory for Environmental Typology (NIfETy) assessments were performed for PREG participants who lived within a 1 hour drive from Richmond, Virginia, and were located in either an urban or suburban area.[25] Participants who lived in rural areas were excluded because the NIfETy tool was designed to rate neighbourhood blocks. All neighbourhood evaluations were performed by pairs of trained raters during the daytime, and raters knew which block to rate but not which house belonged to a participant.

Depth and breadth of environmental and biological data

Environmental assessment

Questionnaires

The initial questionnaire included 1328 questions, took approximately 40–90 min to complete and covered a wide range of topics spanning lifetime exposures to trauma, neighbourhood quality, perceived stress, pregnancy-specific stress, lifestyle, housing and food security and lifetime and current symptoms of depression, anxiety and substance use (see online supplementary table S1). Follow-up pregnancy questionnaires at visits 2–4 were a subset of the baseline and specifically inquired about stressors and experiences that happened since the last PREG questionnaire. They required approximately 30–60 min to complete. For the postpartum visits, questionnaires included 83 questions and took about 30 min to complete. The response rate for key variables in baseline questionnaires is shown in table 2 for PREG participants who met all inclusion criteria and no exclusion criteria. Questionnaire data from the MDP study extension is currently being finalised.

Neighbourhood Inventory for Environmental Typology

The NIfETy instrument provides a detailed assessment of neighbourhood social and environmental factors, including access to public transportation and recreational outlets, indicators of violence and drug use and physical layout (eg, presence of sidewalks, amount of car and foot traffic, etc). In total, the assessment collects >130 variables and provides an objective measure of environmental exposures.

Biological/biomarker measurements

The major focus for the biological data collection was genome-wide DNA methylation (DNAm) measurement from maternal peripheral blood and infant cord blood. DNAm is an epigenetic modification responsible in

Table 2 Demographic characteristics of PREG study participants

	AA	EA	Response rate (AA | EA)	Significance
Participants (%, N)	50.3% (89)	49.7% (88)	–	–
Gestational age at birth*†	272.5 (10.5)	277.6 (8.3)	100% | 100%	<0.0001
Age*	27.0 (5.5)	31.0 (3.4)	100% | 100%	<0.0001
Body mass index*	29.1 (8.9)	25.4 (5.2)	99% | 100%	<0.0001
Current smoker	11.2% (10)	1.1% (1)	94% | 97%	<0.0001
No health insurance before pregnancy	20.2% (18)	2.3% (2)	100% | 100%	<0.0001
First pregnancy	21.3% (19)	43.2% (38)	97% | 97%	0.002
No prenatal vitamin‡	65.2% (58)	10.2% (9)	97% | 97%	<0.0001
Relationship status			100% | 97%	<0.0001
Married/in a relationship	66.3% (59)	94.3% (83)		
Single	29.2% (26)	0.0% (0)		
Other§	1.8% (2)	4.0% (5)		
Did not answer	5.3% (6)	4.8% (6)		
Employment¶			99% | 98%	<0.0001
Full	21.3% (19)	67.0% (59)		
Part	28.1% (25)	18.2% (16)		
Student	21.3% (19)	5.7% (5)		
Unemployed	37.1% (33)	8.0% (7)		
Educational attainment			100% | 98%	<0.0001
Did not complete HS	22.5% (20)	1.1% (1)		
Completed HS/GED	34.8% (31)	3.4% (3)		
Some college	28.1% (25)	12.5% (11)		
Completed college/professional degree	12.4% (11)	80.7% (71)		
Household income			99% | 98%	<0.0001
<US$60 000	73.0% (65)	56.8% (50)		
>US$60 000	2.2% (2)	39.8% (35)		
Do not know/prefer not to answer	23.6% (21)	1.1% (1)		

Categorical tests were performed using two-sided Fisher's exact tests and Student's t-tests were applied to quantitative variables.
*Mean (SD).
†Measured in days.
‡Assessed in SEDH questionnaire given at first visit.
§Separated or never married.
¶Only unemployment status was tested for significance because many full-time and part-time employees were also students.
AA, African-American; EA, European-American; GED, General Equivalency Diploma; HS, Highschool; PREG, Pregnancy, Race, Environment, Genes; SEDH, social and environmental determinants of health.

part for maintaining chromatin structure and modulating gene regulation.[26] Changes in DNAm can result in altered GE, and aberrant DNAm profiles have been associated with constitutional and acquired abnormalities like imprinting disorders[27] and cancer,[28] respectively. Together, the DNAm data from the dyad and maternal SEDH and PSE measurements can provide insight into the degree of similarity between maternal and infant DNAm profiles and how maternal environmental exposures affect infant DNAm and epigenetic age.[29–31]

Genome-wide DNAm measurements were performed on maternal blood at four time points during pregnancy and on cord blood samples using Illumina Infinium HumanMethylation 450 BeadChip (450k).[32] Maternal postnatal DNAm assessments were performed using the EPIC 850 BeadChip (850k),[33] which includes >90% of the 450k probe set. Both microarrays provide standardised measurements of DNAm in intragenic and intergenic regions, covering >99% of RefSeq genes. The DNAm measurements overlapping or near single nucleotide polymorphisms can be leveraged to generate ancestry-relevant principal components, which can be incorporated as covariates in multivariate models.[34] Analytic pipelines developed for 450k data are relevant for analysing data from the 850k,[35] and methods exist for combining datasets with a mixture 450k and 850k data.[36]

In addition to DNAm, GE, telomere length and micronuclei prevalence were measured. GE was assayed in maternal peripheral blood and infant cord blood using the GeneChip Human Genome U133 Plus 2.0 Array (U133 Plus). The U133 Plus microarray assays >47 000 transcripts and variants and includes all of the probe sets from the U133A 2.0, U133A and U133B microarrays.[37 38] These data can be used to determine which DNAm marks are directly associated with GE. Both telomere length and micronuclei prevalence are thought to be biomarkers of cellular ageing and health.[39] Telomeres are repetitive DNA sequences found at the ends of every chromosome. Their lengths have been associated with overall genomic stability,[40] can be a useful marker of long-term stressors, and were measured via quantitative PCR. Micronuclei are small chromatin-containing structures that neighbour parent cells and arise from the exclusion of a whole or partial laggard chromosome following mitotic cell division. This cytological phenomenon is used as an end point to quantify chromosomal instability.[41] The presence of micronuclei has been associated with recent and lifetime exposures to cytotoxic and genotoxic agents.[42 43] Both average telomere length and micronuclei prevalence were assessed only at baseline to estimate the impact of prepregnancy environmental exposures to genomic and cellular health. GE was measured on the visit before pregnancy to assess which DNAm marks measured during previous study visits may have influenced gene transcription levels in maternal peripheral blood later in pregnancy.

FINDINGS TO DATE

Response and retention rates

Of the 240 women enrolled in the PREG study, 177 women (AA=89, EA=88) met full inclusion criteria, and 126 (71.2%) participants completed at least three pregnancy visits in addition to cord blood collection. The PREG study maintained a favourable retention rate (74%), and of the 63 participants who met exclusion criteria, 44 (69.8%) were for a medical reason. The most common cause of exclusion was miscarriage (n=17), followed by presence of pre-eclampsia/PIH/HELLP (n=12). Adequate cord blood samples were collected from n=136 participants (AA=66; EA=70). To date, 139 NIfETy ratings have been completed. Enrolment for the MDP extension concluded with 104 participants consented (AA=50; EA=54).

Demographic structure

Demographic information was collected through self-report questionnaires and medical records abstraction (table 2). Mean age for AA participants was significantly less than for EA participants (27.0 years (SD=5.5) vs 31.0 years (SD=3.4)) and mean BMI was significantly greater in AA participants (29.1 (SD=8.9) vs 25.4 (SD=5.2)). Almost half (42%) of the EA women and 22% of AA women were primiparous. Other significant group differences included use of prenatal vitamins (35% AA used vs 90% EA), relationship status (29% AA single vs 0% EA) and presence of health insurance prior to pregnancy (20% AA did not have health insurance before pregnancy vs 2% EA).

Mean gestational length at birth was significantly less for the AA participants by 5.1 days in PREG study participants (272.5 (SD=10.5) days vs 277.6 (SD=8.3)). Women who continued to the postpartum extension did not differ significantly by race, maternal age, gestational age at birth, income, education, prenatal vitamin use or relationship, unemployment or student status from PREG participants who did not met any pregnancy or birth exclusion criteria. Multiple regression analysis was used to estimate the amount of explained variance of frequently used covariates assessed in socioeconomic models (eg, years of education, relationship status, primiparity, maternal age, maternal BMI and smoking history) on gestational age at birth. These predictors, assessed at baseline, accounted for 9.7% of the variance in gestational age at birth (adjusted R^2=6.3%). This set of variables both explains a modest amount of variance in gestational age at birth and is associated with self-identified race, which makes each variable in the set ideal to test as candidates for mediating the influence of race on gestational age at birth. Further tests can uncover DNAm loci that participate in the identified mediating pathways.

STRENGTHS AND LIMITATIONS

Strengths

Phenotyping quality

A key strength of the PREG study is the scope and quality of repeated PSE measures collected over the course of the peripartum period, which is not typically seen in other studies. This study design allows for modelling relationships between environmental (SEDH and PSE) and biological (eg, DNAm, telomere length) variables, assessing how SEDH and PSE influence DNAm, and testing causal hypotheses for multiple perinatal outcomes of interest, including PTB and MDP. The dataset incorporates objective and self-reported environmental measures and multiple biological measures, including DNAm and GE measures. Very few datasets currently exist outside of cancer research that include genome-wide DNAm and GE measures, and PREG includes paired DNAm and GE measures for both mother and infant. Additionally, daily measurements of environmental contaminants and pollutants, such as carbon monoxide, lead, sulfur dioxide, particulate matter and ozone, have been obtained from the Virginia Department of Environmental Quality archive for 2013–2016.[44] These data come from 11 monitoring stations around Virginia, including one in Richmond, and provide insight into annual and seasonal differences in regional levels of pollution. Other longitudinal studies, such as the Global Alliance to Prevent Prematurity and Stillbirth (GAPPS),[45] also have collected data from self-report questionnaires and multiple tissue sources; however, a major difference between PREG and GAPPS is that the

GAPPS questionnaires focus on medical history, medication use, gynaecological and sexual history and lifestyle choices (eg, diet) rather than constructs important to stress and weathering. Items to measure coping and stress are present, but they are not the emphasis. PREG questionnaires thoroughly interrogated personal history of trauma, coping efforts, self-esteem, social support and job stress occurring before and during pregnancy to capture comprehensive and robust assessments of SEDH and PSE.

The PREG study has the potential to increase understanding of how SEDH and PSE exposures influence risk for adverse perinatal outcomes and affect racial health disparities in perinatal outcomes. Furthermore, the repeated measures of DNAm will provide insight into how DNAm changes over the course of pregnancy and how SEDH and PSE influence DNAm temporal patterns. The PREG study results have the advantage of potentially being generalisable to a large proportion of the US population since the sample includes AA and EA participants and the rates of PTB and depressive symptoms are representative of much of the US population.

Recruitment, retention and response rates

Research studies targeting pregnant women and under-represented populations present unique challenges for recruitment compared with those from the general population. We found that using a sole research nurse to coordinate scheduling and research visits was able to meet ascertainment goals and minimised participant attrition rates. The vast majority of participants who dropped out of the study were excluded for medical reasons. Of the 19 participants who left for non-medical reasons, only 4 participants requested to leave the PREG study. Eight moved out of the area and/or changed providers. Participants did not seem discouraged from enrolling in this study, which included multiple study visits and lengthy questionnaires. One reason for this positive reception was that every effort was made to schedule study visits during prenatal care visits to maximise convenience for participants, including the ability to start the questionnaire in the clinic and the option of finishing at home since the data capture tool use was web-based.

Transparent data processing

All self-report data were collected through REDCap and processed using automated R scripts. Using electronic data collection and automated processing reduces the likelihood of human error during data transfer and provides a highly efficient, transparent and reproducible way to process and analyse data. Moreover, many electronic methods have features that ease collaboration efforts. For example, REDCap can generate data dictionaries that concisely display variable names, which facilitates data requests and sharing.

Limitations

The investment in collecting repeated measurements restricted the total number of pregnant women that could be recruited. While the size of the cohort is modest, it includes approximately 1100 person time points of data. Moreover, longitudinal data from a medium-sized cohort provides insights unavailable even to large cross-sectional studies and provides a framework for testing causal hypotheses. In addition, the presence of at least four time points of data allows for the assessment of non-linear growth models.

Having a single research nurse meant that at every visit, participants saw a familiar face, and in between visits, they had a single point of contact; however, being the sole liaison required extensive planning, especially on days when the research nurse needed to travel to multiple clinics to collect either peripheral blood or umbilical cord blood. Moreover, sometimes it was not possible for the research nurse to be present to collect cord blood samples, which required coordinating with hospital labour and delivery personnel to ensure collection occurred.

Lessons

Maintaining high retention rates

One of the most influential factors that contributed to the retention rate was the research nurse. This person was the primary contact for all PREG study participants regardless of recruitment site. Providing multiple communication options for contacting the research nurse (eg, email, phone, text) combined with establishing a reliable and familiar contact fostered a strong rapport with participants.

Coordinating with clinical recruitment sites

As skilled clinicians, research nurses are ideal team members for studies that either recruit or collect samples in medical environments (eg, hospitals, clinics, etc). Their nursing training allows them to integrate quickly and adapt into varying clinical environments, which facilitates good rapport with clinical staff. Medical doctors and nurses at recruitment sites coordinated with the PREG research nurse so that questionnaires could be distributed and blood draws obtained during normal prenatal care visits without disrupting care.

Including research nurses as part of the study team can also improve participant experience. The clinical training nurses receive can help to identify and handle potentially sensitive situations with care. This skill is especially beneficial for studies recruiting in environments that people could be visiting for very different reasons. For example, whenever possible, the PREG research nurse screened medical charts before approaching potential participants to verify that they were not in clinic for fetal loss or other pregnancy complications.

FUTURE PLANS AND COLLABORATION

Additional biomarker processing is ongoing and includes DNA sequencing, GE analysis, global telomere length measurement and assaying micronuclei frequencies.

Initial NIfETy neighbourhood assessments have been completed, and the secondary ratings to assess reliability will be completed by early 2018. Researchers interested in using PREG data are encouraged to contact the corresponding author for more information regarding data availability.

CONCLUSION

The longitudinal design of the PREG study provides traction to investigate causal relationships between environmental exposures and both DNAm and perinatal outcomes. The scope and quality of data will support investigations in many closely related research areas pertaining to pregnant women, including stress, trauma, racial health disparities and coping efforts.

Author affiliations
[1]Department of Human and Molecular Genetics, Virginia Commonwealth University School of Medicine, Richmond, Virginia, USA
[2]Department of Obstetrics and Gynecology, Virginia Commonwealth University, Richmond, Virginia, USA
[3]Eastern Virginia Medical School, Norfolk, Virginia, USA
[4]Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, USA
[5]Department of Family and Community Health Nursing, Virginia Commonwealth University, Richmond, Virginia, USA
[6]Department of Psychology and Neuroscience, Baylor University, Waco, Texas, USA
[7]Department of Pathology, Virginia Commonwealth University, Richmond, Virginia, USA

Acknowledgements The authors would like to thank the mothers of PREG as well as the Labor and Delivery teams at participating hospitals and clinics who selflessly dedicated their time in the hope of improving birth outcomes for future families.

Contributors TPY conceived the PREG study and in collaboration with AA, PAK, RR-N, SJL and JFS developed the study design, including selecting biological measures, psychiatric instruments and statistical models. PAK developed the MDP extension in collaboration with TPY, AA and CJ-C (and other colleagues: Angela Starkweather, Leroy Thacker). SW was the research nurse responsible for participant recruitment and retention. EO created the REDCap questionnaires. CJ-C performed all DNA extraction and will perform all laboratory work for telomere length and micronuclei assays. DL performed data processing and wrote the first draft of the manuscript with input from TPY. All authors contributed feedback regarding manuscript edits and approved of the final manuscript.

Funding Funding for this study comes from the National Institute on Minority Health and Health Disparities (P60MD002256, PI: York, JFS), The John and Polly Sparks Foundation and Brain and Behavior Research Foundation (24712, PI: TPY), American Nurses Foundation Research Grant (5232, PI: PAK), VCU Center for Clinical and Translational Research Endowment Fund (6-40595, PI: PAK). DL is supported by National Institute of Mental Health T32MH020030. The use of REDCap was supported by Clinical and Translational Science Award (CTSA) award No. UL1TR000058 from the National Center for Advancing Translational Sciences.

Disclaimer The contents of this study are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.

Competing interests None declared.

Patient consent Obtained.

Ethics approval The PREG study and MDP extension were approved by the Virginia Commonwealth University Institutional Review Board (IRB protocol #14000).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data processing for the PREG study is ongoing. Anyone interested in collaboration or data access should contact TPY (corresponding author) for more information.

REFERENCES

1. Beck S, Wojdyla D, Say L, *et al*. The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity. *Bull World Health Organ* 2010;88:31–8.
2. Goldenberg RL, Cliver SP, Mulvihill FX, *et al*. Medical, psychosocial, and behavioral risk factors do not explain the increased risk for low birth weight among black women. *Am J Obstet Gynecol* 1996;175:1317–24.
3. York TP, Strauss JF, Neale MC, *et al*. Racial differences in genetic and environmental risk to preterm birth. *PLoS One* 2010;5:e12391.
4. Komlos J, Lauderdale BE. The mysterious trend in American heights in the 20th century. *Ann Hum Biol* 2007;34:206–15.
5. Gravlee CC. How race becomes biology: embodiment of social inequality. *Am J Phys Anthropol* 2009;139:47–57.
6. Martin J, Hamilton B, Sutton P. Births: Final data for 2003. National vital statistics reports. *Natl Cent Heal Stat* 2005;54.
7. Dominguez TP. Race, racism, and racial disparities in adverse birth outcomes. *Clin Obstet Gynecol* 2008;51:360–70.
8. Mustillo S, Krieger N, Gunderson EP, *et al*. Self-reported experiences of racial discrimination and Black-White differences in preterm and low-birthweight deliveries: the CARDIA Study. *Am J Public Health* 2004;94:2125–31.
9. Anum EA, Retchin SM, Garland SL, *et al*. Medicaid and preterm births in Virginia: an analysis of recent outcomes. *J Womens Health* 2010;19:1969–75.
10. Manuck TA. The genomics of prematurity in an era of more precise clinical phenotyping: a review. *Semin Fetal Neonatal Med* 2016;21:89–93.
11. Class QA, Lichtenstein P, Långström N, *et al*. Timing of prenatal maternal exposure to severe life events and adverse pregnancy outcomes: a population study of 2.6 million pregnancies. *Psychosom Med* 2011;73:234–41.
12. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. *Am Psychiatr Assoc* 2013;991.
13. Kramer MR, Hogue CR. What causes racial disparities in very preterm birth? A biosocial perspective. *Epidemiol Rev* 2009;31:84–98.
14. Messer LC, Kaufman JS. Invited commentary: the socioeconomic causes of adverse birth outcomes. *Am J Epidemiol* 2010;172:135–7.
15. Berg CJ, Wilcox LS, d'Almada PJ. The prevalence of socioeconomic and behavioral characteristics and their impact on very low birth weight in black and white infants in Georgia. *Matern Child Health J* 2001;5:75–84.
16. McGowan PO, Sasaki A, D'Alessio AC, *et al*. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nat Neurosci* 2009;12:342–8.
17. Weaver IC, Cervoni N, Champagne FA, *et al*. Epigenetic programming by maternal behavior. *Nat Neurosci* 2004;7:847–54.
18. Hussain N, Krueger W, Covault J, *et al*. Effects of prenatal tobacco exposure on gene expression profiling in umbilical cord tissue. *Pediatr Res* 2008;64:147–53.
19. Wang H, Ogawa M, Wood JR, *et al*. Genetic and epigenetic mechanisms combine to control MMP1 expression and its association with preterm premature rupture of membranes. *Hum Mol Genet* 2008;17:1087–96.
20. Enquobahrie DA, Williams MA, Qiu C, *et al*. Early pregnancy peripheral blood gene expression and risk of preterm delivery: a nested case control study. *BMC Pregnancy Childbirth* 2009;9:56.
21. Shakhbazov K, Powell JE, Hemani G, *et al*. Shared genetic control of expression and methylation in peripheral blood. *BMC Genomics* 2016;17:278.
22. Collins JW, David RJ, Handler A, *et al*. Very low birthweight in African American infants: the role of maternal exposure to interpersonal racial discrimination. *Am J Public Health* 2004;94:2132–8.
23. Paul K, Boutain D, Manhart L, *et al*. Racial disparity in bacterial vaginosis: the role of socioeconomic status, psychosocial stress, and

neighborhood characteristics, and possible implications for preterm birth. *Soc Sci Med* 2008;67:824–33.
24. Marozio L, Picardo E, Filippini C, *et al*. Maternal age over 40 years and pregnancy outcome: a hospital-based survey. *J Matern Fetal Neonatal Med* 2017:7:1–7.
25. Furr-Holden CD, Smart MJ, Pokorni JL, *et al*. The NIfETy method for environmental assessment of neighborhood-level indicators of violence, alcohol, and other drug exposure. *Prev Sci* 2008;9:245–55.
26. van der Maarel SM. Epigenetic mechanisms in health and disease. *Ann Rheum Dis* 2008;67(Suppl 3):iii97–iii100.
27. Li E, Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol* 2014;6:a019133.
28. Irizarry RA, Ladd-Acosta C, Wen B, *et al*. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 2009;41:178–86.
29. Knight AK, Craig JM, Theda C, *et al*. An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biol* 2016;17:206.
30. Simpkin AJ, Hemani G, Suderman M, *et al*. Prenatal and early life influences on epigenetic age in children: a study of mother-offspring pairs from two cohort studies. *Hum Mol Genet* 2016;25:191–201.
31. Javed R, Chen W, Lin F, *et al*. Infant's DNA Methylation Age at Birth and Epigenetic Aging Accelerators. *Biomed Res Int* 2016;2016:1–10.
32. Bibikova M, Barnes B, Tsan C, *et al*. High density DNA methylation array with single CpG site resolution. *Genomics* 2011;98:288–95.
33. Pidsley R, Zotenko E, Peters TJ, *et al*. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* 2016;17:208.
34. Barfield RT, Almli LM, Kilaru V, *et al*. Accounting for population stratification in DNA methylation studies. *Genet Epidemiol* 2014;38:231–41.
35. Wright ML, Dozmorov MG, Wolen AR, *et al*. Establishing an analytic pipeline for genome-wide DNA methylation. *Clin Epigenetics* 2016;8:45.
36. Aryee MJ, Jaffe AE, Corrada-Bravo H, *et al*. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014;30:1363–9.
37. Shi L, Reid LH, Jones WD, *et al*. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006;24:1151–61.
38. Affymetrix. GeneChip® human genome arrays data sheet. 2003 https://www.affymetrix.com/support/technical/datasheets/human_datasheet.pdf.
39. Zhang L, Hu XZ, Li X, *et al*. Telomere length - a cellular aging marker for depression and Post-traumatic Stress Disorder. *Med Hypotheses* 2014;83:182–5.
40. O'Sullivan RJ, Karlseder J. Telomeres: protecting chromosomes against genome instability. *Nat Rev Mol Cell Biol* 2010;11:171–81.
41. Fenech M, Chang WP, Kirsch-Volders M, *et al*. HUMN project: detailed description of the scoring criteria for the cytokinesis-block micronucleus assay using isolated human lymphocyte cultures. *Mutat Res* 2003;534:65–75.
42. Sawal HA, Asghar K, Bureik M, *et al*. Bystander signaling via oxidative metabolism. *Onco Targets Ther* 2017;10:3925–40.
43. Sram RJ, Svecova V, Rossnerova A. Systematic review of the use of the lymphocyte cytokinesis-block micronucleus assay to measure DNA damage induced by exposure to polycyclic aromatic hydrocarbons. *Mutat Res* 2016;770(Pt A):162–9.
44. Virginia Department of Environmental Quality. Air quality in Virginia. http://vadeq.tx.sutron.com/ (accessed 1 Dec 2017).
45. Lawn JE, Gravett MG, Nunes TM, *et al*. Global report on preterm birth and stillbirth (1 of 7): definitions, description of the burden and opportunities to improve data. *BMC Pregnancy Childbirth* 2010;10(Suppl 1):S1.