Polymethylation scores paper

F blostein

9/13/2021

# Title Page

# Abstract

**Background**:

**Results**

**Conclusions**

# Keywords

# Background

Maternal prenatal smoking is a public health concern: in 2018, 11% of pregnant US women aged 15-44 reported any past-month cigarette use [1]. Exposed children are more likely to have low-birth weight, negative neurodevelopmental outcomes and asthma [2]. Measuring prenatal maternal smoking behavior is challenging. Maternal prenatal smoking is underreported due to stigma [3, 4]. The gold-standard for smoking measurements, serum cotinine, has a short half-life of nine hours in pregnant women [5]. Further, serum cotinine measures during pregnancy are rarely available outside of birth cohorts, limiting clinical application as a marker of prenatal exposures. Developing a portable and reliable biomarker of prenatal maternal smoke exposure would have important implications for research and clinical practice.

Differences in DNA methylation signatures of cordblood and placental tissues by prenatal maternal smoke exposure exist in primarily European cohorts of newborns [6–8]. Blood DNA methylation signatures in cohorts of older children and adults also differ by prenatal maternal smoking exposure [9, 10]. Yet few studies have examined associations at different ages in the same participants. Additionally, while associations are well-replicated in cohorts of European ancestry, only a few, small sample size studies are in non-European ancestry populations. One 2016 study included 186 participants of African or admixed ancestry [11], while a 2018 study of 89 participants had 60% of African or admized ancestry [9]. A larger 2019 study of 572 Latino children replicated associations with prenatal maternal smoking at 7 out of 148 tested *a priori* CpG sites from Joubert et al. at a FDR-adjusted *P*-value<0.05 [6, 12]. Thus, while DNA methylation is associated with prenatal maternal smoking, the consistency of this association across age and ancestry is understudied.

The association between prenatal maternal smoking and DNA methylation suggests DNA methylation could provide a reliable and specific biomarker of prenatal maternal smoke exposure. Importantly, changes in DNA methylation are specific to smoke exposure *in utero*. Previous work has found reasonable performance of cord-blood accuracy ranging from 81% to 91%) [11, 13]. A 2018 study found that DNA methylation based scores from adult peripheral blood could even predict predict prenatal smoke exposure (30 years previously) with an area under the curve of 0.72 (95% confidence interval 0.69, 0.76) [10]. However, the consistency of DNA methylation as a biomarker for prenatal maternal smoking across age and ancestry has not been evaluated. Few studies have compared the effectiveness of DNA methylation as biomarker of prenatal maternal smoking in the same individuals at different times or evaluated if the inclusion of multiple time points improves performance. Additionally, previous work has relied on blood and placental samples. DNA methylation varies across cell-types and tissues, and epigenetic markers of prenatal maternal smoking are known to differ between cord blood and placental samples[7]. Cord blood or placenta samples are unlikely to be available in a clinical setting. While peripheral blood is a more realistic clinical sample, saliva is even easier to collect. To our knowledge, no study has evaluated salivary DNA methylation as a biomarker for prenatal maternal smoking.

DNA methylation changes hold promise as a biomarker for prenatal maternal smoking. However, before translation to clinical applications, we must understand how and if the signal varies across age, ancestry and tissue sample. The goal of this study was to assess the *in vitro* potential of salivary DNA methylation biomarkers for prenatal maternal smoking. We tested associations between prenatal maternal smoke exposure and saliva DNA methylation biomarkers from the Illumina 450K chip, including epigenetic clocks, polymethylation scores, and individual *a priori* CpG sites using longitudinal samples from the Fragile Families and Child Wellbeing study. We performed age- and ancestry-stratified analyses to test the hypothesis that DNA methylation could serve as a consistent biomarker of prenatal maternal smoking across populations and time.

# Results

## Study sample descriptive statistics

Of 1812 samples from 897 unique individuals with DNA methylation data measured on the 450K Illumina array, 1505 samples from 815 individuals had data on additional key covariates (Figure 1). 690 individuals had both age 9 and age 15 samples. Excluded samples were similar to included samples, except that included samples were slightly more likely to be from children of African ancestry and less likely to be from children of Hispanic ancestry (Supplemental Table 1). Twenty percent of the mothers reported any prenatal maternal smoking, 12% reported prenatal alchol use and 5% reported prenatal drug use. The mean income to poverty ratio of mothers at birth was 2.19. Of the children in the analytic sample, 50% were male, 60% were of African ancestry, and 24% were of Hispanic ancestry.

We calculated several common summary measures of DNA methylation, including global methylation, epigenetic clocks (Pediatric and GRIM), and polymethylation scores. We calculated polymethylation scores using regression coefficients for 6074 CpGs made from three different models in Joubert et al.: a regression using newborn cordblood samples controlling for cell-type proportions, a regression using newborn cordblood samples not controlling for cell-type proportions and a regression using peripheral blood samples from older children. We also examined the top probe from Joubert et al., AHRR:cg05575921. Many of the methylation summary measures were correlated with each other (Supplemental Figure 1). Notably, polymethylation scores calculated using coefficients from a cell-type corrected regression as weights did not correlate with estimated cell-type proportions, while scores calculated using coefficients from a regression not corrected for cell-type did. Among the 690 individuals with data from both visits, methylation summary measures from the same individual were correlated between the age 9 and age 15 (Supplemental Figure 2). The correlation was strongest for the polymethylation score calculated using weights from a regression for sustained smoking in newborn cordblood with cell-type control (Pearson =0.9, *P* value<0.0001). The distribution of epigenetic ages shifted between the age 9 and age 15 visit, as expected (Supplemental Figure 3; mean pediatric clock age 9: 9 age 15: 12, mean GRIM clock age 9 25, age 15: 30. The distribution of estimated cell-type proportions also shifted between visits while the distribution of global methylation and polymethylation scores were more consistent (Supplemental Figure 3). Based on these results, we focused on the polymethylation score from newborn cordblood with cell-type control, the pediatric clock, AHRR:cg05575921 and global methylation for bivariate and multivariable analyses.

## Bivariate associations between prenatal maternal smoking and DNA methylation summary measures

Mothers who reported smoking during pregnancy had lower income to poverty ratios and were more likely to report prenatal alcohol use, prenatal drug use and postnatal smoking than mothers who did not report prenatal maternal smoking (Table 1). Children of mothers who reported smoking during pregnancy were more likely to be of European ancestry and had higher prenatal maternal smoking polymethylation scores at both the age 9 and age 15 visit (Table 1). Children of mothers who reported prenatal maternal smoking did not differ from those whose mothers did not in terms of global methylation, epigenetic age estimaated from pediatric clock or estimate cell-type proportions. Compared to children of mothers not reporting prenatal materna smoking, exposed children were significantly more hypomethylated at AHRR:cg05575921 at age 9 (mean methylation `r 0.77 vs 0.78, *P* value=0.04) but not age 15 ( 0.76 vs 0.77 *P* value= 0.21).

## Multivariable models

The association between prenatal maternal smoking and the cell-type controlled polymethylation score was also observed in multivariable models controlling for precision and confounder variables (Figure 3; Supplemental Table 2). In the entire sample, prenatal maternal smoke exposure was associated with a 0.12 (95%CI: 0.08, 0.15) higher polymethylation score at age 9 and 0.11 (95%CI: 0.07, 0.14) higher polymethylation score at age 15. A consistent association was observed when stratifying to only African ancestry children (0.12 (95%CI: 0.08, 0.17)). The direction and magnitude of the association was similar in European and Hispanic ancestry children at age 9, though attenuated at age 15 (Figure 3; Supplemental Tables 2). Prenatal maternal smoking was associated with hypomethylation at AHRR:cg055975921 at age 9 but not age 15 and not associated with global methylation or epigenetic age from the pediatric clock at either age. Results were similar in linear mixed effect models and in sensitivity models controlling for surrogate variables instead of known covariates (Supplemental Tables 3-4; Supplemental Figure 4-5).

## Accuracy of polymethylation scores as a biomarker of prenatal maternal smoking

Next, we compared the accuracy of different DNA methylation summary measures for classifying prenatal maternal smoking using receiver operating curves (Figure 4A). The polymethylation score with cell-type control had the largest area under the curve (AUC) both when used as a single predictor (age 9 AUC; 0.66, age 15 AUC: 0.65 ) and when used in models including child sex, maternal income-poverty ratio, immune cell proportion and batch (age 9 AUC: 0.77, age 15 AUC: 0.77). The AUC for the polymethylation score was significantly larger than using only the top CpG from Joubert et al., AHRR:cg05575921 (age 9 AUC: 0.55 *P* value=0.005 , age 15 AUC: 0.54 *P* value=0.003). Accuracy was not improved by including polymethylation scores from age 9 and age 15 together (Figure 4B).

# Discussion

In the longitudinal Fragile Families and Child Wellbeing birth cohort, we found an association between prenatal maternal smoking and DNA methylation in sequential saliva samples from ages 9 and 15. Polymethylation scores were associated with prenatal maternal smoking across ages and ancestry while global methylation and epigenetic clocks were not, demonstrating a specific and persistent signal of prenatal smoke exposure in DNA methylation. Polymethylation scores had reasonable accuracy for classifying prenatal maternal smoking. This accuracy was more consistent across time than when using some individual *a priori* CpG sites.

Our findings fit within the previous literature on associations between prenatal maternal smoking and DNA methylation. We replicate top hits from previous epigenome wide association studies, including a large metaanalysis of DNA methylation in blood samples [6]. Although there are no previous studies of prenatal maternal smoking and saliva methylation, our findings fit with findings on the consistency of DNA methylation signatures of own-smoking between saliva and blood samples [14, 15].

We contribute to the literature on DNA methylation and prenatal maternal smoking by evaluating the consistency of this association in saliva and across ages and ancestries. Polymethylation scores built using coefficients from metaanalysis of blood DNA methylation still associated with prenatal maternal smoking in our independent population of saliva samples. Similarly, even though the source population for construction of the polymethylation scores was primarily of European ancestry, an association with prenatal maternal smoking was apparent in our primarily African ancestry population. Effect estimates were consistent across ancestries at age 9, although Hispanic and European ancestry age 15 samples had a lower effect estimate. This could reflect higher unreported smoking initiation in European and Hispanic ancestry US teens than African ancestry teens. Genetic ancestry correlates with race, and White and Latino teens have much higher rates of teen smoking and earlier ages at initiation than Black children [16]. Own-smoking could influence DNA methylation at ages 9 and 15, creating outcome misclassification. This could also explain the observed attenuation of hypomethylation at AHRR:CpG05575921 by age 15 in our sample, as methylation at this probe is known to vary by own-smoking exposure as well[14, 15].

Our findings also fit with previous work on DNA methylation as a biomarker for prenatal maternal smoke exposure. Accuracy of the polymethylation scores in our sample (Age 9: 0.67 and Age 15: 0.66) was worse than the performance of biomarkers using cord blood (AUCs ranging from 0.82 to 0.97) [11, 13]. Differences in methylation patterns across tissues and over time could be responsible for this diminished accuracy. However, our accuracy was comparable to accuracy when using peripheral blood from older adults (AUC 0.72) [10]. The diminished accuracy from cord blood may be an acceptable trade-off considering saliva is a more readily available tissue for epidemiological and clinical applications.

Our analysis answers pertinent questions for the development of a clinically applicable methylation biomarker of prenatal maternal smoking. We show that DNA methylation scores calculated from children’s saliva can accurately classify prenatal maternal smoke exposure in non-European ancestry children. Other bomarkers and risk scores are known to differ across ancestry groups, potentially contributing to and exacerbating health disparities. As new biomarkers are developed, understanding their consistency across ancestry groups is critical. We also demonstrated that accuracy is not substantially affected when using coefficients from newborn cordblood (no cell type control, age 9: 0.62, age 15: 0.61) vs older children peripheral blood (no cell type control age 9: 0.63, age 15: 0.60). However, using coefficients from models that incorporated cell-type control at the time of calculation did result in better accuracy, suggesting cell-type control should be instituted both when creating weights and when applying scores to new population. Including samples from both ages did not improve accuracy, thus repeated samplings may be unnecessary for clinical applications. Additionally, accuracy was more consistent when using polymethylation scores than using the single top-hit from Joubert et al, AHRR:CpG05575921 [6]. Salivary AHRR:CpG05575921 methylation was less accurate in our sample at predicting prenatal maternal smoke exposure than when it has been used to classify own smoking in blood or saliva [15]. While using a single CpG site as a biomarker may be clinically pragamatic, single sites may be less robust to time-since-exposure and new environmental exposures. Alternatively, incorporating information from across several CpGs in a polymethylation score may be more robust to these .

Our analysis is not without its limitations. We used maternal self report of prenatal maternal smoker as serum cotinine levels were not available. Due to social desirability bias, this could result in exposure misclassification. We would expect this to bias our results towards the null. In any analysis of a prenatal exposure and postnatal outcome, there is the possibility of selection bias into the cohort due to live birth bias. Selection bias is also possible due to loss-to-follow-up between birth and age 9. Additionally, while we controlled for postnatal secondhand smoke exposure and excluded children who reported any own-smoking, we cannot exclude the possibility of residual confounding in this observational cohort. However, the longitudinal design with repeated measures, exposure-outcome temporality and large sample size are strengths of our analysis.

# Conclusions

Our analysis contributes to the clinical implementation of a DNA methylation-based biomarker of prenatal maternal smoking. We showed that DNA methylation in saliva, a clinically practical tissue, had strong associations with and reasonable accuracy for prenatal maternal smoke exposure. Further, we demonstrated that polymethylation scores could be applied as a biomarker of prenatal maternal smoke exposure across ancestry-groups, an important consideration for health equity. The development and application of biomarker for prenatal maternal smoke exposure has important implications for epidemiological research and clinical practice. Given the difficulty of measuring prenatal maternal smoke exposure, such a biomarker could allow for confounder control in research areas where such control is currently impossible. Given the prevalence and negative health consequences of prenatal maternal smoke exposure, such a biomarker could be used to identify exposed individuals who might especially benefit from certain interventions.

## Methods

### Cohort

The Fragile Families and Child Wellbeing Study (FFCW) is a birth cohort of nearly 5,000 children born in 20 U.S. cities between 1998 and 2000 [17]. FFCW sampled births using a three-stage stratified random sample design, described in detail elsewhere, which oversampled unmarried mothers by a ratio of 3:1 [17]. FFCW excluded the following births from the study: those with parents who planned to place the child for adoption, those where the father was deceased, those who did not speak English or Spanish well enough to be interviewed, those where the mothers or babies were too ill to complete the interview and those where the baby died before the interview could take place. Children were followed longitudinally with assessments at ages 1, 3, 5, 9 and 15; additional follow up is ongoing. Assessments include medical record extraction, biosample collection, in-home assessments, and surveys of the mother, father, primary caregiver, teacher and child. At ages nine and fifteen a saliva sample was taken from the child for DNA and DNA methylation measurements [17].

### Covariates and exposure measurement

Demographic and prenatal maternal substance use variables were derived from maternal self-report questionnaire data at baseline (child’s birth). Demographic variables included maternal income to poverty ratio and child sex. Prenatal maternal alcohol and other drug use were dichotomized to any vs no prenatal maternal alcohol and drug use respectively. Prenatal maternal smoking was ascertained from the categorical interview question

"During your pregnancy, how many cigarettes did you smoke? Did you smoke…:

2 or more packs a day

1 or more but less than 2 packs per day

Less than 1 pack a day

None

Because few participants reported smoking a pack or more a day, this question was also dichotomized to any vs no prenatal maternal smoking.

Postnatal second-hand smoke exposure was encoded in two variables. To encapsulate general early childhood smoke exposure, we used a binary any postnatal exposure to maternal or primary caregiver smoking at ages 1 or 5 , constructed from questions about maternal or primary caregiver smoking in the age 1 and 5 interviews. To encapsulate recent postnatal smoke exposure, we used a categorical packs per day maternal or primary care giver smoking in the month prior to age 9 and 15 maternal interview.

Principal components of child genetic ancestry were calculated from genetic data (CHIP?) from child saliva samples. In non-stratified models, the first two components of genetic ancestry from principal component analysis run in the entire analytic subset were used, in ancestry-stratified models the first two principal components from principal component analysis run within each strata were used.

Confounder variables included principal components of child genetic ancestry, prenatal maternal alcohol and other drug use.

Precision variables (variables strongly associated with DNA methylation controlled for to decrease standard errors of coefficient of interest) included child sex, cell proportions estimated from DNA methylation data

### DNA methylation measurement

Salivary samples from the children in FFCW were taken at ages nine and fifteen using the Oragene•DNA sample collection kit (DNA Genotek Inc., Ontario). Saliva DNA was extracted manually using DNA Genotek’s purification protocol using prepIT L2P. DNA was bisulfite treated and cleaned using the Zymo Research EZ DNA Methylation kit. Samples were randomized and plated across slides by demographic characteristics. Saliva DNA methylation was measured using the Illumina HumanMethylation 450k BeadArray [18] and imaged using the Illumina iScan system. All samples were run in a single batch to minimize technical variability.

DNA methylation image data (IDAT) were processed in R statistical software (3.5) using the minfi package [19]. The IDAT pairs (n=1811) were read into R and the minfi *preprocessNoob* function was used to normalize dye bias and apply background correction. Further quality control was applied using the ewastools packages [20]. We dropped samples with >10% of sites have detection p-value >0.01 (n=43), sex discordance between DNA methylation predicted sex and recorded sex (n=20), and abnormal sex chromosome intensity (n=3). CpG sites were removed if they had detection p-value >0.01 in 5% of samples (n=26,830) or were identified as cross-reactive (n=27,782) [21]. Relative proportions of immune and epithelial cell types were estimated from DNA methylation measures using a childhood saliva reference panel [22].

Global methylation was calculated for each sample as the mean methylation value of each sample across the cleaned probe set. Pediatric epigenetic age was calculated for each sample. To create polymethylation scores for prenatal maternal smoke exposure, 6,074 CpG sites with regression coefficients reported in Joubert et al. 2016 [6] were multiplied by mean-centered methylation beta values from FFCW and summed. We calculated polymethylation scores using regression coefficients from 4 different regressions reported in Joubert et al. 2016 [6] (sustained smoking exposure and DNA methylation in newborn cordblood with cell-type control, sustained smoking exposure and DNA methylation in newborn cordblood without cell-type control, sustained smoking exposure and DNA methylation in older children peripheral blood, any smoking exposure and DNA methylation in newborn cordblood ).

Epigenome-wide association analyses were run separately at age 9 and age 15 using the linear models and the R package limma [23]. We also ran epigenome wide association analyses using DNA methylation measures from both time points using linear mixed effect models with a random intercept for child ID. Gap probes were identified using the function *gaphunter* from the R package minfi [19].

# Declarations

## Ethics approval and consent to participate

## Consent for publication

Not applicable

# References

1. (2019) National survey on drug use and health 2018.

2. Banderali G, Martelli A, Landi M, Moretti F, Betti F, Radaelli G, Lassandro C, Verduci E (2015) Short and long term health effects of parental tobacco smoking during pregnancy and lactation: A descriptive review. J Transl Med 13:327

3. Kvalvik LG, Nilsen RM, Skjærven R, Vollset SE, Midttun O, Ueland PM, Haug K (2012) Self-reported smoking status and plasma cotinine concentrations among pregnant women in the norwegian mother and child cohort study. Pediatr Res 72:101–7

4. Dietz PM, Homa D, England LJ, Burley K, Tong VT, Dube SR, Bernert JT (2011) Estimates of nondisclosure of cigarette smoking among pregnant and nonpregnant women of reproductive age in the united states. Am J Epidemiol 173:355–9

5. Dempsey D, Jacob P 3rd, Benowitz NL (2002) Accelerated metabolism of nicotine and cotinine in pregnant smokers. J Pharmacol Exp Ther 301:594–8

6. Joubert BR, Felix JF, Yousefi P, et al (2016) DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis. Am J Hum Genet 98:680–696

7. Everson TM, Vives-Usano M, Seyve E, et al (2021) Placental DNA methylation signatures of maternal smoking during pregnancy and potential impacts on fetal growth. Nat Commun 12:5095

8. Cardenas A, Lutz SM, Everson TM, Perron P, Bouchard L, Hivert M-F (2019) Mediation by Placental DNA Methylation of the Association of Prenatal Maternal Smoking and Birth Weight. American Journal of Epidemiology 188:1878–1886

9. Tehranifar P, Wu H-C, McDonald JA, Jasmine F, Santella RM, Gurvich I, Flom JD, Terry MB (2018) Maternal cigarette smoking during pregnancy and offspring DNA methylation in midlife. Epigenetics 13:129–134

10. Richmond RC, Suderman M, Langdon R, Relton CL, Davey Smith G (2018) DNA methylation as a marker for prenatal smoke exposure in adults. Int J Epidemiol 47:1120–1130

11. Ladd-Acosta C, Shu C, Lee BK, et al (2016) Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood. Environ Res 144:139–148

12. Neophytou AM, Oh SS, Hu D, Huntsman S, Eng C, Rodrı́guez-Santana JoséR, Kumar R, Balmes JR, Eisen EA, Burchard EG (2019) In utero tobacco smoke exposure, DNA methylation, and asthma in latino children. Environmental Epidemiology 3:

13. Reese SE, Zhao S, Wu MC, et al (2017) DNA methylation score as a biomarker in newborns for sustained maternal smoking during pregnancy. Environ Health Perspect 125:760–766

14. Dawes K, Andersen A, Vercande K, Papworth E, Philibert W, Beach SRH, Gibbons FX, Gerrard M, Philibert R (2019) Saliva DNA Methylation Detects Nascent Smoking in Adolescents. Journal of Child and Adolescent Psychopharmacology 29:535–544

15. Philibert R, Dogan M, Beach SRH, Mills JA, Long JD (2019) AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 183:51–60

16. El-Toukhy S, Sabado M, Choi K (2016) Trends in Susceptibility to Smoking by Race and Ethnicity. Pediatrics 138:e20161254

17. Reichman NE, Teitler JO, Garfinkel I, McLanahan SS (2001) Fragile families: Sample and design. Children and Youth Services Review 23:303–326

18. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics 6:692–702

19. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30:1363–1369

20. Heiss JA, Just AC (2018) Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. Clinical Epigenetics. <https://doi.org/10.1186/s13148-018-0504-1>

21. Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 8:203–209

22. Middleton LYM, Dou J, Fisher J, et al (2021) Saliva cell type DNA methylation reference panel for epidemiological studies in children. Epigenetics 1–17

23. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43:e47–e47