Considerations for Cell Type Heterogeneity in Pediatric Salivary DNA Methylation Analyses: Comparison of Reference Panels & Stratification by Estimated Cell Type Proportion

Meingold Hiu-ming Chan [1,2,3], Mandy Meijer [1,2,3], Sarah M. Merrill[4], Maggie Po Yuan Fu[1,2,3], David Lin[2], Julia L. MacIsaac[2], Jenna L. Riis[5,6], Douglas A. Granger[6,7], Elizabeth A. Thomas[5, 8], & Michael S. Kobor[1,2,3,9]

[1]Department of Medical Genetics, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[2]BC Children's Hospital Research Institute, University of British Columbia, Vancouver, BC, Canada

[3]Centre for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver, BC, Canada

[4]Department of Psychology, University of Massachusetts Lowell, Lowell, MA USA

[5]Institute for Interdisciplinary Salivary Bioscience Research, University of California, Irvine, CA, USA

[6]Department of Health and Kinesiology, University of Illinois Urbana Champaign, Urbana, IL, USA

[7]Johns Hopkins University, School of Medicine, Baltimore, MD, USA

[8]Department of Neurobiology and Behavior, University of California, Irvine, CA, USA

[9]Edwin S. H. Leong Centre for Healthy Aging, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

Conflict of interest: In the interest of full disclosure, DAG is Chief Scientific and Strategy Advisor at Salimetrics LLC and Salivabio LLC and these relationships are managed by the policies of the committees on conflict of interest at the Johns Hopkins University School of Medicine and the University of California at Irvine.

Funding details: The research reported in this publication was supported by the Environmental influences on Child Health Outcomes (ECHO) program, Office of The Director, National Institutes of Health Award Number 7UH3OD023332-04, The Eunice Kennedy Shriver National Institute of Child Health and Human Development Award Number R01HD081252 and P01HD039667. MC was supported by the Canadian Institutes of Health Research (CIHR) Postdoctoral Fellowship (reference no. MFE-194003). MM was supported by a personal grant from the Dutch Research Council (NWO/ZonMW): Rubicon (grant no. 04520232320009).

Acknowledgements:

We thank Kaitlin Smith, Hillary Piccerillo, Tatum Stauffer, and Andrew Huang for technical assistance with salivary biospecimen testing. We would like to express our gratitude to all of the families, participants, and teachers who participated in this research and to the Family Life Project (FLP) research assistants for their hard work and dedication to the FLP. This study is part of the Family Life Project. We thank Clancy Blair for his leadership of the Family Life Project (FLP), mentorship, and contributions to the early stages of this project. We thank Christopher Bartlett and his team for their technical support on the genotyping of the salivary samples.

Abstract

Saliva is a widely used sample in epigenetic research with children due to its non-invasive nature. Since DNA methylation (DNAm) profile is cell type (CT) specific, salivary DNAm associations with exposures may be influenced by CT compositions, which is highly variable in saliva as it contains immune and buccal epithelial cells (BEC). Reference-based CT deconvolution and statistically adjusting estimated CT in DNAm analyses have become an increasingly common practice. However, careful examinations of how different reference panels may affect DNAm results and alternative approaches (e.g., stratification) are lacking. To scrutinize the best analytical strategies on pediatric salivary DNAm, the current study used 529 salivary DNAm samples obtained from children (mean age = 7.26 years, SD = 0.26 years) in the Family Life Project. Our results showed higher estimated CT variability with child than adult reference panels and highlighted the impact of these estimated CT discrepancies on DNAm associations with social variables (socioeconomic status). Stratifying salivary DNAm samples by BEC proportions detected a larger number of significant associations with biological variables (sex) and tissue-specific effect with cotinine level, a tobacco smoke-exposure biomarker. We provide analytical recommendations for future epigenetic research involving pediatric saliva samples.

Keywords: pediatric saliva; DNA methylation; cell type heterogeneity; epigenome-wide association study; reference-based cell deconvolution

Considerations for Cell Type Heterogeneity in Pediatric Salivary DNA Methylation Analyses: Comparison of Reference Panels & Stratification by Estimated Cell Type Proportion

Epigenetics has gained increasing interest as a potential biological mechanism through which environmental exposures and other biopsychosocial factors can contribute to long-term health and development. DNA methylation (DNAm) is an epigenetic mark that is well-characterized, easily accessible, establishes and maintains cellular identity, and commonly studied in humans (Jones et al., 2018). Most human DNAm research involves bulk tissues, which are composed of different cell types (CTs). Since DNAm is integral in cellular differentiation and maintenance, each CT has a distinct DNAm profile, and therefore DNAm associations with environmental exposures and health states are highly CT-specific (Jones et al., 2018). To address this issue, CT proportions can be bioinformatically estimated from DNAm data with a reference panel and adjusted for in an EWAS model (Langie et al., 2017; Middleton et al., 2022; Zheng et al., 2018). Tissue- and occasionally developmentally-specific (i.e., pediatric vs. adult) CT reference panels were developed for accurate CT estimations in different tissue sources. As such, one key consideration in epigenome-wide association studies (EWAS) identifying DNAm links to exposures is to account for cellular heterogeneity in statistical analyses, among other biological and technical covariates known to drive DNAm variations (Qi & Teschendorff, 2022).

In most human studies, DNAm are obtained from easily accessible, heterogenous, peripheral surrogate tissues. Oral samples, like saliva, are preferred in biosocial and/or epidemiological research involving children since collection methods are minimally invasive, create low levels of discomfort, and are more accepted by parents (Hamilton et al., 2022; Nemoda, 2020). Additionally, it can be collected at home and mailed at room temperature, resulting in lower demand on technical training (Nemoda, 2020). Although saliva as a biological

sample has unique strengths and has generated considerable opportunities in pediatric DNAm investigations, its high CT heterogeneity presents an analytical challenge that needs to be addressed to ensure rigorous DNAm analyses.

Saliva is composed of a heterogeneous cell population, containing both buccal epithelial cells (BECs) from the oral mucosa and immune cells, including those resides in the mucosal immune system and others that translocate from blood vessels into the oral cavity (Nemoda, 2020). Furthermore, salivary cellular composition is highly variable across individuals, ranging from 20 and 100% of BECs (Theda et al., 2018), which could potentially be explained by social exposure and oral hygiene (Slavish et al., 2015; Tsukamoto & Machida, 2014). Importantly, salivary CT proportion can also vary across the lifespan (Wong et al., 2022). Specifically, BEC proportion decreases while immune cells proportion increases with age potentially, in part, due to increasing prevalence of gingivitis and other oral health issues (Merrill et al., 2024; Theda et al., 2018; Wong et al., 2022). Alternatively, such proportional change may be a result of cellular structural differences such as changes in the shape of oral neutrophils with age (Merrill et al., 2024). Therefore, DNAm markers associated with these cells could also be different with age.

Since salivary CT composition and the DNAm markers associated with CTs could potentially be developmentally sensitive (Wong et al., 2022), the age of the samples with which the reference panels were developed could be highly relevant to salivary CT proportion estimation. Choosing reference panels developed from samples most characteristically similar to the study samples could help improve the accuracy of CT proportion estimation and enhance the parsing of DNAm variation contributed by CT proportion variability. The reference panel primarily used for estimating CT proportions in saliva is a reference constructed from eleven adult epithelial cell lines (EpiFibIC reference) available in the EpiDISH R package (Zheng et al., 2018). Recently, a reference panel specific to pediatric saliva was developed on 22 children’s

salivary samples (aged 7-16 years) (Middleton et al., 2022). When compared across the two reference panels with the same salivary DNAm, BEC proportion estimated with the child saliva reference panel was significantly lower and had a greater range than that estimated with the adult reference panel (Middleton et al., 2022). Such discrepancies were speculated to impact downstream DNAm analyses and the identification of differentially methylated sites in saliva (Langie et al., 2017).

Given the large range and variability of CT proportions in pediatric saliva, research must also question whether statistically adjusting estimated CT proportions in DNAm analyses is sufficient. One could speculate that a saliva sample with 0% BECs (and 100% immune cells) will behave more similarly to tissues of primarily immune cell composition, like blood, while those with 100% BEC may be more similar to tissues with a largely epithelial compositions, such as cheek swabs. Therefore, the differences between these salivary samples may even parallel tissue differences. This issue likely has a differential impact on DNAm analyses depending on the variable of interest and could be particularly pertinent for variables of interest that may mainly affect a specific tissue or are more strongly associated with a certain CT. For example, some exposures, such as pollution and stress, and health-related behaviors, like smoking, may be more associated with DNAm in immune cells than BECs in saliva due to their role in inflammation and immune regulation (Bauer et al., 2016). This could make interpretation of DNAm findings from salivary samples with such high CT heterogeneity challenging and alternative solutions may need to be considered, such as stratifying the salivary samples by CT proportions to reduce noise that originated from cell type heterogeneity and generate more replicable and interpretable results.

Thus, the current study used 529 salivary samples collected from 7-year-olds to address these analytical challenges. First, we characterized and compared the CT proportions predicted

with the adult and child reference panels to elucidate potential differences in estimated CT proportions emerging from reference panels. Second, we examined the effect of using different estimated CT proportions on EWAS outcomes. To maximize the applicability of our findings across disciplines, and to illuminate the importance of choosing an appropriate reference for DNAm analysis, we chose three variables that are commonly of interest in biosocial research, spanning across the spectrum of biological to social. These included 1) biological sex, a variable of primarily biological influence, 2) salivary cotinine level, which is indicative of tobacco smoke exposure and a variable of a combination of biological and social influences, and 3) SES, a variable of primarily social influence. All these variables have been widely studied in the context of DNAm across disciplines (Bauer et al., 2016; Govender et al., 2022; Lam et al., 2012; McDade et al., 2017). Third, we compared these EWAS results from all samples (with high intersample estimated CT heterogeneity) and stratified samples (with fewer intersample estimated CT heterogeneity) to inform on the potential need for stratifying salivary DNAm data. Similarly, we conducted these analyses across the three variables of interest with varying degree of expected CT-specific DNAm associations (Bauer et al., 2016). For example, given the link between smoking and immune cell activity, we expected that stronger associations in the stratified samples with high immune cell proportion could be observed. The current findings served as an important foundation for the consideration of CT heterogeneity in pediatric saliva for DNAm investigation.

Methods

Participants

The current study used archival salivary samples collected as part of the Family Life Project (FLP) 7-year follow-up assessment. As previously described (Vernon-Feagans, 2014), a representative sample of 1,292 families were originally recruited at the time that mothers gave

birth. A subset of 743 FLP families provided consent for analysis of the salivary samples. After quality control (QC), DNAm data from 529 children (mean age = 7.26 years, SD = 0.26 years, 49.9% boys) and survey data from their primary caregivers (98.4% biological parents) were used in our analyses (**Table 1**). There were no statistically significant differences between the FLP sample at 2 months, which was the time point with the most complete data (n=1292), and our current sample (n=529) in percentage of low-income families (75.2% in current sample vs. 77.6% in whole sample, p = .313), mother-reported child race as Black/African American (39.7% in current sample vs. 42.6% in whole sample, p = .330), and child biological sex (49.9% in current sample vs. 49.1% in whole sample, p = .644) based on chi-square tests. There was a small but significant difference in age (current sample = 7.26 vs. whole sample = 7.29, p = 0.03) between the whole and current samples.

Table 1. Participant characteristics (n = 529)

Variables	**n**	**Mean**	**SD**	**Range**
Age (years)	529	7.26	0.26	6.56-8.12
Biological sex	529	49.9% male	–	–
Mother-reported child race	529	59.5% white 39.7% Black/African America	–	–
Family income to needs ratio	529	1.91	1.61	0-11.18
Salivary cotinine (ng/mL)	394	2.53	3.74	0.16-47.08

Parent-Reported Measures: Caregiver-reported socioeconomic status & children's sex

Primary caregivers reported on children's biological sex at birth, which was confirmed by the DNAm (described in DNAm preprocessing) in the final samples used for analyses. Primary caregivers also reported their income from all sources and any income from other household

members, which was used as an estimate of total household income. Income-to-needs ratio, as an index of socioeconomic status (SES), was calculated by dividing total household income by the federal poverty threshold for the year 2011, adjusted for the number of persons in the home. An income-to-needs ratio of 1.00 or below indicates family income at or below the poverty level, adjusted for household size (Vernon-Feagans, 2014).

Collection of salivary samples

Unstimulated whole saliva samples were collected from children using the passive drool method (Granger et al., 2012) during the 90 month visit to participants' homes. At the time of collection, samples were frozen at -20C and then transferred to the Institute for Interdisciplinary Salivary Bioscience Research at the University of California, Irvine for archiving at -80˚C until used. At the time of use, saliva samples were thawed and centrifuged (5,000 g; 10 min; 4C) to remove insoluble material and cellular debris. Supernatants were collected and used for other assays separate from the current study. The resulting cell pellet was stored at -80°C until DNA extraction.

DNA extractions

Pellet fractions from the centrifuged saliva samples were resuspended in 200 ul of PBS. DNA was then isolated using the QIAamp DNA Mini Kits (QIAGEN, Cat #1304) following the manufacturer's instructions, with the exception that twice as much Proteinase K was used and samples were incubated with Proteinase K for 2 hours. DNA quality and quantity was determined using a NanoDrop Spectrophotometer (ThermoFisher).

DNA methylation profiling by Illumina MethylationEPIC BeadChips

DNAm profiling by the llumina MethylationEPIC BeadChips (EPIC array, Illumina, CA, USA) were performed at the University of British Columbia (Vancouver, BC, Canada) as

previously described (McEwen et al., 2018). Briefly, 750 ng of genomic DNA was bisulfite-converted using EZ DNA Methylation Kit (Zymo Research, CA, USA). Subsequently, 160 ng of bisulfite-converted DNA was applied to the EPIC array, following manufacturer's protocols (Illumina, CA, USA). Processed EPIC arrays were scanned with an Illumina iScan (Illumina, CA, USA).

DNA methylation preprocessing

QC of the EPIC array DNAm data was performed in R (4.2.2) using a combination of the `minfi` package (v1.44.0), which plots the log median intensity of a sample in both the methylated and unmethylated channels, and `ewastools` (v1.7.2), which provides seventeen QC metrics as described by Illumina in their BeadArray Controls Reporter Software Guide document packages. A total of 200 samples were removed after all QC steps (see **Figure 1a** for details). One possible reason for this high number of removed samples is bacterial contamination as indicated by the bacterial DNA assay showing lower cycle threshold (Ct) value (i.e., faster amplification) in a subset of the failed samples (mean = 10.95) than samples that passed QC (mean = 11.71), $t(24) = 3.08$, $p = .005$. Four additional samples with age mismatch (>10 years difference) based on the epigenetic age calculated with the Pediatric Buccal (PedBE) and Pan-tissue Horvath clocks in the `methylclock` R package and reported chronological age were removed. A total of 529 high-quality samples were included for downstream analyses (**Supplementary Figure S1**). The removed samples did not differ from the final sample on demographic variables, including child sex, low-income status, and parent-report child ethnicity (*p*s = .601 - .962).

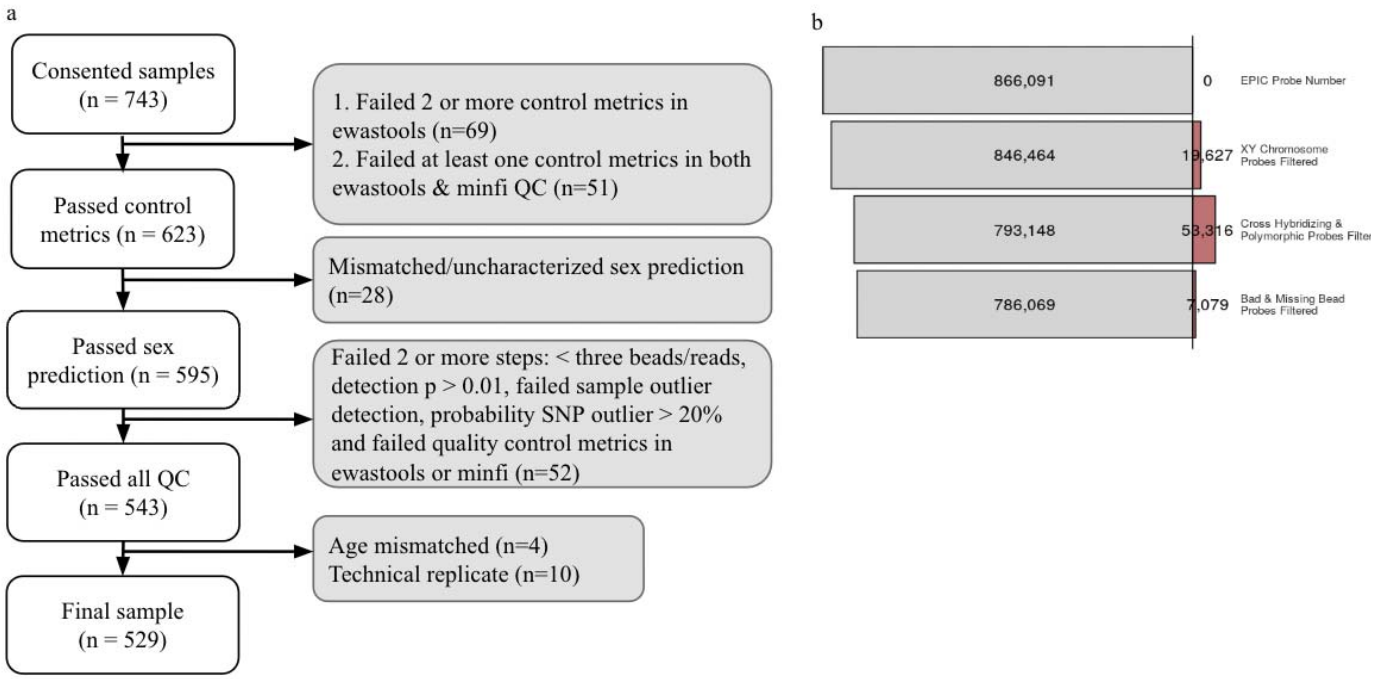

Figure 1 DNA methylation preprocessing. Quality control (QC) was performed on the DNA methylation data. a) QC was performed on a total of 743 consented samples and the number of samples failed at each step. A final high-quality sample of n=529 individuals that passed all QC steps were included in the analyses in the current study. b) The number of probes filtered at each filtering step. A total of 786,069 probes remained after probe filtering.

To account for type I and type II probe differences on the EPIC array, DNAm data was normalized using functional normalization by *minfi::preprocessFunnorm*. Probe filtering was conducted on the normalized DNAm data (see **Figure 1b**). In total, 41,787 cross-reactive probes were removed. Next, polymorphic probes containing a single nucleotide polymorphism (SNP) at the target DNAm site (n probes = 11,270) or at the base extension sites (48 and 49 base pairs for Infinium Type I and II probes, respectively) (n□=259) with minor allele frequency (MAF) >5%, based on the 1000 genome project, including genotypes from 26 different populations, were also removed. Then, DNAm sites which overlapped with probes located on the sex chromosomes, and probes with a detection p-value > 0.01, as well as those with NAs in more than 2% of

samples using *watermelon::pfilter* were also removed. A total of 786,069 probes were included for downstream analyses. We performed *sva::ComBat* to account for remaining technical effects on plate and chip (Jones et al., 2018).

Whole genome genotyping & genetic principal components

Among our high quality DNAm samples (n=529), 468 samples also had genotype data. Illumina Global Screening v1.0 arrays were used to generate 654,027 SNP genotypes. QC was conducted as in Simmons et al. (2011), including individual/SNP genotype characteristics, relationship checking, and ancestry. Briefly, SNPs with MAF < 1%, SNP-level missingness > 1% or a Hardy-Weinberg equilibrium test with $p < 1x10^{-6}$ were filtered. Individual samples with excess/reduced heterozygosity, individual-level missingness > 1%, or incorrect sex determination were filtered.

We analyzed population ancestry using principal components analysis (PCA) calculated with the smartpca function from the EIGENSOFT package that computes PC scores from SNPs (Price et al., 2008). We analyzed all samples that passed quality control along with all 1000 Genomes samples to provide clear reference populations for the major continental groupings. The first three principal components were visualized graphically for our samples along with the 1000 Genomes samples, all color-coded by ancestry. African American samples displayed variation in the degree of admixture, as expected. As many of these samples did not clearly fall into either continental population, according to Price et al. (2008), it is necessary to use the quantitative information on ancestry to control for this potentially confounding variation. Controlling for these effects of genetic admixture, we included the first three PCs to in all analyses.

Salivary cotinine measure

Salivary cotinine was assayed in duplicate using a commercially-available, ELISA kit (Salimetrics LLC, Carlsbad, CA, Cat ##1-2002) following the manufacturer's protocol and per our previous studies (Gatze-Kopp L, J Expo Sci Environ Epidemiol, 2023). This assay has a test volume of 20 µL, range of standards from 0.8 to 200 ng/L, and lower limit of detection (LLD) of 0.15 ng/L. Cotinine high and low controls were run on every assay plate. The intra-assay and inter-assay coefficients of variation (CVs) are 7.1% and 9.1%, respectively. Concentrations below the LLD (n=135) were treated as missing data in the main analyses. A concentration below 10 ng/ml is indicative of potential passive environmental tobacco exposure (Rao et al., 2023), which is in line with our observation in this cohort (mean = 2.53 ng/ml).

Statistical analyses

Characterizing and comparing cell type proportion estimates across adult and pediatric saliva reference panels

An overview of the analytical plan is presented in **Figure 2**. First, we estimated the CT proportions of BECs and immune cells with robust partial correlation (RPC) and both an adult (Zheng et al., 2018) and a pediatric saliva-based reference panel (Middleton et al., 2022). For the adult reference, three major subsets (epithelial cells, fibroblasts, and immune cells) were first deconvoluted with the *EpiDISH::*centEpiFibIC.m reference matrix. Subsequently, we used *EpiDISH::hepidish* to estimate seven cell type proportions based on epigenomic deconvolution (B-cells, CD4+ T-cells, CD8+ T-cells, NK cells, monocytes, neutrophils, and eosinophils) with the *EpiDISH::*centBloodSub.m reference matrix. As for the pediatric reference, the saliva samples were first deconvoluted into the BEC versus immune cell proportions, followed by deconvolution of the immune cells also with the *EpiDISH::*centBloodSub.m reference matrix. We assessed the median and interquartile range (IQR) as well as the variance and distributions of

estimated BEC proportions. We also examined the appropriateness of the reference panels using the CEll TYpe deconvolution GOodness (CETYGO) score (*CETYGO*::projectCellTypeWithError), a metric to assess the accuracy of cellular deconvolution when actual cell count is not available (Vellame et al., 2022). The CETYGO score was defined as the root mean squared error (RMSE) between the measured DNAm profile in the sample and the expected profile from the reference panels. Therefore, a perfect estimate would have a CETYGO score of 0 (i.e., lower scores are more accurate), while higher values reflect less accurate estimations of CT compositions. A CETYGO score < 0.1 indicates an appropriate reference while > 0.1 suggests the reference panel may not be relevant for the tissue being profiled (Vellame et al., 2022).

Examining the effect of using different estimated cell type proportions as covariates on EWAS outcomes

To evaluate the effect of different estimated CT proportions on EWAS outcomes, a series of EWASs were conducted. To reduce the dimensionality of our CT proportion variables, we performed isometric log-ratio transformation on the estimated CT proportion, followed by robust principal component analysis (PCA) with robCompositions::pcaCoDa. The first two PCs explained > 90% of the variance (94.6% for EpiDISH; 97.2% for Middleton) and were included in the EWAS models. We selected variables that were previously associated with DNAm and represented a spectrum of biological and social factors to provide insight into a range of research areas, including biological sex, salivary cotinine level, and SES (**Figure 2**). To address the potential effect of the missing data in salivary cotinine on our results, sensitivity analyses were conducted to account for the censored data below the assay's LLD (0.15 ng/L) by imputing these data with half the LLD of the assay (i.e., 0.075 ng/L; total sample N for these analyses = 523

with mean = 1.93 ng/L, range = 0.075 - 47.08 ng/L). Given that the main results did not differ substantially between the analyses with these two variables, the results of the sensitivity analyses were included in the **Supplementary Figure S6**.

A total of 614,686 variable probes, where the DNAm level (β value) varies by at least 5% across samples in the 5th and 95th percentile, were included for the EWASs. Two EWAS models were run for each variable of interest, adjusting for different estimated CT proportions, either with a) a child or b) an adult CT reference panel (six EWAS models in total). The other covariates for both models were the same - the first three genetic PCs and child's age. The robust linear regression model used for the EWAS was:

$$CpG \sim Variable\ of\ interest + Age\ at\ Sample\ Collection + Genetic\ PC1 + Genetic\ PC2 + GeneticPC3 + Cell\ Type\ PC1 + Cell\ Type\ PC2 + \varepsilon$$

Bias and inflation in our EWAS models was assessed with BACON, a Bayesian method (van Iterson et al., 2017). Differences in DNAm associated with the variable of interest was quantified by the difference in beta value (Δβ) (Jones et al., 2018). In our analyses for chromosomal sex, Δβ was the regression coefficient of the variable from the robust linear regression described above. For cotinine and SES, Δβ was calculated by extracting the regression coefficient of the variable of interest from the robust linear regression models described above and then multiplying the coefficient with the range of cotinine or SES values between the 5th and 95th percentiles to reduce the effects of outliers. The resulting Δβ value represents a change in DNAm β value at each site associated with cotinine or SES while adjusting for other covariates.

We used the Benjamini-Hochberg False Discovery Rate (FDR) control method for multiple-test correction (Benjamini & Hochberg, 1995). A statistical threshold was set at FDR < 0.05 to determine significant associations between DNAm at each site with each variable of interest. We also set a technical threshold of |Δβ| > .035, which was greater than technical noise

as operationalized by the RMSE of the technical replicates after preprocessing (0.023-0.034). Therefore, the effect sizes of the DNAm sites that passed these thresholds were likely to be greater than technical noise. Across the two reference panels, we compared the sites that were identified to be statistically significantly associated with each variable of interest.

Comparing EWAS results between stratified samples by BEC proportions

To answer the question whether CT proportions influence EWAS outcomes despite statistically adjusting for it in the model, we conducted the same EWAS on stratified samples based on BEC proportions. Using a data-driven approach, we stratified the samples into three groups – low-, mid-, and high BEC groups (**Table 2**), consistently suggested by both Jenk's natural break (Jenks, 1977) and k-mean clusters (Likas et al., 2003).

Table 2. Summary statistics across BEC stratified groups

Stratified subset	**BEC proportion**	**n**	**% of male**	**Cotinine (**ng/mL) **mean (SD)**	**Income-to-needs ratio mean (SD)**
Low BEC	0 - 27.5%	123	52.8%	3.01 (5.41)	1.80 (1.64)
Mid BEC	27.9% - 70.0%	211	49.3%	2.14 (2.54)	2.03 (1.74)
High BEC	70.3 - 100%	195	47.7%	2.63 (3.36)	1.83 (1.43)

SD = standard deviation; BEC= buccal epithelial cell;

SES= socioeconomic status

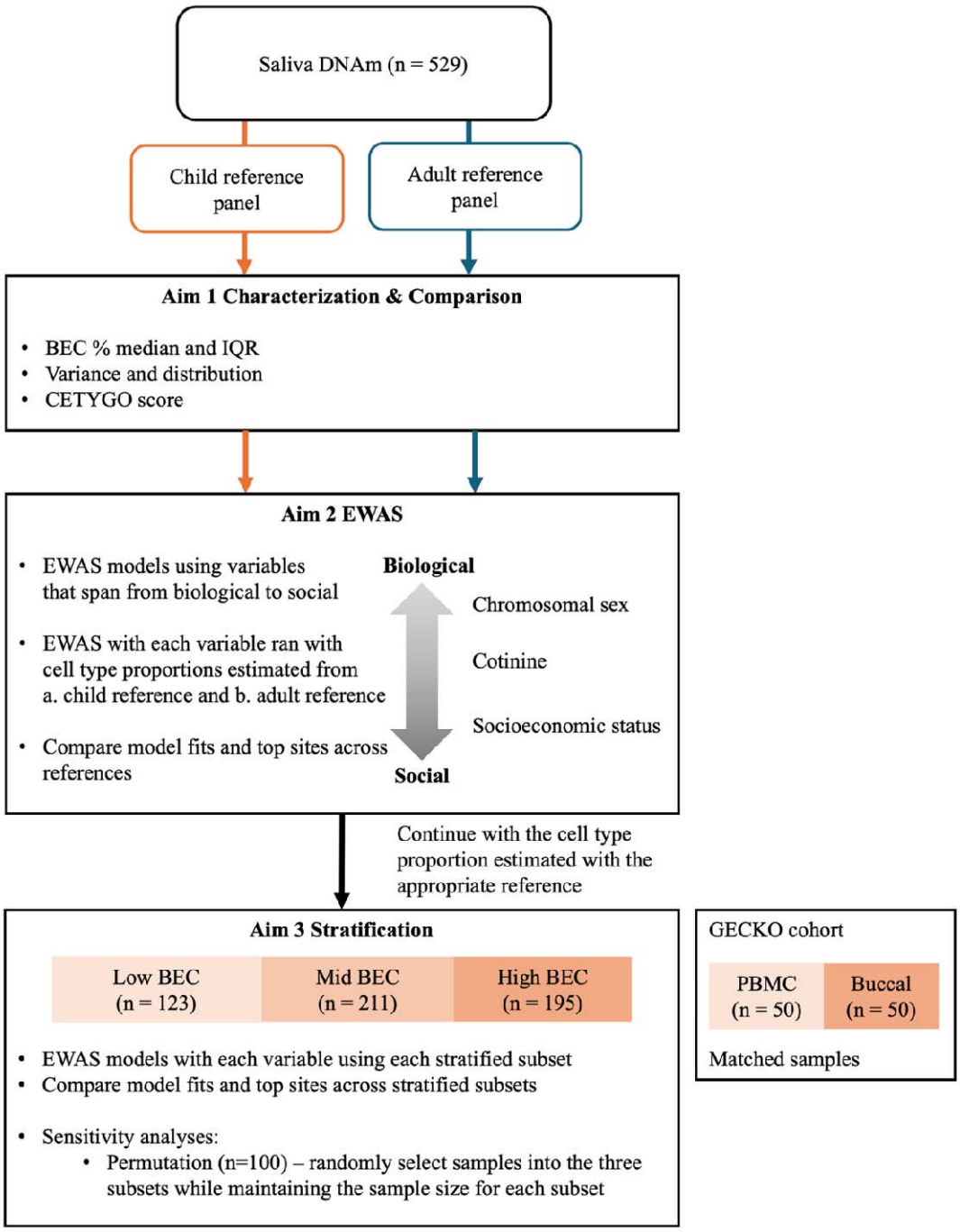

Figure 2 Study Overview. Cell type (CT) proportions were estimated with both a child and an adult salivary reference panel. The first aim of the study was to compare and characterize the estimated CT proportions across reference panels. The second aim was to examine the effect of

CT proportions estimated from different reference panels on EWAS results of chromosomal sex, cotinine levels, and socioeconomic status. The third aim was to investigate the effect of stratification by buccal epithelial cell (BEC) proportion on EWAS results of the same variables.

With the same three variables of interest (biological sex, cotinine level, and SES), one EWAS for each variable was run with each stratified subset (i.e., nine EWAS models). Using similar criteria as above, we assessed whether differences in BEC proportions across the subsets could be sufficiently accounted for by statistically adjusting CT proportions in an EWAS model. We reported the unique number from each BEC subset and the number of overlapping sites across the three subsets associated with each variable of interest. Since the sample sizes of the three subsets were unequal, we ran 100 randomized permutations in which we randomly assigned an individual to one of the three BEC groups, independent of estimated BEC proportions. With the new subsets in each permutation, we ran the same EWAS as described above (i.e., 100 permutations for each of the three EWAS). The resulting value is the permutation p-value that represents the likelihood that the unique number of associations of each group as well as the number of overlapping associations across all groups were observed due to chance. If the BEC group with a larger sample size showed a higher number of significant associations, we also ran 100 permutations to randomly draw a sample matching with the size of the smallest group. The resulting permutation p-value represents the likelihood that a smaller number of associations are observed with a smaller sample size.

Validation of Cross-Tissue Comparison

To compare the extent to which these EWAS outcomes across stratified subsets may be akin to tissue effect, we leveraged the 50 matched buccal and PBMC samples from the GECKO cohort, a Canadian cohort of children aged 7-13 years old, previously described elsewhere

(43.1% female, 72.5% White as reported by parents) (Chan et al., 2023; Islam et al., 2019). EWAS on chromosomal sex was conducted separately in the buccal and PBMC samples while adjusting for similar covariates as the primary analyses, including child age, genetic ancestry, and batch effect. Significant associations were determined by the same thresholds across tissues (statistical: FDR < .05 and technical: $|\Delta\beta| > .05$). The number of overlapping significant DNAm sites across the GECKO matched tissue samples was compared with that across the FLP stratified subsets.

Results

The child reference panel is more appropriate than the adult reference panel when estimating salivary cell type proportions in pediatric samples

BEC proportions were estimated to be the highest and the most variable among all CT proportions with both reference panels (**Supplementary Figure S2**). Furthermore, the estimated proportion of immune cells, primarily neutrophils, had a significant and strong negative correlation with estimated BEC proportion, $r(527)= -0.93$, $p < 2.2e-16$). Therefore, estimated BEC proportion was likely to capture most of the variability of estimated CT proportions in saliva and were used in our remaining analyses. Although the estimated BEC proportions was significantly correlated across reference panels (Pearson's $r = 0.950$, $p < 2.2e-16$), the median and intersample variability differed (**Figure 3**). Specifically, the child reference panel estimated a lower median of BEC proportion (51.34%) and higher IQR (71.18%) than the adult reference panel (median: 80.46%; IQR: 42.29%). Estimated BEC proportion means were different as indicated by paired-sample t-test, $t(528) = 31.788$, $p < 2.2e-16$. Additionally, the distributions were significantly different from each other as suggested by an asymptotic two-sample

Kolmogorov-Smirnov test, D(528) = 0.29, $p < 2.2e-16$. Furthermore, the CETYGO score of the child reference panel was lower than 0.1 (mean = 0.074, range = 0.03-0.09), while the CETYGO score of the adult reference panel was higher than 0.1 (mean = 0.15, range = 0.05 – 0.20), suggesting that the child reference panel is a more appropriate panel for pediatric samples in contrast to the adult reference panel.

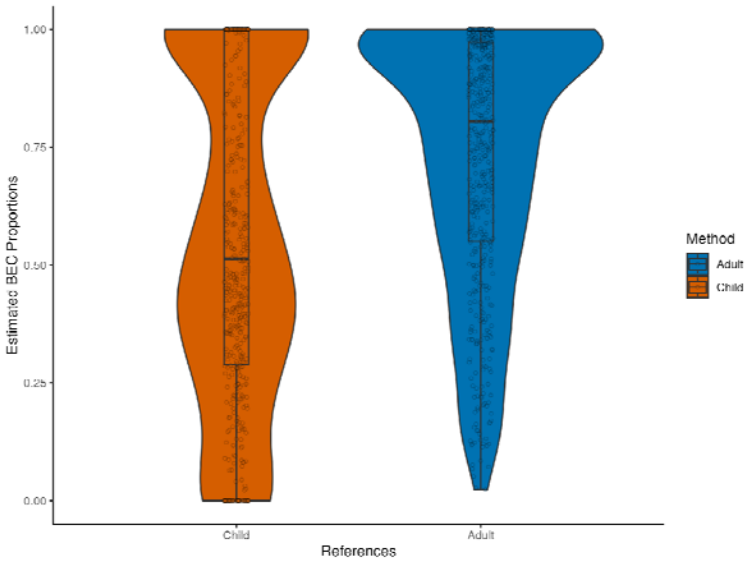

Figure 3 Estimated buccal epithelial cell (BEC) proportions with child and adult reference panels. A violin plot of BEC proportions in pediatric saliva as estimated with child and adult reference panels using the DNAm-based deconvolution tool EpiDISH. Orange (left) reflects the child reference panel, and blue (right) reflects the adult reference panel.

Reference panels produced different outcomes for an EWAS of socioeconomic status and similar results for biological variables

To assess the capability of the different reference panels to account for test statistic inflation in EWAS and allow for accurate evaluation of the association between variables of interest and DNAm, we performed different EWASs, in which we used CT composition derived

from either reference panels. We compared the EWAS results for outcome variables along the spectrum of biological and social (see **Figure 2**). EWAS for more biological variables, sex and cotinine concentration, had high overlaps in statistically significantly associated sites (FDR < .05 and $|\Delta\beta| > .035$, 92% and 100%, respectively) and comparable model fits for estimations from both references (**Figure 4a & b; Supplementary figures S3 and S4**). However, differences in EWAS outcomes with SES (social variable) were observed across reference panels (only 18% overlap). The SES EWAS model adjusting for the child reference panel estimated CT proportions has a higher number of unique significant DNAm associations (76 and 4, respectively) and better model fit (**Figure 4c and Supplementary figures S3 and S4**) than that adjusting for the adult reference panel. Nonetheless, the majority (80%) of the significant DNAm associations found with adult reference panel was also present in the results with the child reference panel.

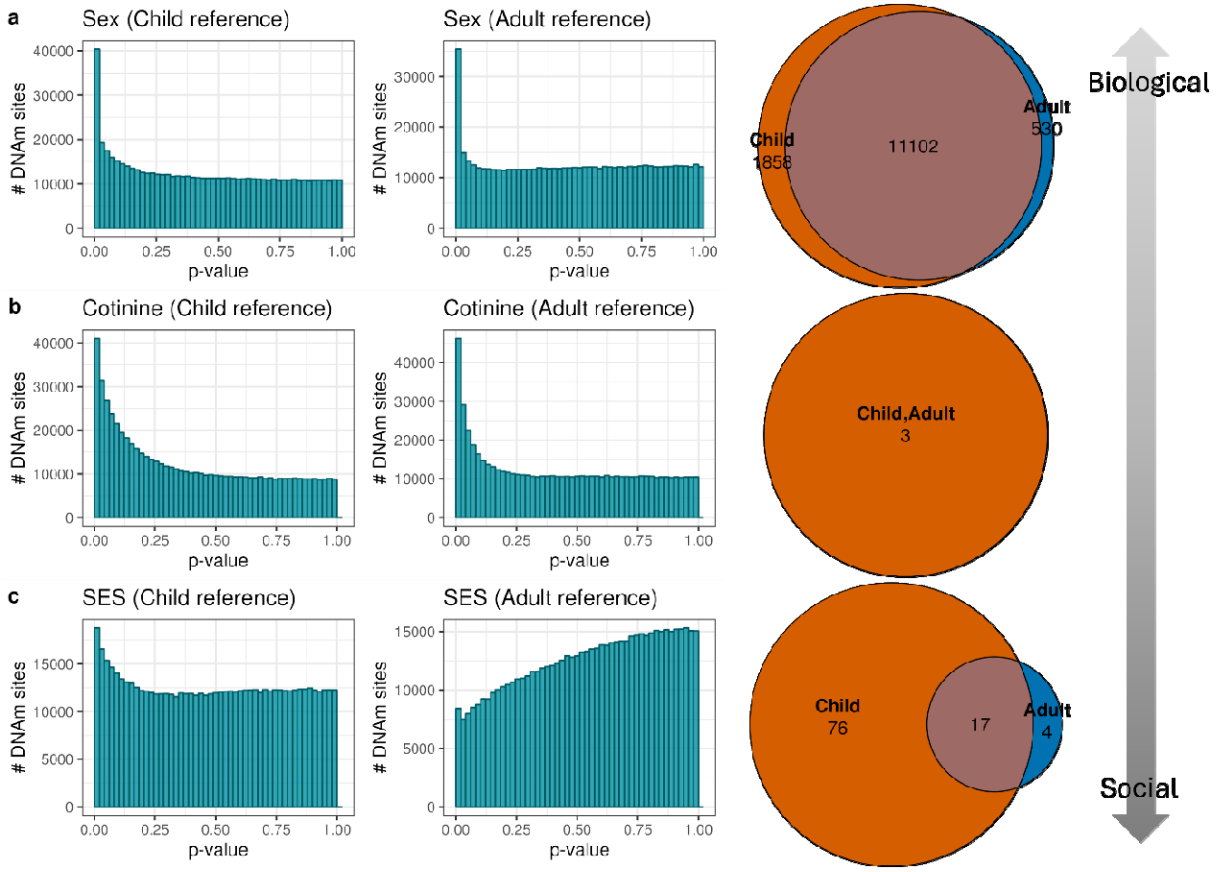

Figure 4 Number of overlapping significantly associated DNA methylation sites with variables of interest spanning from biological to social. Each Venn diagram represents the unique and overlapping sites associated with each variable of interest (a: sex, b: cotinine concentration, and c: socioeconomic status) across the child and adult reference panels. The range of Δβ extracted from EWAS (with child-reference estimated cell type proportions) also varied across variables (Sex: -0.394 to 0.290, cotinine: -0.095 to 0.090, SES: -0.0967 to 0.147)

Stratifying samples by estimated BEC proportion detected larger number of significant associations with biological sex and cotinine but not socioeconomic status

We examined whether statistically adjusting for estimated CT proportions sufficiently accounted for saliva CT heterogeneity, especially given the large variance observed in BEC

proportions estimated by the child reference panel. To determine if stratifying and analyzing these samples by their most abundant CT (i.e., primarily BEC, primarily neutrophils, or an approximately even mix of both) may result in differing EWAS results, we ran the same analyses on three stratified subsamples by estimated BEC proportion, herein described as high, mid, and low BEC subsamples.

EWAS for sex performed in the high and low, but not the medium BEC subsamples, showed an increased number of significant associations (n associations=3,007 and 1,980, respectively) compared to the full sample (n associations =1,780) (**Figure 5a**), indicating an increase in power to detect DNAm associations with sex when samples were stratified by estimated BEC proportion. Across the three BEC samples, the high BEC subsamples had the largest number of significant associations with sex, while the medium BEC subsamples, despite having the largest sample size, had the least number of sites associated with sex (**Figure 5a**). To assess whether the difference in the amount of statistically significant associations were driven by differences in sample sizes (given that the high BEC subsamples [n = 195] was higher than the low BEC subsamples [n = 123]), we conducted 100 permutations of the EWAS on biological sex by randomly drawing 123 samples (i.e., the lowest sample size involved in the EWAS comparisons) from the high BEC subsamples. The results of the permutations suggested no difference in the number of statistically significant associations found in high BEC subsamples when the sample size was reduced from n = 195 to n = 123 (median n associations = 1,009; p = 0.14).

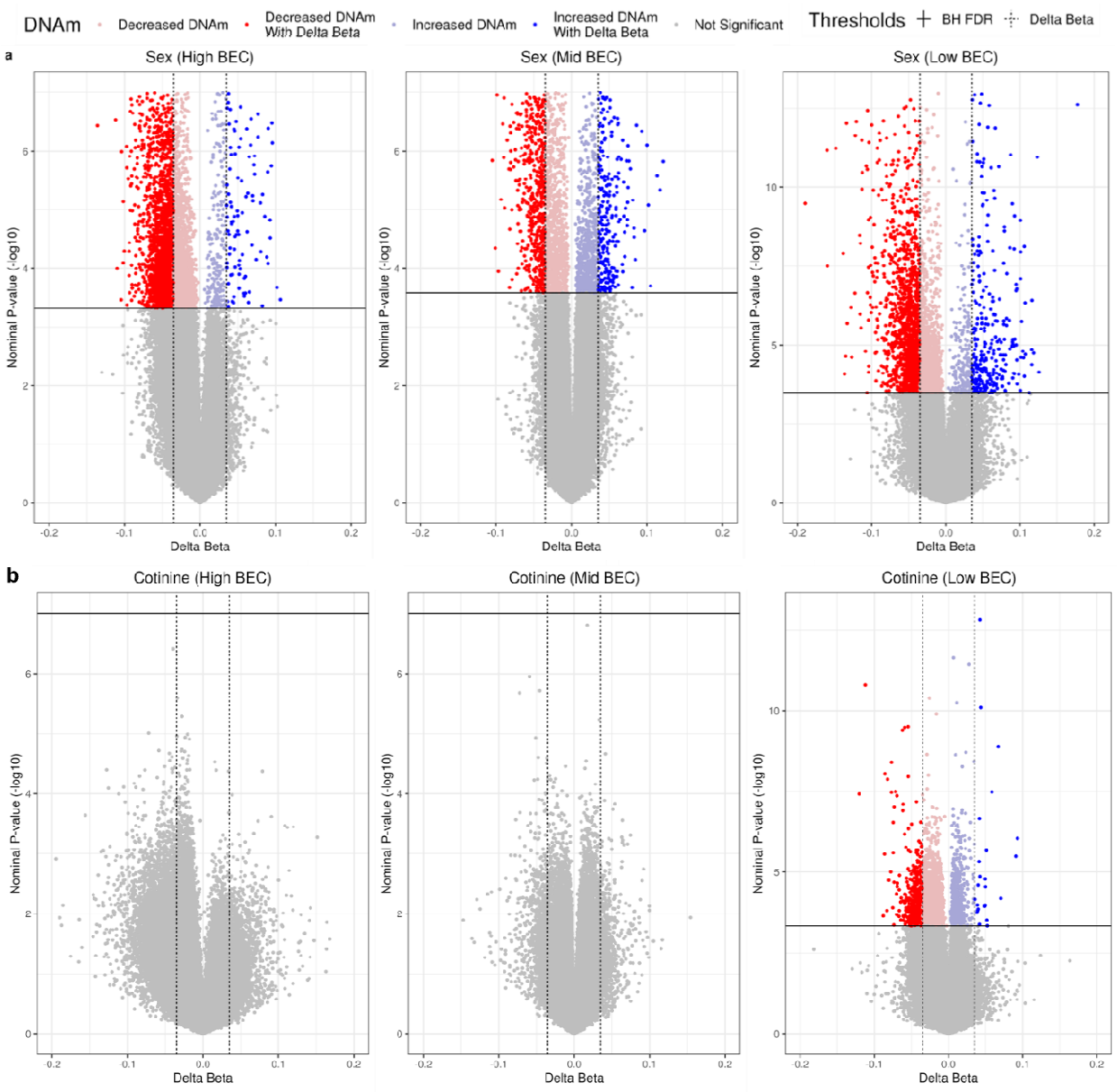

Figure 5. DNA methylation associations with biological sex and salivary cotinine across all three BEC subsamples. Volcano plots displaying DNAm association with a) biological sex and b) salivary cotinine in the low, middle, and high BEC subsamples. The colored dots represented significant DNA methylation associations with biological sex, passing the statistical threshold of FDR < .05 and $|\Delta\beta| > .035$. Red dots represent DNAm sites with a negative association, whereas blue dots represent DNAm sites with a positive association.

We observed 11.9% overlapped significantly associated sites (559 sites) across the three subsamples (**Figure 6a**). Next, we determined whether the three groups yielded significantly different sets of DNAm associations. In other words, we tested whether it is likely to find a percentage of overlapping associations lower than 11.9% by chance. We performed 100 randomized permutations, which showed that a percentage of overlapping associations as low as or lower than 11.9% across the three groups were not due to chance or random grouping of samples (p = 0.03; median = 26.00%; range = 5.22-36.45%). Similarly, the unique number of associations (3,007 sites) in the high BEC was unlikely to be observed by chance with random grouping (p = 0.03).

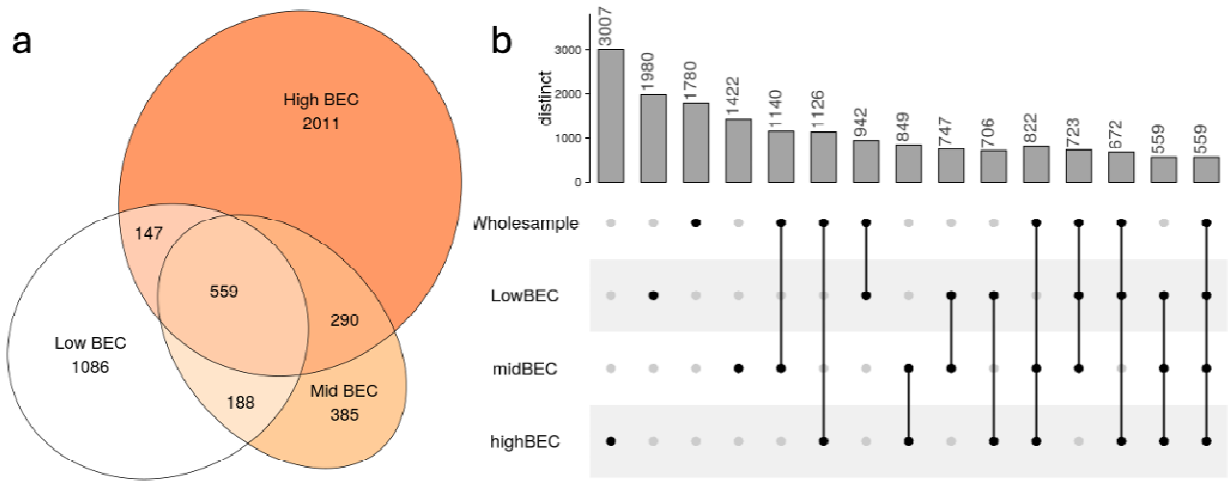

Figure 6 Significant associations with sex across stratified and full samples. EWAS for sex in the high, mid, and low BEC subsamples adjusting for cell type proportions estimated by the child reference panel. Panel *a* depicts the unique and overlapping significant associations across the three BEC groups. Panel *b* shows the unique and overlapping significant associations in the stratified samples compared with the full sample.

To examine whether the overlapping associations across the three BEC subsamples were comparable to that across tissues, we leveraged matched buccal and PBMC samples from GECKO, our validation cohort, which showed a lower percentage of overlapping DNAm associations with sex (5%, **Supplementary Figure S5**) than that across the three BEC subsamples in our main analyses with the FLP, the discovery cohort.

When examining DNAm associations with cotinine levels in the three different BEC subsamples, we only observed significant associations in the low BEC group (**Figure 6b; Supplementary Figure S6**). Interestingly, despite having a smaller sample size, a higher number of sites were significantly associated with cotinine in the low BEC stratified sample compared to the full sample (n associations = 784 vs. 3, respectively). One hundred randomized permutations indicated that the number of unique statistically significant associations in the low BEC subsamples was not significantly higher than by chance ($p = 0.09$). Lastly, we did not find any significant associations with SES when the samples were stratified by BEC proportions (see **Supplementary Figure S7**).

To explore whether DNAm patterns associated with sex, cotinine, and SES differed across stratified samples, we compared the $\Delta\beta$ of all DNAm sites in the low and high BEC subsets and identified sites with diverging DNAm effects across stratified samples (i.e., $\Delta\beta > .035$ in one and $< -.035$ in another) (**Supplementary Figure S8**). We identified 145, 473, and 888 diverging DNAm sites for sex, cotinine, and SES, respectively.

Discussion

The current study addressed the analytical challenges of EWAS in high CT heterogeneity pediatric salivary samples. To this end, we compared two existing reference panels for estimating CT in pediatric saliva – a commonly employed adult epithelial and immune cell

reference and a child saliva reference – and two analytical approaches to account for intersample CT heterogeneity – adjusting for estimated CT proportion in full or stratified samples in addition to the CT adjustment. To summarize, we found the pediatric reference panel was more appropriate than the adult reference panel for estimating CT proportion in pediatric saliva and predicted a higher interindividual CT variability and a lower mean BEC proportion. Subsequently applying the estimates from the child reference panel for statistically adjusting for CT in EWAS produced a well-behaved p-value histograms on the social variable (i.e., SES), and stratifying samples by BEC detected CT-specific associations with DNAm for biological variables (i.e. sex and cotinine).

Reference panels influenced estimated epithelial cell proportions

Statistically adjusting for estimated CT proportions has become a common practice in EWAS analyses to account for the influence of CT heterogeneity in easily accessible bulk tissues. Two reference panels have been developed that may be suitable for pediatric saliva based on different populations (adult vs. child samples) and tissues (eleven epithelial cell lines vs. whole saliva samples). We investigated to what extent the CT estimations from these reference panels were comparable with each other. The estimated BEC proportions in our current childhood saliva sample were lower and more variable when estimated with the child reference compared to the adult reference and showed highly comparable estimations with findings from a previous study (i.e., BEC medians = 51.3% and 80.5%, estimated with child and adult reference, respectively, in the current sample, and 56.1% and 81.7%, respectively, in Middleton et al., 2022). Since there was no external validation cohort for the generation of the pediatric saliva reference panel, the current study validated the this previously reported patterns and demonstrated that such discrepancy between child and adult reference estimated CT was not only

found among the samples used to develop the saliva reference panel (Middleton et al., 2022). The current estimated BEC proportions also closely aligned with those reported in a previous study that used cytology to count cell types in saliva samples (58% BEC) (Wong et al., 2022). In the same study, when comparing between cell-counted proportions and DNAm deconvolution-based proportions with the adult reference panel, the latter method underestimated BEC proportions in oral samples as compared to the former method (Wong et al., 2022). Nonetheless, since the discrepancies between the two methods were not as large in oral samples collected *without* Oragene devices, the reported underestimation may be driven by the saliva samples collected with the devices. The chemical process involved in this device have been speculated to lead to the selective extraction of immune cell DNA, resulting in lower BEC proportions estimated through DNAm deconvolution (Middleton et al., 2022). The wider range of estimated BEC proportions with the child reference compared to the adult reference is likely reflective of the natural biological variation of pediatric saliva as indicated by a highly variable CT proportion observed with actual cell count data (Middleton et al., 2022; Theda et al., 2018). As further evidence of the pediatric reference being more appropriate to pediatric saliva samples, it had a lower CETYGO score in contrast to the adult reference, indicating the cellular deconvolution by the child reference was more accurate (Vellame et al., 2022). Our findings demonstrated discrepancies between estimated salivary CT proportions based on the available tissue-appropriate reference panels and indicated that the child saliva reference panel was indeed most appropriate for CT estimations in our child salivary samples. This supports the broader notion that even when other available CT reference panels can also provide reasonable estimates, the CT panel including samples from the same tissue and developmental stage as the studied

population will be more accurate, likely due to important biological factors such as cellular compositions and age effect on cell type identity DNAm markers.

Cell-type estimation discrepancy differentially affected DNA methylation associations with biological and social variables

Our study demonstrated that these discrepancies in CT estimation could impact the outcome of statistical models which are investigating high dimensional DNAm associations differentially across variables that span across the biological and social spectrum. Specifically, we found that adjusting for CT estimated by the pediatric reference panel produced a more well-behaved p-value histogram, and allowed for detection of a larger number of significant associations for SES as compared to the adult reference panel. However, the DNAm associations with biological sex and cotinine did not differ across reference panels. Therefore, not all variables were equally affected by the estimated CT discrepancies across references. Although the current analyses could not empirically differentiate true positive from false positive findings, it is tempting to speculate that adjusting for potentially more appropriate CT estimations generated in data of a similar tissue and age range was able to better capture variability of DNAm associated with CT and increased the power to detect significant associations with social variables such as SES (Jaffe & Irizarry, 2014). Yet, it is unclear why such impact was only observed with the most biologically distal and socially proximal variable in our study. Future investigations on other social variables may be needed to scrutinize the effect of the CT estimation discrepancy. Nonetheless, our findings add to the growing body of literature creating awareness around the considerations of CT proportions in DNAm analyses (Jones et al., 2018; Turinsky et al., 2019).

Stratified samples by epithelial cell proportions allowed detection of tissue-specific effects

Although adjusting for CT in DNAm analyses may be sufficient for tissues with less CT heterogeneity (e.g., buccal with 70-100% BEC) (Zheng et al., 2018), saliva samples are heterogeneous and may require further considerations since exposures and phenotypes could be differentially associated with DNAm across cell populations. Explorations of more homogenized cell populations could assist in our understanding of which CTs are most affected by exposures and phenotypes. Yet, these effects will likely be overlooked when the full samples of an extremely heterogenous tissue are analyzed together, even when CT proportion was adjusted in the model. First, DNAm associations across CTs could potentially cancel out each other if they are of opposing directionality. This issue may be more influential in tissues with high interindividual CT heterogeneity like saliva. For example, the positive DNAm associations in immune cells that are more abundant in half of the samples may be offset by the negative effect sizes found for BECs that are most abundant in another half of the samples. Second, if the variable of interest is only significantly associated with DNAm of a certain CT that is less abundant in most samples, these effects may also be left undetected. In these cases, analyzing samples with high CT-heterogeneity together may not be the most optimal approach. Therefore, we took one step further beyond only adjusting for the estimated CT proportions in the full sample and explored the effect of creating more CT homogenous samples via stratified analyses by child reference-based estimated BEC proportion in addition to adjustment of CT proportions.

With stratification, two major changes to the data, sample size and CT variability of the sample, should be considered when interpreting the findings. First, the stratified sample sizes were smaller resulting in decreased power to detect effects. Second, there was substantially less CT heterogeneity among the stratified samples as compared to the full sample, and the most abundant CT in each stratified sample differed: 1) the low BEC subset that was primarily

immune cells), 2) mid BEC that retained the most CT heterogeneity and contained a mix of BECs and immune cells, and 3) high BEC that was primarily BECs. Overall, our results showed that stratifying samples for EWAS had differential effects across variables of interest, such that a larger number of significant DNAm associations were detected with both biological sex and cotinine, but not with SES. In the following discussion, we drew on past studies across tissues, including adult blood, saliva, and buccal samples, which have similar CT compositions as our three subsets, respectively. However, it is important to acknowledge that even though the low BEC subsets have higher immune cell proportions, they are not a direct comparison of blood due to the presence of BECs albeit small proportion, the presence of oral-specific immune cells such as oral neutrophils, the lower proportion of other immune cells such as lymphocytes and monocytes, and the different tissue-environments these cells reside in (Landzberg et al., 2015; Rijkschroeff et al., 2018).

Sex differences in DNAm levels have been widely studied in different tissues, including blood, saliva, and buccal swabs (Grant et al., 2022; Moore et al., 2020; Protti et al., 2023; Reiner et al., 2023). In our current saliva samples, we found a larger number of sex-associated DNAm sites in both high and low BEC saliva stratified samples than the mid BEC subsamples and the full sample. Despite the smaller sample size in the high and low BEC stratified samples as compared to the full sample, the reduction in CT heterogeneity in these subsets likely increased the power to detect significant DNAm associations with biological sex, potentially by decreasing noise and/or increasing observable effect size. Furthermore, stratification allowed for detection of tissue-specific sex associations. First, we found that a subset of DNAm sites showed diverging effects across stratified subsets supporting the idea that some tissue-specific effect may be cancelled out due to opposite directionality across cell compositions, and as such could indicate

larger observable effect sizes when only considering a stratified sample composed largely of one primary CT. In addition, the high BEC subset had the largest number of significant sex associations, which converged with a previous study reporting a greater number of sex-associated co-methylated regions (CMRs) in buccal samples than other tissues (Gatev et al., 2021). Furthermore, we found that the overlap of sex-associated sites across the three subsets was significantly lower than would be expected by chance based on random group assignment. These consistent DNAm sites overlapped with all five CMRs that were previously reported to be associated with sex across multiple tissues, including buccal and blood tissues (Gatev et al., 2021). The low overlap across subsets, however, prompted the question of whether these subsets produced differential associations that were comparable to tissue-specific associations. This is an important question since the discordance in DNAm associations across tissue have been well-documented (Jiang et al., 2015), and caution in the interpretation of salivary DNAm associations is needed if the high interindividual CT differences in saliva samples has created as much discordance as that across tissues. Nevertheless, we found that the differential associations with sex across the three saliva subsets were not as prominent as across matched blood and buccal samples (11.9% vs. 5.0% of overlapping significant associations, respectively).

We also investigated the DNAm associations of salivary cotinine levels, which may reflect both biological and social impact of second-hand smoking in children and has been highly replicated across tissues in the adult DNAm literature. We identified three significant sites, out of which two (cg05549970, cg14588422 annotated to the *C4orf50* and *PARD3* genes, respectively) have been reported to be associated with first-hand smoking (Sikdar et al., 2019; Teschendorff et al., 2015). After stratification by BEC proportions, we identified significant DNAm associations with children salivary cotinine only in the low BEC subsample. The number of significant

cotinine-DNAm associations identified in the low BEC stratified sample, which had a smaller sample size yet a more homogeneous and immune-focused CT composition, was higher than that identified in the full sample. This finding is in line with past studies that identified tissue-specific DNAm associations with smoking, with blood showing stronger effects than saliva (Dawes et al., 2021; Philibert et al., 2020).

Lastly, we interrogated the effect of stratification on DNAm associations of SES, a commonly investigated social variable in EWAS. In the full sample analysis, we identified 95 significant DNAm associations and the effect sizes of significant SES-associations were comparable with cotinine yet smaller than biological sex. However, after stratification, we did not identify any significant SES-associations. One possible explanation is that with a smaller effect size than biological sex, the reduction of sample size may have decreased the power to detect the subtle effect. Furthermore, unlike cotinine, SES may not have a strong immune CT effect in saliva samples, or required further CT specificity than could be achieved by stratification in this, and hence did not benefit as much from stratification by CT. Yet, we observed a large number of diverging SES-associated DNAm sites across high and low BEC subsets, suggesting potential tissue specific effects. Past studies have also found differential associations across tissues with sociodemographic variables, for example one study comparing DNAm associations in buccal and blood samples have indicated that blood DNAm has stronger association than buccal DNAm (Jiang et al., 2015). However, in our salivary samples, the low BEC subset, which has the highest immune cell proportion, had the smallest sample size (n =123) and therefore, the least power to detect significant SES-associations if they were present. As such, it is unclear whether the lack of significant SES-associations in the low BEC subset was due to the small sample size or the inherent difference between blood and saliva with high

immune cell proportions given the differences in oral immune cells, which cannot be tested with our current sample but should be investigated in future studies.

Limitations & Recommendations

The findings this study should be interpreted while acknowledging its limitations. First, our study did not include cell count data to compare the accuracy of CT predictions across reference panels. Therefore, our investigation focused on the impact of differential CT predictions on downstream analyses. Second, our data only included saliva that was collected as whole saliva. Some collection methods, such as the Oragene Kit, can potentially yield different CT compositions and level of CT heterogeneity due to the chemical processes involved (e.g., 69-100% immune cells with Oragene Kit) (Middleton et al., 2022; Raffington et al., 2023; Reiner et al., 2023). Therefore, it is crucial to first assess the degree of cell type heterogeneity in the given saliva samples; with lower levels of CT heterogeneity, stratification of samples may not be necessary. Third, some of our samples had salivary cotinine concentrations that were too low to be reliably measured by the assay, which led to missingness in our data. This was not unexpected given the young age of our sample. Importantly, our main findings were supported by sensitivity analyses that used imputed cotinine concentration for samples with censored data, indicating the robustness of our findings. Given that the main objective of the current paper was to examine how different methods might affect DNAm analyses using cotinine concentrations as an example, we did not intend to draw conclusions from the significant DNAm sites identified in this study. Future studies should consider using other censored data approaches or categorizing cotinine concentrations to index different tobacco-smoke exposure levels alongside with our recommendations to identify DNAm biomarkers. Lastly, our study employed stringent QC to ensure high-quality DNAm samples for our investigation, which has removed a substantial

portion of samples. However, the QC process is cohort-dependent and can be influenced by many factors, including the extent to which child participants were able to follow instructions during sample collection and their oral health conditions, such as tooth decay, tooth loss, and cold sores, which are out of the scope of the current study (Nemoda, 2020; Padilla et al., 2020). Despite these challenges, saliva still represents a wide range of strengths in epidemiological research including its non-invasive sampling methods and opportunities for interdisciplinary research ranging from stress hormones to immune markers, from oral analytes to microbiome, in addition to genetics and epigenetic research (Hamilton et al., 2022).

To inform future epigenetic research using pediatric saliva, we summarized recommendations on analytical approaches to handle the high CT heterogeneity of pediatric saliva in DNAm analyses (see **Table 3**). In conclusion, CT references built on a population and tissue matched with the targeted samples produced more accurate estimations and the discrepancies in CT estimations had differential effects on DNAm associations with commonly investigated variables in the field. Further, stratification of samples by CT proportions to obtain DNAm data with more homogenous CT composition can help identify tissue-specific effects and allow for more powerful analyses for suitable variables of interest and sample characteristics. The current study contributes to future pediatric DNAm research by offering practical guidance to leverage this complex yet highly accessible biological sample in children.

Table 3. Recommendations for analyses with pediatric saliva DNA methylation and corresponding pros and cons

Conditions	Recommendations	Pros	Cons
Examining DNAm associations with social variables	Use the pediatric saliva reference panel for cell type deconvolution	More power to detect significant associations	Not as comparable with past saliva DNAm analyses that used the adult references
Detecting tissue-specific effects	Stratify your sample by BEC proportion	Help identify what CT or tissue may be more sensitive to the effect of the variable of interest	Less appropriate for smaller initial sample size and/or expected effect sizes of the variable of interest
Identifying generalizable saliva biomarkers	Keep full sample for analyses	More power and detection of effects that are similar across CT	High CT heterogeneity may impede interpretation of DNAm findings

*Note***.** BEC = buccal epithelial cell; CT = cell type; DNAm = DNA methylation

References

Bauer, M., Fink, B., Thürmann, L., Eszlinger, M., Herberth, G., & Lehmann, I. (2016). Tobacco smoking differently influences cell types of the innate and adaptive immune system—Indications from CpG site methylation. *Clinical Epigenetics*, *8*(1), 83. https://doi.org/10.1186/s13148-016-0249-7

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Chan, M. H., Merrill, S. M., Fatima, F., MacIsaac, J. L., Obradović, J., Boyce, W. T., & Kobor, M. S. (2023). Cross-Tissue Specificity of Pediatric DNA Methylation Associated with Cumulative Family Adversity. *bioRxiv*, 2023.10.04.559423. https://doi.org/10.1101/2023.10.04.559423

Dawes, K., Andersen, A., Reimer, R., Mills, J. A., Hoffman, E., Long, J. D., Miller, S., & Philibert, R. (2021). The relationship of smoking to cg05575921 methylation in blood and saliva DNA samples from several studies. *Scientific Reports*, *11*(1), 21627. https://doi.org/10.1038/s41598-021-01088-7

Gatev, E., Inkster, A. M., Negri, G. L., Konwar, C., Lussier, A. A., Skakkebaek, A., Sokolowski, M. B., Gravholt, C. H., Dunn, E. C., Kobor, M. S., & Aristizabal, M. J. (2021). Autosomal sex-associated co-methylated regions predict biological sex from DNA methylation. *Nucleic Acids Research*, *49*(16), 9097–9116. https://doi.org/10.1093/nar/gkab682

Govender, P., Ghai, M., & Okpeku, M. (2022). Sex-specific DNA methylation: Impact on human health and development. *Molecular Genetics and Genomics*, *297*(6), 1451–1466. https://doi.org/10.1007/s00438-022-01935-w

Granger, D. A., Fortunato, C. K., Beltzer, E. K., Virag, M., Bright, M. A., & Out, D. (2012). Focus on Methodology: Salivary bioscience and research on adolescence: An integrated perspective. *Journal of Adolescence*, *35*(4), 1081–1095. https://doi.org/10.1016/j.adolescence.2012.01.005

Grant, O. A., Wang, Y., Kumari, M., Zabet, N. R., & Schalkwyk, L. (2022). Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina EPIC array. *Clinical Epigenetics*, *14*(1), 62. https://doi.org/10.1186/s13148-022-01279-7

Hamilton, K. R., Granger, D. A., & Taylor, M. K. (2022). Science of interdisciplinary salivary bioscience: History and future directions. *Biomarkers in Medicine*, *16*(14), 1077–1087. https://doi.org/10.2217/bmm-2022-0452

Islam, S. A., Goodman, S. J., MacIsaac, J. L., Obradović, J., Barr, R. G., Boyce, W. T., & Kobor, M. S. (2019). Integration of DNA methylation patterns and genetic variation in human pediatric tissues help inform EWAS design and interpretation. *Epigenetics and Chromatin*, *12*(1), 1–18. https://doi.org/10.1186/s13072-018-0245-6

Jaffe, A. E., & Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, *15*(2), R31. https://doi.org/10.1186/gb-2014-15-2-r31

Jenks, G. F. (1977). Optimal data classification for choropleth maps. *Department of Geographiy, University of Kansas Occasional Paper*.

Jiang, R., Jones, M. J., Chen, E., Neumann, S. M., Fraser, H. B., Miller, G. E., & Kobor, M. S. (2015). Discordance of DNA methylation variance between two accessible human tissues. *Scientific Reports*, *5*, 8257. https://doi.org/10.1038/srep08257

Jones, M. J., Moore, S. R., & Kobor, M. S. (2018). Principles and challenges of applying epigenetic epidemiology to psychology. *Annual Review of Psychology*, *69*, 459–485. https://doi.org/10.1146/annurev-psych-122414-033653

Lam, L. L., Emberly, E., Fraser, H. B., Neumann, S. M., Chen, E., Miller, G. E., & Kobor, M. S. (2012). Factors underlying variable DNA methylation in a human community cohort. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(SUPPL.2), 17253–17260. https://doi.org/10.1073/pnas.1121249109

Landzberg, M., Doering, H., Aboodi, G. M., Tenenbaum, H. C., & Glogauer, M. (2015). Quantifying oral inflammatory load: Oral neutrophil counts in periodontal health and disease. *Journal of Periodontal Research*, *50*(3), 330–336. https://doi.org/10.1111/jre.12211

Langie, S. A. S., Moisse, M., Declerck, K., Koppen, G., Godderis, L., Vanden Berghe, W., Drury, S., & De Boever, P. (2017). Salivary DNA Methylation Profiling: Aspects to Consider for Biomarker Identification. *Basic & Clinical Pharmacology & Toxicology*, *121*(S3), 93–101. https://doi.org/10.1111/bcpt.12721

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, *36*(2), 451–461.

McDade, T. W., Ryan, C., Jones, M. J., MacIsaac, J. L., Morin, A. M., Meyer, J. M., Borja, J. B., Miller, G. E., Kobor, M. S., & Kuzawa, C. W. (2017). Social and physical environments early in development predict DNA methylation of inflammatory genes in young

adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(29), 7611–7616. https://doi.org/10.1073/pnas.1620661114

McEwen, L. M., Jones, M. J., Lin, D. T. S., Edgar, R. D., Husquin, L. T., MacIsaac, J. L., Ramadori, K. E., Morin, A. M., Rider, C. F., & Carlsten, C. (2018). Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. *Clinical Epigenetics*, *10*(1), 1–9.

Merrill, S. M., Konwar, C., Fatima, F., MacIsaac, J. L., Letourneau, N., Giesbrecht, G. F., Dewey, D., England-Mason, G., Lewis, C. R., Wang, D., Ailing, T., Meaney, M. J., Gonzalez, A., Bush, N. R., Stewart, S. E., & Kobor, M. S. (2024). *Pediatric Buccal Epithelial Cell Proportions Decrease Reliably With Age: Considerations for PedBE in Pediatric Research*. https://doi.org/10.21203/rs.3.rs-4219789/v1

Middleton, L. Y. M., Dou, J., Fisher, J., Heiss, J. A., Nguyen, V. K., Just, A. C., Faul, J., Ware, E. B., Mitchell, C., Colacino, J. A., & M. Bakulski, K. (2022). Saliva cell type DNA methylation reference panel for epidemiological studies in children. *Epigenetics*, *17*(2), 161–177. https://doi.org/10.1080/15592294.2021.1890874

Moore, S. R., Humphreys, K. L., Colich, N. L., Davis, E. G., Lin, D. T. S., MacIsaac, J. L., Kobor, M. S., & Gotlib, I. H. (2020). Distinctions between sex and time in patterns of DNA methylation across puberty. *BMC Genomics*, *21*(1), 1–16. https://doi.org/10.1186/s12864-020-06789-3

Nemoda, Z. (2020). The Use of Saliva for Genetic and Epigenetic Research. In D. A. Granger & M. K. Taylor (Eds.), *Salivary Bioscience: Foundations of Interdisciplinary Saliva Research and Applications* (pp. 115–138). Springer International Publishing. https://doi.org/10.1007/978-3-030-35784-9_6

Padilla, G. A., Calvi, J. L., Taylor, M. K., & Granger, D. A. (2020). Saliva Collection, Handling, Transport, and Storage: Special Considerations and Best Practices for Interdisciplinary Salivary Bioscience Research. In D. A. Granger & M. K. Taylor (Eds.), *Salivary Bioscience: Foundations of Interdisciplinary Saliva Research and Applications* (pp. 21–47). Springer International Publishing. https://doi.org/10.1007/978-3-030-35784-9_3

Philibert, R., Dogan, M., Beach, S. R. H., Mills, J. A., & Long, J. D. (2020). AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *183*(1), 51–60. https://doi.org/10.1002/ajmg.b.32760

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., & Taylor, K. D. (2008). Long-range LD can confound genome scans in admixed populations. *The American Journal of Human Genetics*, *83*(1), 132–135.

Protti, G., Rubbi, L., Gören, T., Sabirli, R., Civlan, S., Kurt, Ö., Türkçüer, İ., Köseler, A., & Pellegrini, M. (2023). The methylome of buccal epithelial cells is influenced by age, sex, and physiological properties. *Physiological Genomics*, *55*(12), 618–633. https://doi.org/10.1152/physiolgenomics.00063.2023

Qi, L., & Teschendorff, A. E. (2022). Cell-type heterogeneity: Why we should adjust for it in epigenome and biomarker studies. *Clinical Epigenetics*, *14*(1), 31. https://doi.org/10.1186/s13148-022-01253-3

Raffington, L., Schneper, L., Mallard, T., Fisher, J., Vinnik, L., Hollis-Hansen, K., Notterman, D. A., Tucker-Drob, E. M., Mitchell, C., & Harden, K. P. (2023). Salivary Epigenetic

Measures of Body Mass Index and Social Determinants of Health Across Childhood and Adolescence. *JAMA Pediatrics*. https://doi.org/10.1001/jamapediatrics.2023.3017

Reiner, A., Bakulski, K. M., Fisher, J. D., Dou, J. F., Schneper, L., Mitchell, C., Notterman, D. A., Zawistowski, M., & Ware, E. B. (2023). Sex-specific DNA methylation in saliva from the multi-ethnic Future of Families and Child Wellbeing Study. *Epigenetics*, *18*(1), 2222244. https://doi.org/10.1080/15592294.2023.2222244

Rijkschroeff, P., Loos, B. G., & Nicu, E. A. (2018). Oral Polymorphonuclear Neutrophil Contributes to Oral Health. *Current Oral Health Reports*, *5*(4), 211–220. https://doi.org/10.1007/s40496-018-0199-6

Sikdar, S., Joehanes, R., Joubert, B. R., Xu, C.-J., Vives-Usano, M., Rezwan, F. I., Felix, J. F., Ward, J. M., Guan, W., Richmond, R. C., Brody, J. A., Küpers, L. K., Baïz, N., Håberg, S. E., Smith, J. A., Reese, S. E., Aslibekyan, S., Hoyo, C., Dhingra, R., … London, S. J. (2019). Comparison of Smoking-Related DNA Methylation Between Newborns from Prenatal Exposure and Adults from Personal Smoking. *Epigenomics*, *11*(13), 1487–1500. https://doi.org/10.2217/epi-2019-0066

Simmons, T. R., Flax, J. F., Azaro, M. A., Hayter, J. E., Justice, L. M., Petrill, S. A., Bassett, A. S., Tallal, P., Brzustowicz, L. M., & Bartlett, C. W. (2011). Increasing Genotype-Phenotype Model Determinism: Application to Bivariate Reading/Language Traits and Epistatic Interactions in Language-Impaired Families. *Human Heredity*, *70*(4), 232–244. https://doi.org/10.1159/000320367

Slavish, D. C., Graham-Engeland, J. E., Smyth, J. M., & Engeland, C. G. (2015). Salivary markers of inflammation in response to acute stress. *Brain, Behavior, and Immunity*, *44*, 253–269. https://doi.org/10.1016/j.bbi.2014.08.008

Teschendorff, A. E., Yang, Z., Wong, A., Pipinikas, C. P., Jiao, Y., Jones, A., Anjum, S., Hardy, R., Salvesen, H. B., Thirlwell, C., Janes, S. M., Kuh, D., & Widschwendter, M. (2015). Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. *JAMA Oncology*, *1*(4), 476–485. https://doi.org/10.1001/jamaoncol.2015.1053

Theda, C., Hwang, S. H., Czajko, A., Loke, Y. J., Leong, P., & Craig, J. M. (2018). Quantitation of the cellular content of saliva and buccal swab samples. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-25311-0

Tsukamoto, K., & Machida, K. (2014). Effects of psychological stress on neutrophil phagocytosis and bactericidal activity in humans—A meta-analysis. *International Journal of Psychophysiology*, *91*(2), 67–72. https://doi.org/10.1016/j.ijpsycho.2013.12.001

Turinsky, A. L., Butcher, D. T., Choufani, S., Weksberg, R., & Brudno, M. (2019). Don't brush off buccal data heterogeneity. *Epigenetics*, *14*(2), 109–117. https://doi.org/10.1080/15592294.2019.1581592

van Iterson, M., van Zwet, E. W., the BIOS Consortium, & Heijmans, B. T. (2017). Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology*, *18*(1), 19. https://doi.org/10.1186/s13059-016-1131-9

Vellame, D. S., Shireby, G., MacCalman, A., Dempster, E. L., Burrage, J., Gorrie-Stone, T., Schalkwyk, L. S., Mill, J., & Hannon, E. (2022). *Uncertainty quantification of reference based cellular deconvolution algorithms* (p. 2022.06.15.496235). bioRxiv. https://doi.org/10.1101/2022.06.15.496235

Vernon-Feagans, L. (2014). THE FAMILY LIFE PROJECT: AN EPIDEMIOLOGICAL AND DEVELOPMENTAL STUDY OF YOUNG CHILDREN LIVING IN POOR RURAL COMMUNITIES. *Monographs of the Society for Research in Child Development*.

Wong, Y. T., Tayeb, M. A., Stone, T. C., Lovat, L. B., Teschendorff, A. E., Iwasiow, R., & Craig, J. M. (2022). A comparison of epithelial cell content of oral samples estimated using cytology and DNA methylation. *Epigenetics*, *17*(3), 327–334. https://doi.org/10.1080/15592294.2021.1950977

Zheng, S. C., Webster, A. P., Dong, D., Feber, A., Graham, D. G., Sullivan, R., Jevons, S., Lovat, L. B., Beck, S., Widschwendter, M., & Teschendorff, A. E. (2018). A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. *Epigenomics*, *10*(7), 925–940. https://doi.org/10.2217/epi-2018-0037