Navarro-Delgado *et al. Genome Biology* (2025) 26:421
https://doi.org/10.1186/s13059-025-03864-4

Genome Biology

SOFTWARE Open Access

RAMEN: Dissecting individual, additive and interactive gene-environment contributions to DNA methylome variability in cord blood

Erick I. Navarro-Delgado[1,2,3], Darina Czamara[4], Karlie Edwards[1,2,3], Maggie P. Fu[1,2,3], Sarah M. Merrill[2,3,5], Chaini Konwar[1,2,3], Julie L. MacIsaac[1,2,3], David T.S. Lin[2,3], Piush Mandhane[6,7], Elinor Simons[8], Padmaja Subbarao[9], Theo J. Moraes[9], Jari Lahti[10,11], Gregory E. Miller[12], Elisabeth B. Binder[4], Katri Räikkönen[13,14], Stuart E. Turvey[2,15], Keegan Korthauer[2,3,16*] and Michael S. Kobor[1,2,3,17*]

*Correspondence:
Keegan Korthauer
Keegan.Korthauer@bcchr.ca
Michael S. Kobor
michael.kobor@ubc.ca

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

Abstract

Genetic variation and environmental exposures are amongst the main factors associated with inter-individual DNA methylation variability. However, the prevalence and genomic context of individual, additive, and interactive gene-environment effects remains unclear. We present RAMEN, an R package that dissects genome-exposome contributions to microarray Variably Methylated Loci using machine learning and statistical techniques. Analyzing cord blood samples from CHILD and PREDO (overall $n = 1662$), we identify genetic variants as key contributors to DNA methylation variability, usually in additive and interactive combinations with the environment. We provide a detailed catalogue of cord blood Variably Methylated Loci and the gene-environment contribution to their variability.

Keywords DNA methylation, DNA methylome variability, Prenatal exposome, Gene-environment interaction, Multi-omics integration, R package, Gene-environment contribution, GxE, Software

Background

The prenatal period is a time of heightened sensitivity that marks the beginning of an individual's health trajectory. Developmental programming, a concept often studied under the Developmental Origin of Health and Disease (DOHaD) framework, describes how exposures and experiences during the prenatal period or shortly after can lead to long-lasting biological effects, potentially impacting future health outcomes [1]. Initially developed in the context of fetal malnutrition during gestation associated with lower birth weight and a future increased risk for cardiometabolic diseases [1, 2], contemporary research has further identified genetic variants as important mediators of this particular pathway [3, 4]. At the molecular level, several processes may facilitate

developmental programming, including DNA methylation (DNAme). DNAme is an epigenetic mark defined by the addition of a methyl group to the DNA primarily in a cytosine-guanine (CpG) context [5, 6]. Briefly, DNAme is a plausible candidate for biological embedding of environments due to its role in regulating gene activity [7, 8] and its longitudinally-stable yet environmentally-malleable nature [9–12]. Furthermore, DNAme changes have been associated with a wide variety of environmental exposures [6, 7, 12] and health outcomes [6–8]. However, the factors that contribute to shaping DNAme variation across the human genome are just beginning to be understood. Characterizing these factors is crucial to better understand disease-associated DNAme patterns, which might provide innovative avenues to refine precision medicine approaches.

Inter-individual DNAme differences in human populations are primarily, but not exclusively, associated with two factors: genetic variants [13] and environmental exposures [14]. Single Nucleotide Polymorphisms (SNPs) associated with DNAme levels, termed methylation quantitative trait loci (mQTLs), tend to be mostly stable across the lifespan [15] and associated with a significant proportion of CpGs in the human methylome (in the range of 1 to 48%) predominantly when they are in *cis* (reviewed in Villicaña & Bell [13]). Besides genetic contributors, multiple associations between environmental exposures and DNAme changes have been reported. Specifically in early life, DNAme signatures have been associated with biological, nutritional, socioeconomical, and pollutant exposures [14, 16, 17].

In addition to the individual associations of genetic variants and the environment, environmental exposures could have differential influence on DNAme depending on the underlying genetic makeup. Recent studies in blood have reported gene-environment interactions (GxE) associated with DNAme levels in psychological and environmental exposures [18–22]. These interactions add a layer of complexity in the study of DNAme, since genome-wide investigation requires a larger sample size and more computationally expensive statistical methods [23]. However, characterizing these interactions offers a highly promising avenue to understand the etiology of complex human traits [24]. Overall, there is robust evidence pointing at both genetic variation and environmental exposures as sources of DNAme variation.

To understand the extent to which genetic and environmental differences contribute to DNA methylation variation, several earlier studies have leveraged the power of classical twin models. By contrasting the DNAme correlation between pairs of monozygotic and dizygotic twins, these models are a valuable tool for estimating genetic contributions, and informing complementary environmental contributions without profiling the genome or exposome [25, 26]. These studies have found additive genetic effects to explain on average from 5% to 19% of DNAme variance across the genome [27–31]. However, a limitation of twin models is that they assume residual effects (i.e., not explained by genetic differences) to be unique environmental effects. As residual effects are a mix of stochastic variation, measurement error, and environmental effects, the environment contribution estimates in twin models are indirect and thus require careful interpretation [26, 32]. Furthermore, these classical approaches can underestimate shared environmental effects, and do not model gene-environment interactions [25, 26]. Therefore, population-based studies that integrate genomic, exposomic, and epigenomic data can offer valuable complementary insights by directly estimating the individual,

joint and interactive contribution of genetic and environmental factors to DNA methylation variation, and by providing specific candidate contributor variables.

Recent integrative studies in perinatal population cohorts have pioneered the direct interrogation of genetic and environmental contributions, as well as their interaction, to DNAme variation [33–35]. These studies observed that the DNAme levels of Variably Methylated Regions (VMRs; genomic regions with high levels of inter-individual DNAme variation) are better explained by additive or interactive mathematical models that include genetic and prenatal environmental variables as opposed to only one of these factors. While these studies [33–35] laid the groundwork for modelling the methylome-wide contributors to DNAme variability, recent advances in DNAme analysis methods and study design now offer the opportunity to refine these estimates. For instance, due to the lack of available tools at the time, some early analyses did not explicitly adjust the models for cell type proportions [34, 35] or population stratification [34]. Additionally, the existing studies measured a limited number of prenatal exposome variables ranging from 8 to 19 [33–35], which could result in an underestimation of the environment's contribution to DNAme variability. Furthermore, the framework for examining the genome-exposome contribution to methylome variability could potentially improve its estimations through statistical and analytical updates, such as taking into account the DNAme microarray design to best capture methylome variability patterns, the selection of metrics to model DNAme variation, and controlling for the spurious models expected by chance.

In this study, our main objective was to characterize DNAme variability patterns in cord blood samples from general populations (i.e., not enriched for any specific disease) and identify whether they were best explained by differences in genetic variants (G), environmental exposures (E), their independent additive effects (G + E), or their interaction (GxE) through a reproducible bioinformatic framework that addresses the limitations of previous works mentioned in the previous paragraph. Building upon previous methodological ideas [33–35], we integrated genome, exposome, and methylome data from 699 cord blood samples of the Canadian Healthy Infant Longitudinal Development (CHILD) study [36–38]. This cohort possesses a comprehensive characterization of prenatal exposures, capturing 94 curated prenatal exposome variables including maternal psychosocial factors, built environment, maternal nutrition, and maternal health measurements. To allow for a reproducible and structured analysis of such data, we developed the R package RAMEN (*Regional Association of Methylome variability with the Exposome and geNome*) a Findable, Accessible, Interoperable, and Reusable (FAIR) [39] tool to conduct a genome-exposome contribution to DNA methylome variability analysis. Specifically, in this paper we aimed to: (1) identify genomic regions with high DNAme variability in cord blood; (2) identify the model (G, E, G + E, GxE or null) that best explained the methylome variability per region; (3) characterize the regions explained by each model in CHILD; and (4) apply our methodology and compare results with those previously published on the Pre-eclampsia and Intrauterine Growth Restrictions (PREDO) cohort [33]. Using a FAIR methodology, our comprehensive characterization on how and where in the genome genetic variants and environmental exposures associate with DNAme variability sheds light on the complex interplay between these factors during a highly sensitive developmental period.

Results

RAMEN provided a reproducible framework to estimate the gene-environment contribution to DNA methylation across genomic regions

We developed *RAMEN (Regional Association of DNA Methylome with the Exposome and geNome)*, an R package that provides a FAIR and user-friendly tool to dissect the factors associated with genome-wide DNA methylation variability patterns. Building upon previous methodological ideas [33–35], *RAMEN* integrates matched microarray DNA methylome data, genomic variants, and exposome data, and analyzes them in a reproducible framework that can be conceptually divided into three main steps.

The first step identifies the top 10% most variable DNAme probes in a given data set based on their variance. In the second step, these Highly Variable Probes (HVPs; top 10% probes with the highest variance) are grouped into Variably Methylated Loci (VML). To best capture methylome variability patterns in microarrays, we analyzed two types of VML: Variably Methylated Regions (VMRs) and sparse Variably Methylated Probes (sVMPs). In one hand, we defined VMRs as two or more proximal HVPs (< 1 kb apart) with correlated DNAme levels (r >0.15). Modelling DNAme variability through regions rather than individual CpGs provides several methodological advantages in association studies, since CpGs display a significant correlation for co-methylation when they are close (≤ 1 kilobase) [40, 41]. Some of these advantages include increasing statistical power by testing redundant probes only once, reducing false-positives driven by one problematic probe in a region, and improving comparability between studies that analyze the same genomic region but measure distinct CpGs due to microarray design differences [42]. In contrast with Differentially Methylated Regions (DMRs), a different class of regional construct used in the field, VMRs denote regions with high inter-individual variability in methylation levels within a single population, while DMRs represent regions where DNA methylation differs significantly across a variable of interest [42].

In addition to traditional VMRs, we also identified sVMPs, a second type of VML that takes into account the sparse and non-uniformly distributed coverage of CpGs in microarrays to tailor our analysis to this DNAme platform. sVMPs aimed to retain genomic regions with high DNAme variability measured by single probes, where probe grouping based on proximity and correlation is therefore not applicable. This is particularly relevant in the Illumina EPIC v1 array, where most covered regulatory regions (up to 93%) are represented by just one probe. Notably, based on empirical comparisons with whole-genome bisulfite sequencing data, these single probes are mostly representative of local regional DNAme levels due to their positioning (98.5–99.5%) [43]. By taking into account the probes positioning design during the VML identification step, we aimed to improve the capture of methylome variability patterns in microarrays.

Following the identification of VML, the individual and joint contribution of genetic and environmental variables to each locus is estimated through machine learning and statistical techniques (Fig. 1A; see Methods for further details). We implemented the RAMEN framework in 6 core functions compatible with parallel computing that identify and summarize VML in DNAme microarrays, select G and E variables potentially associated with DNAme, identify the best model per VML (G, E, G + E or GxE), decompose DNAme variance, and simulate a null distribution to identify VML where gene-environment terms did not improve the baseline model (i.e., the model with covariates only) more than what we would expect by chance (Fig. 1B).

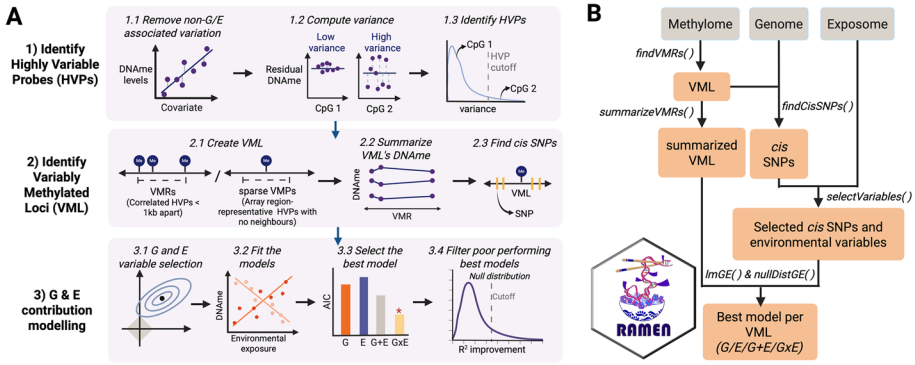
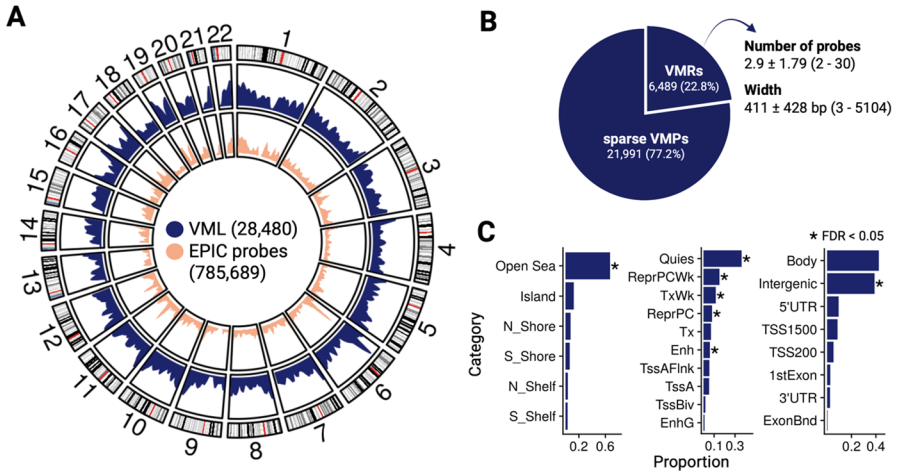


Fig. 1 RAMEN provides a FAIR framework to model gene-environment contributions to DNA methylome variability. **A** Conceptually, the first step involves identifying the 10% most variable probes in the DNA methylation microarray based on their variance (steps 1.1–1.3). Next, the HVPs are grouped into VML, which can be VMRs based on genomic proximity and DNAme correlation, or sparse VMPs based on the absence of proximal probes in the array (step 2.1). The DNAme levels of each VMR are summarized per individual by taking the median DNAme level of the constituting probes (step 2.2). After identifying the SNPs in *cis* for each VML (step 2.3), a machine learning algorithm is used to screen *cis* SNPs and environmental exposures and select potentially relevant variables per VML (step 3.1). Selected variables are then used to fit G, E, G + E and GxE models (step 3.2), and the best model is chosen by comparing their Akaike Information Content metrics (step 3.3). Finally, a permutation analysis is conducted to simulate a null distribution that is used to filter out winning models that do not perform better than expected by chance (step 3.4). **B** The RAMEN package implements the methodology in six functions: *findVMRs* (steps 1–2.1.1), *summarizeVMRs* (step 2.2), *findCisSNPs* (step 2.3), *selectVariables* (step 3.1), *lmGE* (steps 3.2–3.3), and *nullDistGE* (step 3.4). See Methods for further details

Cord blood VML were enriched in quiescent chromatin, open sea, intergenic, and gene body regions

Putting RAMEN into practice, we took advantage of the CHILD cohort, which has DNAm profiles from 699 cord blood samples along with genotyping and extensive prenatal exposome characterization. After removing variation associated with concomitant variables (sex, gestational age, estimated cell type proportion, and population stratification), we identified 78,569 HVPs . These HVPs captured 70% of the 3,885 previously reported tissue- and ethnicity-independent human highly variable CpGs (hvCpGs) in the EPIC array [44], as expected based on the study's definition of hvCpGs (top 5% variable CpGs in ≥ 65% data sets [44]). Furthermore, HVPs were significantly enriched for probes with bimodal and trimodal patterns in our dataset ($p < 0.005$, OR bimodal = 112.4, OR trimodal = 148.7), despite the low number of probes in each category (bimodal = 403, trimodal = 105). Probes with bimodal and trimodal patterns can be indicative of genetic effects [45], which was consistent with the majority of them being also detected as mQTLs (bimodal = 93.3%, trimodal = 93.3%).

After identifying the HVPs, they were grouped into 28,480 VML that were distributed roughly evenly across the autosomes (Fig. 2A; additional details in Methods), and primarily composed of the sparse VMP type (Fig. 2B). VML were predominantly located in open sea, gene body and intergenic segments, and quiescent chromatin states, and, compared to the array background probe distribution, VML were significantly enriched in all those categories except for gene body (FDR < 0.05; Fig. 2C). Additionally, we found probes in VML to be significantly associated with a higher probe reliability compared to all other probes based on the metrics reported by Sugden et al. [46] (Fig. S1, Additional File 1). This observation agrees with previous studies that report an association between CpG biological variability and ICC values [47].



Notes: chrom state: only top 10 most frequent states. Enrichment with one side Fisher's test, significant = fdr < 0.05); all categories have counts, but some are small and seem like they dont

Fig. 2 Cord blood VML were mainly located in open sea, quiescent, gene body, and intergenic regions. **A** Density of VML across the genome. The background distribution of EPIC array probes is displayed in the most inner circle; circos plot was created with *circlize* [48]. **B** Composition of VML types, and average number of probes and width for VMRs. **C** Characterization of VML regarding CpG island context, chromatin state and gene annotation. The FDR corresponds to a one side Fisher's test to identify enrichments in the VML probes relative to probes in the Illumina EPIC array. For the chromatin state, only the top 10 most frequent states are shown. Definition of categories: Shore = 0.2 kb from island, Shelf = 2–4 kb from island, N = upstream(5´) of CpG island, S = downstream (3´) of CpG island; Quies - quiescent/low; ReprPCWk = weak repressed PolyComb; TxWk = weak transcription; ReprPC = repressed PolyComb; Tx = strong transcription; Enh = enhancers; TssAFlnk = flanking active transcriptional start site (TSS); TssA = active TSS; TssBiv = bivalent/poised TSS; EnhG = genic enhancers; TSS200 = 0–200 bases upstream of the TSS; TSS1500 = 200–1500 bases upstream of the TSS; 5'UTR = Within the 5' untranslated region, between the TSS and the ATG start site; Body = Between the ATG and stop codon, irrespective of the presence of introns, exons, TSS, or promoters; 3'UTR = between the stop codon and poly A signal

RAMEN reduced and balanced the number of genome and exposome variables through feature selection

Following the identification of cord blood VML, we summarized the DNAme levels of each VMR by obtaining their median methylation value (Fig. 1A, step 2.2), as VMRs contained from 2 to 30 probes. Then, we modeled the contribution of the genome and exposome to the DNAme levels of each VML in four different scenarios: DNAme levels being best explained by a genetic difference (G), an environmental exposure (E), or the combination of genetic variants and the environment in an additive (G + E) or an interactive (GxE) manner. To model the exposome contribution, we used 94 prenatal exposome variables characterized in the CHILD cohort that encompass four main fetal exposure dimensions: maternal health (48.9%), maternal nutrition (30.9%), maternal psychosocial state (10.6%), and built environment (9.6%; Additional File 2). To model the genetic contribution, we utilized imputed DNA microarray genotyping data and included the SNPs in *cis* of each VML (*cis*: < 1 Mb up or downstream; mean *cis* SNPs per VML = 980.3, *SD* = 415.3, range = [0–6172]).

In the CHILD data set, we observed a substantially higher number of G variables compared to E across VML. This could potentially bias the results towards a higher frequency of G models, as the number of VML with a G association could be higher compared to E simply by chance. In addition, the high number of variables could lead to a high computational burden associated with fitting all combinations of G, E, G + E, and GxE models, which were on the order of 10 [9]. To balance the number of genome and

exposome variables and remove uninformative variables, we implemented in RAMEN a scalable machine learning feature selection strategy based on *Least Absolute Shrinkage and Selection Operator* (LASSO) before fitting the models (for more details see *Methods*, *Genome and exposome variable selection*). After feature selection, we observed a reduction in the number of *cis* SNPs and environmental exposures considered for modelling per VML from 980 to 16 (G) and 94 to 3 (E) on average, and a reduced difference between average G and E variables per VML (initial: 886, post-selection: 13; Fig. 3A).

Despite reducing the initial difference in the number of genome and exposome variables, we still found a higher number of selected G features compared to E across the VML. To explore if this could be an artifact driven largely by genetic contributors having a higher number of variables (initial mean G = 980, initial mean E = 94), which could lead to more selected features by chance, we analyzed the number of selected G and E features in a permuted scenario. In a permuted scenario, any association between DNAme and G or E is broken through data shuffling; after the permutation, differences in the observed number of selected features between groups can be assumed to be driven by the variable imbalance, isolated from the effect of true underlying associations. We observed an overall similar distribution of the number of selected variables, with a median of 0 as expected for a permutated scenario, but a heavier tail for the genome compared to the exposome (genome: mean = 3.7 SPNs per VML, *SD* = 6.9; exposome: mean = 2.2 variables per VML, *SD* = 4.2; Fig. S2A, Additional File 1). These results showed that the initial difference in the number of variables exclusively has a minimal overall effect on the number of selected genome and exposome features. The distribution of the number of

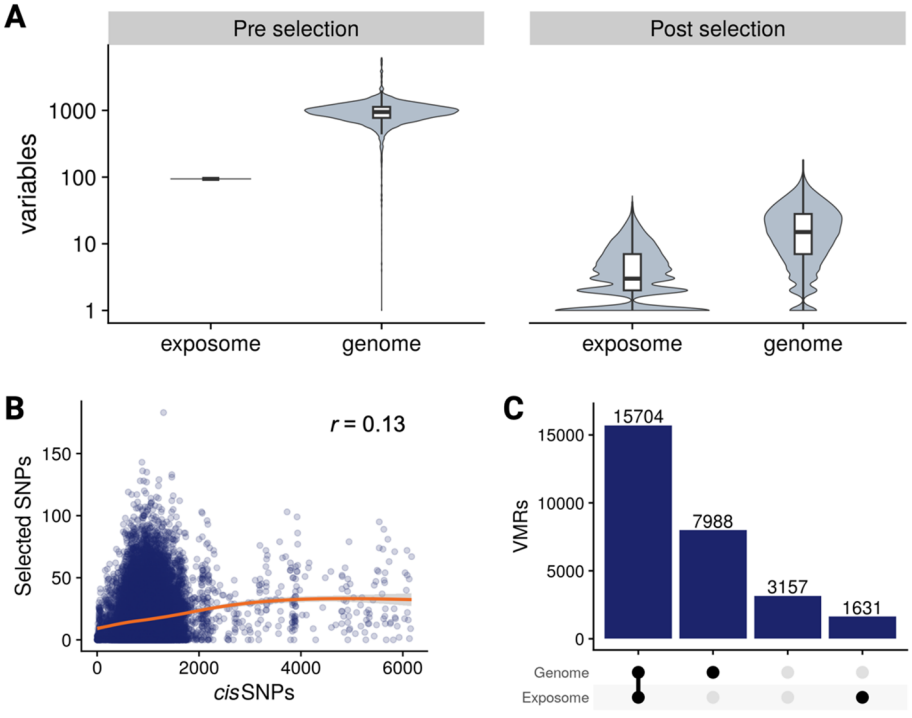


Fig. 3 Feature selection reduced and balanced the number of genome and exposome variables per VML. **A** Number of exposome and genome variables in the set of VML before (left) and after (right) the feature selection strategy based on LASSO. The initial number of variables for each VML corresponds to the SNPs in *cis* (genome) and all the 94 available exposome variables. **B** Relation of initial and selected SNPs across the VML. **C** From left to right: number of VML with at least one variable selected in both the genome and exposome, only the genome, none, or only the exposome

selected features was highly similar between G and E when we equaled the initial number of genome and exposome variables via random subsampling of G prior to the selection (Fig. S2B, Additional File 1).

When characterizing the selected features, we found a significant enrichment of selected G variables (i.e., SNPs) for *cis* mQTLs ($p < 0.05$, OR = 1.86). We found no relation between the initial and post-selection number of G variables in the VML ($r = 0.13$; Fig. 3B); this characteristic is especially important to prevent the potential bias towards G models winning in VML within densely genotyped regions at the modelling stage, and further supports the robustness of the feature selection strategy to the initial number of variables. Compared to the initial exposome data set, selected features were not enriched in any prenatal exposome dimension (post selection: maternal health = 51.6%, maternal nutrition = 27.5%, maternal psychosocial status = 9.9%, built environment = 11%). Comparing VMRs and sparse VMPs, we observed a small difference in the initial number of *cis* SNPs (VMR mean = 1023.9, sVMP mean = 967.5; $p < 0.05$). After the variable selection, the difference of SNP variables persisted (VMR *mean* = 22.1, sVMP *mean* = 14.6; $p < 0.05$). Similarly, we observed a small difference in the number of environmental variables post-selection between the VML categories (VMR *mean* = 3.5, sVMP *mean* = 2.9; $p < 0.05$; Fig. S3A, Additional File 1).

By the end of the variable selection, 15,704 (55%) VML had both E and G variables selected, 7,988 (28%) only G, 1,631 (6%) only E, and 3,157 (11%) neither E nor G (Fig. 3C). These proportions were similar between VMRs and sparse VMPs (Fig. S3B, Additional File 1). VML with no G and E selected variables occurred when no genome or exposome variables substantially improved the model performance compared to having only the concomitant variables. Because of this, the basal model (B; i.e., only concomitant variables) was assigned as the best model to VML with no G and E variables selected in our data set (11%).

Genetic variants were included in the majority of VML best models and explained the largest proportion of DNAme variance

After selecting the potentially relevant variables through the data-driven screening based on LASSO, we proceeded to compare and select the best model for each of the VML with G or E variables selected (89% of them). The evaluated models included one G and/or one E variable at a time to reduce the statistical complexity and computational time involved in assessing higher-order GxE interactions. We fitted all the possible G, E, and pairwise G + E and GxE models using the variables obtained in the feature selection step. For each VML, the model with the lowest AIC, a penalized likelihood metric, was identified as the best fit.

Following the identification of best models by AIC, we wanted to discard the ones that, despite having the lowest AIC, had a performance similar to what we would expect by chance. To do so, we implemented a permutation-based strategy. Briefly, we shuffled the data, re-ran the analysis multiple times, and computed the model R^2 improvement in G, E, G + E, and GxE winning models, compared to the baseline model (i.e., only covariates), that we would expect by chance (i.e., null distribution; more details in *Methods*). VML best models selected by AIC in CHILD for which R^2 improvement (compared to their baseline model) was below the 95th percentile of the null distribution were assigned to

the Baseline group (B; 30.1%), in addition to the VML that had no relevant G or E variables right after the feature selection step (11%; total = 41.1%).

Out of the informative models (58.9%; i.e., other than Baseline), the G + E model best explained most of the VML (27.1%), followed by G (18.3%), GxE (13.4%), and an almost negligible fraction of E (0.1%; Fig. 4A). This order was consistent between VMRs and sparse VMPs, with the difference of sparse VMPs having a higher proportion of Baseline models (46%) compared to VMRs (24%; Fig. S4A, Additional File 1). In almost all cases (>99.99%) where our selection strategy picked both G and E variables, the best model in the AIC comparison contained both factors (i.e., G + E or GxE models; Fig. S4B, Additional File 1), as opposed to only one (i.e., G or E models). This observation suggested a good performance of LASSO's variable screening property [49].

After identifying the VML with informative best models, we characterized the selected variables and VML within model groups to unpack specific contributions and understand the underlying fabric of each group. We found an enrichment of mQTLs in the set of SNPs from winning models with a genetic component (i.e., G, G + E, and GxE) compared to the set of selected *cis* SNPs ($p < 0.05$, OR = 5.2), which were already enriched for mQTLs. The exposure dimension with the highest proportion of variables selected by informative models was maternal health (49.5%), followed by maternal nutrition (28.1%), built environment (11.7%) and maternal psychosocial state (10.7%). This trend

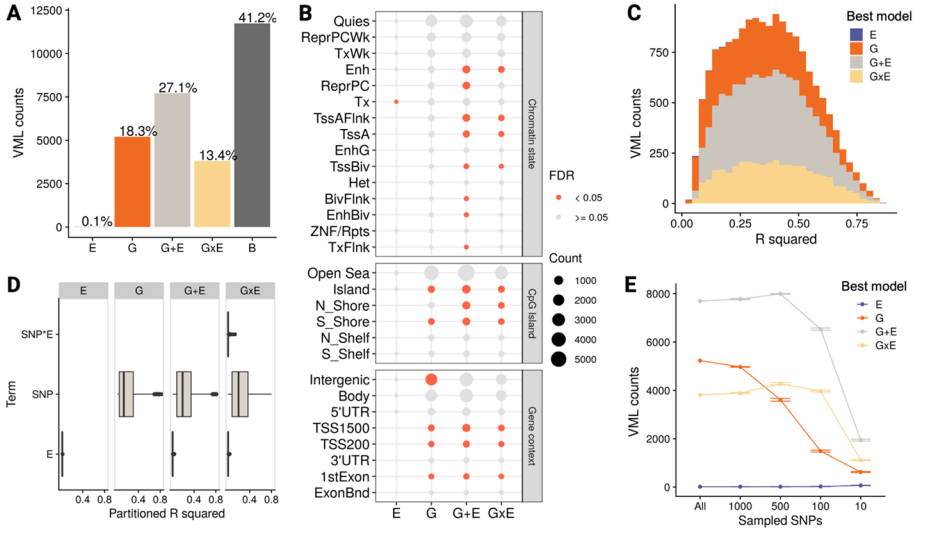


Fig. 4 Genetic contributors were included in most VML best models and explained the largest proportion of DNAme variance. **A** Best explanatory models for the set of VML in CHILD cord blood samples. *B* = basal model. **B** Characterization of model groups regarding CpG island context, chromatin state, and gene annotation. Definition of categories: Shore = 0.2 kb from island, Shelf = 2–4 kb from island, N = upstream(5′) of CpG island, S = downstream (3′) of CpG island; TssA = active transcriptional start site (TSS); TssAFlnk = flanking active TSS; TxFlnk = transcribed state at the end of the 5′ and 3′ genes; Tx = strong transcription; TxWk = weak transcription; EnhG = genic enhancers; Enh = enhancers; ZNF/Rpts = ZNF genes and repeats; Het = heterochromatin; TssBiv = bivalent/poised TSS; BivFlnk = flanking bivalent TSS/Enh; EnhBiv = bivalent enhancer; ReprPC = repressed PolyComb; ReprPCWk = weak repressed PolyComb; Quies = quiescent/low; TSS200 = 0–200 bases upstream of the TSS; TSS1500 = 200–1500 bases upstream of the TSS; 5′UTR = Within the 5′ untranslated region, between the TSS and the ATG start site; Body = Between the ATG and stop codon, irrespective of the presence of introns, exons, TSS, or promoters; 3′UTR = between the stop codon and poly A signal. **C** R^2 distribution of best explanatory models that passed the permutation analysis threshold. **D** Partial R^2 of the genetics, environmental, and interaction terms estimated with the LMG method. **E** Sensitivity of the results to the information reduction in the genomic data set, estimated through random under-sampling of SNPs (detailed information in methodology). Error bars represent *SD* ($n = 5$). The number of B models were not plotted to improve the readability of the graphic, but this can be obtained by subtracting the number of G, E, G + E, and GxE models from the number of total VML (28,480) in each under-sampled experiment

was consistent in VMRs and sparse VMPs (Fig. S5, Additional File 1), and recapitulated the pattern observed in the number of variables selected. We found distinct enriched genomic and functional categories for each model group. G winning-models were enriched in categories related to active Transcription Start Sites (TSSs), G + E in active and bivalent TSSs and enhancers, GxE in active TSSs and enhancers, and E in strong transcription despite the small number of VML (Fig. 4B).

To have a more granular understanding of the contribution of genetic and environmental factors to DNAme, we dissected the explained variance in our results. Informative models explained on average 37% of DNAme regional variance (mean $R^2 = 0.37$, $SD = 0.18$) with a similar range across model groups except for E, which tended to have smaller values (Fig. 4C). When decomposing the variance of each VML best model to estimate the contribution of each term, we found genetic variants to consistently explain a significantly higher amount of variance (mean partitioned $R^2 = 0.22$, $SD = 0.18$) compared to the environmental (mean partitioned $R^2 = 0.01$, $SD = 0.004$; $p < 0.05$) and interaction (mean partitioned $R^2 = 0.01$, $SD = 0.01$; $p < 0.05$) terms (Fig. 4D; full results in Additional File 3). This result was again consistent across VMRs and sparse VMPs (Fig. S6, Additional File 1).

An important consideration when comparing the contribution of the genome and the exposome to DNAme variability is that capturing varying environmental exposures in a population is methodologically more challenging than measuring genetic variants. To explore the sensitivity of our results to this information imbalance, we simulated scenarios with a varying degree of "captured" information of genetic variants by reducing the number of available *cis* SNPs per VML by random under-sampling. When repeating the RAMEN analysis in this context, we observed an inverse relationship between the initial number of SNPs and the number of B models. Specifically, individual models of the under-sampled factor (i.e., G models) were especially sensitive to the information loss, while joint models (G + E and GxE) remained relatively stable (Fig. 4E). Notably, the number of VML best explained only by the unchanged factor (E models) was robust to the different degrees of information in G. These results showed how datasets with a limited characterization of the genome or exposome are likely to underestimate the contribution of their respective genetic or environmental factors, and that our method prevents this under-characterization from causing an overestimation of the complementing factor's contribution.

Genome-exposome contribution trends were replicated in the PREDO cohort with a complete agreement in the involvement of genetic contributors

Next, we wanted to test whether our genome-exposome contribution results from CHILD replicated in another cohort. To do this, we used RAMEN to reanalyze data from the PREDO cohort, which was previously investigated in an integrative study interrogating the genetic and environment contributions to DNAme in cord blood in 2019 (mentioned in the introduction; referred to here as C_2019) [33]. Briefly, the PREDO data was composed of two cord blood sets that were profiled for DNAme with the Illumina 450k (PREDO I; n = 817) or EPIC array (PREDO II; n = 146). The PREDO data sets captured 8 prenatal environmental variables from the maternal psychosocial and health dimensions (2 and 6 respectively; 6/8 also measured in CHILD). Beyond an attempt at replication, using PREDO allowed us to compare RAMEN with one of the methodologies upon

Table 1 Gene-environment contribution to methylome variability in the CHILD and PREDO studies using RAMEN

Cohort	*n*	VML	E models	G models	G + E models	GxE models	B models
CHILD (EPIC)	699	28 480 *(6 489 VMRs; 21 991 sVMPs)*	18 *(0.06%)*	5 203 *(18.3%)*	7 710 *(27.1%)*	3 808 *(13.4%)*	11 741 *(41.2%)*
PREDO I (450k)	817	9 644 *(4 464 VMRs; 5,180 sVMPs)*	3 *(0.03%)*	1 906 *(19.76%)*	1 620 *(16.8%)*	722 *(7.49%)*	5 393 *(55.92%)*
PREDO II (EPIC)	146	24 176 *(6 598 VMRs; 17 578 sVMPs)*	8 *(0.03%)*	4 706 *(19.47%)*	1 824 *(7.54%)*	1 953 *(8.08%)*	15 685 *(64.88%)*

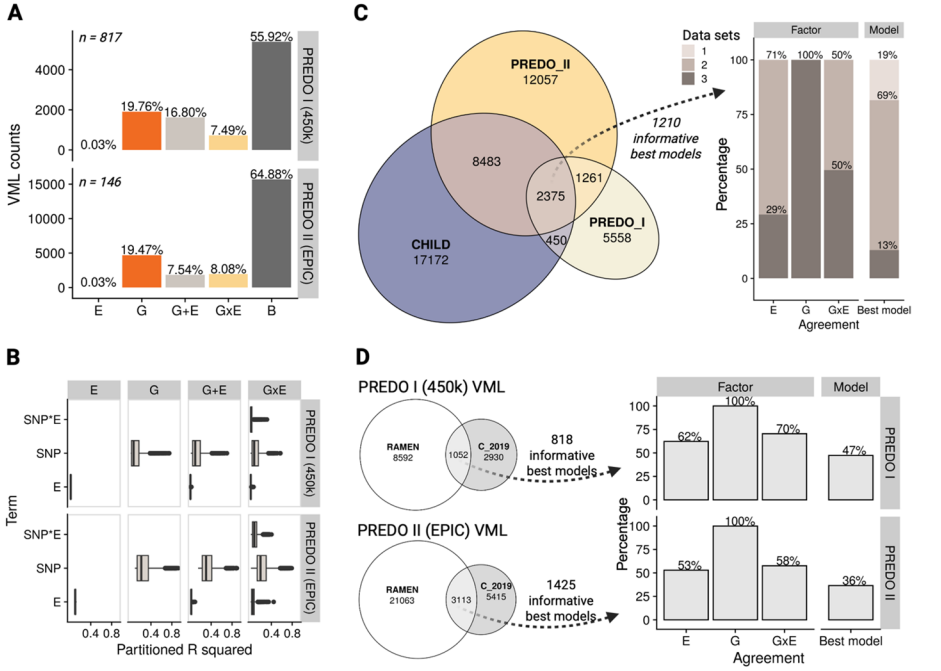


Fig. 5 Gene-environment contributions to DNA methylation variation in PREDO recapitulated CHILD results. **A** Best explanatory models for the set of RAMEN VML in the PREDO cord blood data sets. **B** Partial R^2 of the genetics, environmental, and interaction terms. **C** Factor and model agreement in overlapping VML between the CHILD, PREDO I, and PREDO II data sets using RAMEN. Note that, in the factor agreement, since we analyzed three data sets and we were comparing a binary state (absent/present), the minimum number of agreeing data sets that can exist is 2. **D** Factor and model agreement in overlapping VML between the RAMEN and C_2019 methodology in the PREDO data sets. White fill in plots correspond to results obtained with RAMEN. All the C_2019 results are VMRs, as this method excluded sparse VMPs from the analysis

which we built our ideas. Despite being conceptually similar to the C_2019's approach, RAMEN presents substantial differences as it takes into account the microarray probe design and pairwise CpG correlations to identify sparse VMPs, and implements machine learning and permutation techniques to robustly identify winning models.

Using RAMEN, we identified a lower number of VML in the PREDO data sets compared to CHILD, although the number and composition had higher similarity between data sets derived from the same array platform (i.e., EPIC in CHILD and PREDO II; Table 1). Out of the informative models in PREDO I, G was the most common winning model, followed by G + E, GxE, and E (Additional File 4). In PREDO II, the rank order was G, GxE, G + E, and E (Fig. 5A; Table 1; Additional File 5). Additionally, we observed

a higher proportion of B models in both PREDO data sets compared to CHILD, and a consistent proportion of G and E models across the PREDO and CHILD data sets (Table 1). When dissecting the DNAme variance explained by each term in the winning models, as observed in CHILD, we found the SNP term (PREDO I: *mean* = 0.08, PREDO II: *mean* = 0.16) to explain a significantly higher amount of variance ($p < 0.05$) compared to the interaction term (PREDO I: *mean* = 0.02, PREDO II: *mean* = 0.06) and environmental term (PREDO I: *mean* = 0.01, PREDO II: *mean* = 0.03; Fig. 5B).

To compare the similarity between the results obtained in the PREDO and CHILD data sets, we evaluated the agreement across VML. We assessed agreement using two approaches: (1) factor involvement, where we considered a VML to be consistent for each individual factor (E, G, or GxE) if said factor was both absent or both present in the best model, and (2) overall model, where we considered a VML to be consistent if the best model matched in all factors. For example, if the best model of an overlapping VML was E in the first set of results (model: DNAme ~ E + covars) and G + E in the second (model: DNAme ~ E + G + covars), said VML would be labelled as agreeing in the E factor (since the E term is present in both models), agreeing regarding the GxE factor (since an interaction term is not included in the winning models), not agreeing regarding the G factor (since the G term is not involved in the first set of results but included in the second), and not agreeing in the overall model (since in the first set of results the model is E and in the second is G + E).

We identified 2375 shared VML in the three data sets. Out of the 1210 informative best models (i.e., not belonging to the Basal group in any data set), all of them were in full agreement regarding the G involvement (100%), and less so in GxE (50%) and E (29%). Additionally, 82% of the overlapping VML had a consistent overall model in two to three of the data sets (Fig. 5C). Probes in the overlapping VML were further enriched for probes with bimodal ($p < 0.005$, OR = 12.6) and trimodal patterns ($p < 0.005$, OR = 8.9) in CHILD. Compared to the C_2019 results, RAMEN displayed an identical inter-data set agreement metric regarding G involvement (100%), an increase in the GxE (+ 19%) and overall model agreement (+ 2%), and a decrease in the E model agreement (–12%; Fig. S7, Additional File 1).

Finally, since the C_2019 study used a conceptually similar methodology compared to RAMEN, we compared the results from both methods using the same data (either PREDO I or II) to evaluate their consistency. When comparing the VML obtained with the C_2019 and RAMEN methods, we observed a modest overlap in the PREDO I (1,052; 26%) and PREDO II (3,113; 37%) data sets (Fig. 5D). To evaluate model agreement, we analyzed only overlapping VMRs with informative models in RAMEN, as the Basal best model category is absent in the C_2019 method.

We found a full agreement between the results of the RAMEN and previous C_2019 method regarding the involvement of genetic contributors (PREDO I: 100%, PREDO II: 100%), a modest agreement in the environmental (PREDO I: 62%, PREDO II: 53%) and interaction involvement (PREDO I: 70%, PREDO II: 58%), and a lower agreement in the overall model (PREDO I: 47%, PREDO II: 36%; Fig. 5B).

Discussion

In this study, we analyzed multi-omics data to understand the individual and joint contribution of genomic and prenatal exposome differences to cord blood DNAme variability in general populations. We developed RAMEN, a user-friendly FAIR [39] tool built upon previous methodological approaches in the DNAme field to model regional variability with linear models [33, 34, 42], while incorporating new strategies to tailor the extraction of regional DNAme variability information for microarrays, and to improve the computational efficiency and genome-exposome contribution estimates. We analyzed the CHILD and PREDO cohorts and found genetic variants to be a consistent central contributor to cord blood DNAme variation, frequently in additive and interactive combinations with the prenatal environment. Our results furthermore provided a detailed catalogue of genomic regions with highly variable DNAme, along with the estimated contribution of genetic variants, environmental exposures, and gene-environment interactions to their variability. These findings may inform precision medicine approaches on the potential factors associated with the DNAme variation in CpGs of interest, and improve our understanding of the relation between genetic variants, environmental exposures, and early life DNAme variability patterns.

Compared to conceptually similar works that inspired RAMEN [33–35], our analysis incorporated minor and major changes to refine the analytical framework. Specifically, we used M values to model DNAme, as their properties are better suited for statistical analyses in comparison with beta values [50]. Further, we used variance instead of Median Absolute Deviation (MAD) to identify HVPs, as variance can detect probes with substantial DNAme variation in less than half of the individuals, which would be missed by MAD. When identifying VMRs, we explicitly evaluated the within-region probe correlation to discard proximal probes with dissimilar DNAme patterns. Furthermore, previous studies used the most variable CpG to represent the DNAme level of a VMR; we instead used the median DNAme value across all probes within a region, which aimed to provide a regional summary. In addition to these minor improvements, we deployed the analytical framework in a FAIR software that incorporates additional steps: the analysis of sparse VMPs to improve the capture of methylome variability patterns; a data-driven variable selection step to reduce the computational burden; a variance decomposition analysis to improve the interpretation of the results; and a permutation analysis to discard results where the best explanatory model, despite performing better than other models, still performed as expected by chance.

We modelled a higher number of VML compared to previous perinatal studies even after implementing a within-region probe correlation filter (Table S1, Additional File 1). The number of VML found with RAMEN and their type composition suggests that the increment is driven mainly by the inclusion of sparse VMPs. Sparse VMPs consider the sparse and non-homogeneously covered design of the microarray, by far the most used technology to profile DNA methylomes in epidemiological cohorts. This type of VML are single probes in the array that were filtered out in previous methodologies solely due to sparsity. Taking these sites into account is especially important in the EPIC v1 array, since the DNAme levels of most regulatory regions are measured targeting single probes that accurately represent the methylation state of the surrounding sites, as reported in empirical comparisons with whole-genome bisulfite sequencing data [43]. Our framework optimized the extraction of variability patterns by considering this design, which

led to a 3-fold increase in the number of genomic regions analyzed. Even though the VML identification criteria was created for the EPIC array, we observed them to be equally informative in the 450k microarray. Additionally, the non-random VML genomic distribution we observed is consistent with previous studies reporting highly variable DNAme sites being enriched in open sea, intergenic, transcription, enhancer, quiescent, and repressed regions [33, 44, 51].

A key strength in our study was the computationally efficient detection of models with low evidence of gene-environment effects based on our data (Baseline models), where the genetic and/or environmental contributions behave like spurious associations expected by chance. With this methodological step, we expect our results to provide a likely smaller, yet more refined estimation of the DNA methylome-wide influence of genetic variants and the environment. We found genetic variants as key contributors in most informative cord blood VML models, primarily in combination with the environment, whereas the environment alone explained a notably small proportion of DNA methylation variability. This trend aligned with similar studies in cord blood, umbilical cord, and placenta [33–35]; however, our G, E, G + E, and GxE proportion estimates differ substantially.

We found 36–59% of methylome-wide VML associated with genetic variants, significantly lower compared to previous reports in perinatal integrative studies estimating jointly the contribution of genetic variants and the environment (96–100%) [33–35]. The substantial proportion of variable DNAme sites with a genetic contribution is consistent with recent studies that pinpoint genetic variants as an important component of the DOHaD hypothesis [3, 4]. Additionally, our estimate aligned more closely with studies using significantly different methodological approaches to interrogate specifically the association of genetic variants with DNAme, such as the largest mQTL blood study to date (44% of CpGs under *cis* genetic control) [52] and the whole blood narrow-sense heritability estimates in twins (41% of CpGs with significant additive genetic effects) [30]. Similarly, our estimated proportion of VML with gene-environment interactions on DNAme (~ 10%) is lower than those reported in gene-environment integrative studies in umbilical cord (75%) [34], cord blood (53% on average across 3 cohorts; SD = 8.6) [33], and placenta (70%) [35].

VML best explained by G, E, G + E, and GxE were enriched in distinct locations important for genomic regulation. Consistent with our findings, twin studies in blood at age 18 and epidemiological cohorts at birth in cord blood and placenta have reported DNAme sites with a genetic contribution to be enriched in regions within a 5 kb window upstream of TSSs [31], and sites with an environmental and genetic contribution (G + E) to be enriched in the immediate vicinity of TSS, the 1 st exon of genes, CpG islands, bivalent enhancers, and regions repressed by Polycomb [31, 33, 35]. These findings suggest distinct, potentially conserved pathways through which genetic and environmental factors might associate with variability in the DNA methylome and potentially affect cellular processes across the first 18 years of life.

When dissecting the contribution of each factor in the winning models, we found genetic variants to explain a higher amount of DNAme variance in cord blood VML compared to the environment and interaction terms, and to be fully consistent in overlapping VML across the CHILD and PREDO cohorts. The estimated contribution of genetic variants to DNAme variance across VML (*mean* = 22%) was generally consistent

with the average narrow-sense DNAme heritability estimations in blood family studies (i.e., proportion of variance that can be attributed to additive genetic effects), which ranges from 0.1 to 0.3 [13, 28, 30, 53–56]. We also observed linear GxE interactions to have small contributions (mean partitioned R^2 = 0.01). This suggested that large sample sizes might be required to detect GxE interactions in the DNA methylome, which may also be related to the limited number of significant GxE associations documented in the few DNAme studies looking at epigenome wide GxE [18–22]. VML best explained by GxE models in our results may further be used in follow-up studies to test candidate gene-environment interaction hypotheses, which could help decrease the multiple testing burden and refine precision medicine approaches [23].

The contribution of environmental factors was less consistent than SNPs in overlapping VML across cohorts (i.e., E and GxE terms). This might be a consequence of the E and GxE factors having a lower average contribution to the DNAme variance, requiring a greater statistical power to be consistently detected. One explanation for this low contribution could be the presence of specialized extra-embryonic tissues during gestation, such as the placenta, that protect the fetus and directly regulate the prenatal environment [57, 58]. These extra-embryonic tissues might dampen the overall environmental exposures effect in cord blood. Alternatively, the low concordance and environmental contribution might result from known biases affecting the power to detect and replicate exposome effects, such as the small number of measured variables, heterogeneity of measurements and their errors, reliance on indirect exposures assessments (e.g., self-reported questionnaires), difficulty capturing both the effector exposures at the relevant timepoint and its duration, lack of standardized protocols to measure the heterogeneous and dynamic nature of the exposome, low effect sizes, and differences in the exposures measured across epidemiological studies [59–62].

A significant proportion of VML were best explained by the Baseline (B) model, which means that none of our G and E variables, individually or jointly, substantially improved the model performance. We observed that this proportion increased with higher rates of missing information, which was evident in the SNP sub-sampling experiment in CHILD and the higher B proportion in PREDO I compared to CHILD (similar sample sizes but fewer E variables in PREDO I). Additionally, we observed a higher proportion of B models in PREDO II compared to PREDO I (same number of G and E variables but a smaller sample size in PREDO II). As model selection based on AIC is not part of a statistical inference framework, a formal power analysis in our work is not applicable (for guidance on statistical power in DNAme analyses see Mansell et al. [63]); yet, this observation suggests that sample size plays a role in the capacity to detect informative models. Lastly, the DNAme of some VML might be governed by stochastic processes [64]. Therefore, we hypothesize that non-explained variance in the B models may be associated with a combination of unmeasured variables, sample size limitations, and stochasticity.

Our study has limitations and challenges that must be considered when interpreting results. Regarding the method, our framework is designed to assess gene-environment interactions only if the corresponding SNP and exposure are potentially relevant individually, since we conduct feature selection before fitting the models. Additionally, in cases of high correlation between individual G or E variables (e.g., groups of SNPs with a large amount of linkage disequilibrium or environmental exposures highly correlated), our feature selection strategy will select only one at random for each VML, meaning our

results should be interpreted as associations at the G or E level. Further studies with refined strategies and study designs are needed to fine-map causal factors and gene-environment combinations, which can be particularly challenging in cases of gene-environment correlation (i.e., different genotypes systematically experiencing specific environments) [65].

To reduce computational time and model complexity, we modeled variation as univariate in G and E, although combinations of multiple G or E factors might contribute to DNAme variability to a smaller degree. Given the polygenic nature of DNAme [52], approaches incorporating multiple G or E variables and their interactions could improve the variance explanation, although the number of possible models was considered too high to incorporate into our framework. Additionally, our method estimated the variance explained by each G, E and GxE factor in the best informative models after accounting for the variance explained by the concomitant variables; the variance explained by genetic and environmental factors might be underestimated when collinear with the concomitant variables.

Finally, our results provide a comprehensive characterization of gene-environment effects on cord blood DNA methylome variation in general populations. As DNAme presents cell type-, developmental-, population- and disease-specific patterns [7–9, 12, 66], further studies in different tissues and cohorts will be helpful to understand this phenomenon in humans. From a temporal perspective, our study focused on the prenatal period. As twin studies have suggested that environmental effects on DNAme variance increase with age in 10.4% of CpGs [30], longitudinal studies will be instrumental to assess the gene-environment contribution across the human lifespan. Furthermore, in our study we used microarrays to profile the methylome, which are designed to target primarily non-repetitive regions. DNAme at repetitive regions, such as sub-telomeric repeats and rDNA sequences, have been reported to be especially responsive to the environment [67, 68]. Long-read sequencing technologies are needed to explore the G and E contribution in these genomic elements. Lastly, our cohorts primarily included socio-economically privileged individuals from Western, Educated, Industrial, Rich and Democratic (WEIRD) countries. Studying genetically and/or socioeconomically homogeneous populations can bias the characterization of VML and the estimation of environmental and genetic variant effects. Systemic efforts to discontinue colonialist and unethical practices are needed to include historically underrepresented populations in research, clarify the role of genetic variants and the environment in the DNA methylome and conduct research relevant to all individuals in global societies.

Conclusions

We identified genetic variants as key contributors to DNAme variability in cord blood, frequently in combination with the prenatal environment, supported by a large fraction of informative models including G, a high proportion of DNAme variance explained, and full consistency across cohorts. We reported a detailed genomic map of regions with high interindividual DNAme variation and identified the potential factors explaining their variation, which could be a powerful resource to annotate CpGs, inform precision medicine approaches, and identify candidate GxE interactions in perinatal cohorts. We also introduced a FAIR tool to estimate the genome-exposome additive and interactive contribution to the DNA methylome. Our findings highlight the importance of

accounting for genetic effects in DNAme studies, and contribute to the understanding of how and under which conditions individual genetic susceptibility and environmental variables may work together to influence DNAme in a highly sensitive developmental period.

Methods

All analyses were performed using the R Statistical Software v4.2.2 [69] and the tidyverse framework [70].

Cohort descriptions

CHILD

The CHILD study is a national longitudinal birth cohort following infants from prenatal to age 8 years (with older assessments ongoing) with participants across four provinces in Canada: British Columbia, Alberta, Manitoba, and Ontario. Inclusion and exclusion criteria can be found in the cohort description publication [36]. For the current study we used a subset of 699 cord blood samples from term (>36 weeks) newborns with genome, methylome, and exposome data that passed quality control; this subset of samples within the larger CHILD cohort was selected for multi-omic profiling because of the availability of longitudinal biological samples at birth, age one, and age five. The self-reported ethnicity of these newborn's mothers was 76.7% Caucasian White, 13.2% Asian (including East Asian, South Asian, and South East Asian), 3% First Nations, 1.6% Black, 1.1% Hispanic, and 4.4% other (including multiethnic and Middle Eastern), and their average age was 33.3 (*SD* = 4.6). The assigned sex at birth of the individuals was 46% female and 54% male. Gestational age at delivery was 39.7 weeks on average (*SD* = 1.2 weeks). The distribution of these demographic variables were similar to the ones in the full study [37].

PREDO

The PREDO study is a longitudinal multicenter perinatal cohort from Finland [71]. Briefly, the study recruited pregnant women and their singleton children born alive between 2006 and 2010. Inclusion criteria were either pregnant women with known clinical risk factor status for preeclampsia and intrauterine growth restriction (IUGR), or pregnant women who volunteered to participate who did not present risk factors. Further details about study design can be found in the cohort description publication [70]. For this study, we used a subset of 963 cord blood samples from ethnically homogeneous Finnish newborns with genome, methylome, and prenatal exposome data.

DNA methylation

CHILD sample collection and DNAme profiling

Specimens were collected directly from an umbilical vein before the placenta was delivered in heparinized vacutainers, pooled in a 50 mL conical tube, aliquoted, and frozen [37]. DNA was extracted from cord blood samples using the DNeasy Blood & Tissue Kit (Qiagen), samples were bisulfite converted using EZ-96 DNA Methylation kit (Zymo Research) and DNA methylation profiles of the samples were measured with the Infinium MethylationEPIC BeadChip v1 array (Illumina).

CHILD DNAme pre-processing

Cord blood DNAme data was pre-processed and subjected to quality control checks in R v4.0.3. Sample exclusion criteria were the following: failure in *EWAStools* v1.7 [72] control metrics, reported-predicted sex mismatch inferred using sex chromosome intensities (*minfi* v1.44 [73]), EPIC-GSA SNP mismatch, maternal cord blood contamination [74], detection *p* value, >1% missing probes measured, and outliers (detected with the *lumi* [75] package). Normalization was next conducted using BMIQ with noob to account for probe type bias and background correction implemented in the *wateRmelon* [76] v2.4 and *minfi* v1.44 packages, respectively. Probe exclusion criteria were the following: SNP probes, detection *p* value >0.01 in >5% samples, present in Pidsley et al.'s cross-hybridizing and polymorphic EPIC probes (i.e., overlapping with common genetic variants) [43], used in Haftorn's 2021 EPIC gestational age clock [77] or in the *FlowSorted.Blood.EPIC* [78] cord blood IDOL's probes, and located in non-somatic chromosomes. As a result of these steps, 785,689 out of 866,836 CpG probes and 813 out of 828 samples were retained. Finally, batch effects associated with technical variation (chip and row) were removed using the ComBat function from the *sva* package [79].

PREDO DNAme

The DNAme data pre-processing was conducted as described previously [33]. Briefly, the data set consisted of 817 cord blood samples that were ran in Illumina 450k Methylation arrays, and 146 in Illumina EPIC arrays. We conducted quality control filters regarding minfi [73] control metrics, median intensity, reported-estimated sex and maternal DNA contamination, CpG detection p value. Methylation beta-values were normalized using the *funnorm* [80] function, and variation associated with batch effects was further removed using the ComBat function from the *sva* package [79]. Probes in sex chromosomes and reported to be cross-hybridizing were excluded. The final data set consisted of 418,790 and 794,414 CpG probes for the Illumina 450 K and EPIC arrays respectively.

Genotyping

CHILD pre-processing

DNA was extracted from cord blood samples as mentioned in the DNAme section, and genotyping was performed using the GSA v3 + Psych v1 array (Illumina). Following GenomeStudio v2.0.4 (Illumina) preliminary quality control checks, we exported the genotyping data to R v4.2.0 for further QC guided by Illumina recommended metrics [81]. Detailed thresholds can be found in Supplementary Methods (Additional File 1). Sample relatedness was checked by calculating kinship coefficients (identity by descent) based on Maximum Likelihood Estimation (MLE) in the R SNPRelate package [98]. Based on a kinship coefficient score of 0.5, 3 pairs of related samples were identified which were technical replicates and, being sample duplicates, were expected to be identical. Only one sample from the 3 pairs was retained for downstream analysis. Probes in non-autosomic chromosomes were excluded from the analysis. Imputation was conducted using the ENIGMA imputation protocol [99] based on the Michigan Imputation Server pipeline [82]. After imputation, imputed SNPs with an $R^2 \leq 0.8$ and MAF ≤ 0.01 were filtered out using *bcftools* v1.16 [83]. Finally, we conducted LD pruning using *plink* 1.9 [84] with a window size of 50 variant counts, a 5 variant count window shift and a

0.5 pairwise R^2 threshold. The final genotyping consisted of a data set of 1,260,703 SNPs (492,062 directly measured) and 824 out of 826 individual samples.

PREDO genotyping pre-processing

The genotype data pre-processing was conducted as described previously [33]. Briefly, genotyping was performed on Illumina Human Omni Express Exome Arrays containing 964,193 SNPs. We conducted quality control filters regarding call rate, MAF, HWE, relatedness, reported-predicted sex, and heterozygosity outliers. Following imputation, we ran another round of quality control regarding SNP info score, MAF, HWE, *p* value, and call rate. Next, the genotype data set was pruned using a threshold of r2 of 0.2 and a window-size of 50 SNPs with an overlap of 5 SNPs. The final data set consisted of 983 samples with 788,156 SNPs.

Prenatal exposome

CHILD prenatal exposome

We used 94 prenatal environmental variables collected in the CHILD cohort encompassing four fetal exposure dimensions: maternal psychosocial state, maternal nutrition, maternal health, and built environment. Briefly, the maternal health dimension included 46 variables such as conditions during pregnancy (e.g., cold, diabetes, allergy, hypertension, etc.) and smoking; the maternal nutrition dimension included 29 variables such as dietary patterns, Healthy Eating Index 2010 scores, and dietary constituents that mediate 1-carbon metabolism; the parental psychosocial dimension included 10 variables such as maternal depressive symptoms and socio economic status; and the built environment dimension included 9 variables such as NO_2, $PM_{2.5,}$ and O_3. Detailed information about the pre-processing of these variables can be found in supplementary methods (Additional File 1).

PREDO prenatal exposome

Of the PREDO data set collected, we included 10 prenatal environmental variables described previously [33]: maternal depressive and anxiety symptoms, maternal treatment with betamethasone, delivery mode, parity, maternal age at delivery, pre-pregnancy BMI, hypertensive pregnancy disorders (including gestational hypertension, chronic hypertension, and preeclampsia), gestational diabetes, and glucose values during a 75 g 2 h oral glucose tolerance test (OGTT).

Prenatal exposome pre-processing

For CHILD, each dimension was pre-processed independently but using the same criteria as described: (1) Individuals with > 30% of missing values were removed; (2) Variables that displayed > 15% of missingness were removed; (3) Variables that measured a similar object and were highly correlated ($r > 0.8$) or displayed low variability (< 10 cases in binary variables) in the dataset were removed; (4) Remaining missing values were single imputed using the *mice* package in R [85] with 100 iterations and the default method. For the psychosocial dimension, household income when the child was 1 year old was included as a covariate in the imputation due to its high correlation with income at birth and therefore its potential to improve the imputation accuracy. After this pre-processing, the four dimensions were combined into a single dataset, and individuals

with missing data in one or more dimensions were removed, which resulted in a dataset of 699 out of 790 individuals with complete environmental exposure information of 94 variables (maternal nutrition = 29, maternal health = 44, maternal psychosocial = 10, built environment = 9). A full list of the variables included for this study and their descriptive statistics can be found in Additional File 2. The pairwise correlation between variables can be found in Fig. S8.

The PREDO data set was pre-processed in the same way treating all the variables as a single dimension. The final data set consisted of 8 environmental variables: all those mentioned in the PREDO prenatal exposome section except for betamethasone and OGTT.

Concomitant variables

CHILD

The following DNAme confounders, which can lead to biased or spurious results, were considered "concomitant" variables: predicted cell type composition and population stratification (captured by global genetic ancestry). Additionally, the following sources of DNAme variation not directly related to genetic or environmental variation were also considered concomitant: sex (Female/Male) and gestational age (days at delivery). Concomitant variables were included as covariates in all models of DNAme variability.

Cord blood cell type proportions of CD8 + T cells, CD4 + T cells, Natural Killer, B cells, Monocytes, Granulocytes, and nucleated red blood cells were estimated from raw DNAme data using the *FlowSorted.Blood.EPIC* R package [79] with the Noob pre-processing method and both the IDOL and hypo/hypermethylated probes to distinguish cell types. To address the compositional and multicollinear nature of cell type proportions, we applied an isometric logratio transformation followed by a robust PCA as proposed by Filzmoser et al. in 2009 [86]. The first four components' eigenvectors were used, which explained 92% of the variance.

Global genetic ancestry was estimated by projecting each sample's genotypes on the principal components calculated using the reference 1000 Genomes Project phase 3 data [87]. The eigenvectors of the first four principal components, which captured the population stratification by visual assessment, were used.

PREDO

Like CHILD, the newborn's sex (F/M), predicted cord blood cell type proportions, genetic ancestry, and gestational age at birth (days) were used as covariates in all models. To compare our results with the ones obtained previously using this cohort, the variables were obtained as described before [33]. Briefly, cord blood cell counts were estimated and incorporated directly into the models (all but one to address their compositional nature). The eigenvectors of the first two genetic PCs were used to capture genetic ancestry.

Genome-exposome contribution to methylome variability analysis

The genome-exposome contribution to methylome variability analysis was conducted with the RAMEN v1.0.0 R package (Fig. 1). The sections below detail its methodological steps.

Identification of Highly Variable Probes

To identify HVPs without the effect of known DNAme confounders, we first regressed out the linear effect of the concomitant variables (sex, gestational age at delivery, cell type proportions, and genetic ancestry) on DNAme M values (Fig. 1A, Step 1.1). The residuals of this regression were used to compute the variance of each measured CpG probe (Fig. 1A, Step 1.2), and a cutoff of the 90th percentile of the distribution was used to identify HVPs (Fig. 1A, Step 1.3). Residual M values were only used for HVP identification; unadjusted M values were used in the rest of the analysis.

Identification of Variable Methylated Loci

HVPs were grouped into Variable Methylated Loci (VML; Fig. 1A, Step 2.1). We divided microarray VML into two categories: VMRs and sVMPs. VMRs are regions that meet traditional grouping criteria: two or more HVPs that are close together (chained in < 1 kb windows) and share a similar DNAme profile (r >0.15 based on Gatev et al.'s [88] simulation to empirically determine DNAme regional correlation thresholds). On the other hand, we created the sVMP category to account for the sparsity and non-homogeneous coverage in the EPIC microarray design. This array has a high number of DNA regulatory regions measured by single probes that accurately represent regional DNAme levels [43]. With this category, we aimed to recover valuable information from HVPs that had been excluded in previous studies solely because of the genomic array design. DNAme level of each VMR was further summarized per individual by taking the median DNAme value per region (Fig. 1A, Step 2.2).

Genome and exposome variable selection

For each VML we pre-selected their *cis* SNPs (within a 1 Mb upstream and downstream window; Fig. 1A, Step 2.3). Then, we implemented a scalable variable-selection strategy prior to the model fitting and comparison to reduce the computational burden and address the substantial imbalance in the number of genome and exposome variables, which can introduce a methodological model selection bias towards the category of variables with a substantially higher number of variables. We conducted the variable selection to keep the potentially relevant genome and exposome variables using a method based on Least Absolute Shrinkage and Selection Operator (LASSO; Fig. 1A, Step 3.1). LASSO is an algorithm with a variable screening property that penalizes more complex models (i.e., with more variables) in favor of simpler ones (i.e., with less variables), but not at the expense of reducing the predictive power. This decreases the number of variables in a model, the downstream computational time, and can improve model accuracy [49]. Briefly, for each VML we conducted a LASSO regression in three scenarios: (1) in the presence of only the *cis* SNPs, (2) in the presence of only environmental variables, and (3) in the presence of both *cis* SNPs and environmental variables. Concomitant variables were included in all models and were unpenalized during LASSO. Further details on the implemented LASSO approach can be found in supplementary methods (Additional File 1). For each VML, all variables selected in scenarios 1–3 (i.e., with a coefficient >0) were pooled and moved to the model fitting stage.

Table 2 Models fitted and compared in RAMEN per VML

Model	Name	Abbreviation
DNAme ~ SNP_i + covars	Genetic variants	G
DNAme ~ EE_j + covars	Environmental exposure	E
DNAme ~ SNP_i + EE_j + covars	Additive	G + E
DNAme ~ SNP_i + EE_j + SNP_i*EE_j + covars	Interaction	GxE

Selection of best explanatory model

For each VML, all selected genetic variant (SNP_i) and environmental exposure (EE_j) variables were individually used to fit the models in Table 2 (Fig. 1A, Step 3.2). Following the model fitting, the best explanatory model for each VML was selected based on the models' lower Akaike Information Content (AIC; Fig. 1A, Step 3.3). The use of this metric assumed that most of our models were a simplification of the underlying truth (where more than one G and E variable is likely to contribute to the methylome variance). Briefly, AIC penalizes models with higher number of terms, and excels in problems where all models in the model space are considered incorrect and in cases where the selected model could belong to a very large class of functions describing complex relations [89]. Following the model selection, the variance explained by the SNP and E terms in the best models was estimated using the *relaimpo* R package [90] which uses the Linda, Merenda and Gold (LMG) method [91].

DNAme represents the summarized regional DNAme level, *SNP*EE* the gene-environment interaction term, and *covars* the previously described concomitant variables (sex, gestational age at delivery, cell type proportions, and genetic ancestry).

Finally, to filter out best explanatory models which G/E contribution could be explained by chance, we conducted a permutation analysis (Fig. 1A, Step 3.4) of the G and E contribution to VML's DNAme. To do so, we shuffled the G and E variables together in the data set and repeated the evaluation of G and E contribution (Fig. 1A, Steps 3.1–3.3) 10 times. We used the R^2 G/E increment (i.e., $R^2_{BEM} - R^2_{BM}$; where *BEM* = Best Explanatory Model and *BM* = Baseline Model including only the concomitant variables) of those permutations to create a null R^2 increment distribution. We obtained a bimodal R^2 G/E increment distribution that we stratified into two groups: marginal (G and E), and joint (G + E and GxE) models. The 95th percentile of the null distributions was used as a cut-off to remove poor performing best models in their corresponding strata and assign the best explanatory model of the corresponding VML as Baseline.

It is worth mentioning that instead of the permutation analysis, we first tried to remove false positives from the set of Best Explanatory Models selected with AIC by using an F-test. For each VML, we contrasted the selected G, E, G + E, or GxE model with the nested Baseline one (i.e., only the covariates) to identify models that, despite having the lowest AIC, were not significantly better than the baseline. We discarded this strategy as it produced a highly right skewed p-value distribution where all tested models were better than the baseline (Fig. S9, Additional File 1), likely due to the variable and model selection steps that were conducted before.

Total number of G, E, G + E, and GxE models without variable selection

We computed the total number of models that would have been fitted and compared without the variable selection stage as follows:

$$Total\ models = \sum_{i=1}^{n} G_i + E_i + 2\,(G_i * E_i)$$

Where n is the total number of VML and i its index, G is the number of *cis* SNPs in a given VML, and E is the number of environmental exposures (94 for all VML). In lay terms, for each VML, the number of models fitted is the sum of (a) one model for each SNP in *cis*, (b) one model for each environmental exposure (E = 94), and (c) two models (one additive and one interaction) for each possible SNP and E combination. The sum of this was 5,279,537,086 models.

Comparison of HVPs with derakhshan's highly variable CpGs

To compare the HVPs in our study with previously reported human tissue- and ethnicity-independent highly variable CpGs (hvCpGs) [44], we downloaded the list of probes from the supplementary data in Derakhshan et al.'s publication. In summary, they analyzed 30 data sets that covered 8 ethnicities and 19 tissues/cell types spanning a wide range of ages (0.75–108 years old). There, hvCpGs were defined as probes covered in at least 15 of the 30 data sets and with a methylation variance in the top 5% of all (non-removed) CpGs in ≥ 65% of data sets where the CpG was covered (after qc). Since this hvCpGs data set was made with Illumina 450k array data, we conducted the comparison only in the subset of probes present in both the EPIC and the 450k array.

mQTL analysis

We conducted an mQTL analysis using the R package *Matrix eQTL* v2.3 [92] on a superset of 790 cord blood samples profiled for DNAme and genotyped as described above. These 790 samples correspond to the 699 individuals used for the genome-exposome contribution to methylome analysis plus 91 samples from the same cohort (CHILD) that were excluded from said analysis due to incomplete exposome information. The covariates included in the association test are the same ones described in the *Concomitant variables* section above. *Cis* (SNP-CpG pars within 1 Mb) mQTLs were detected with a p-value threshold of 1×10^{-5}. The significant *cis* mQTLs found in the 207,199 CpGs (26.47% of the probes) can be found in Additional Files 6 and 7.

Identification of probes with trimodal and bimodal patterns

We applied the *gaphunter* [45] function present in the *minfi* v1.44 package [73] to the beta DNAme values of the 785,689 probes in our analysis using the default arguments.

Enrichment analyses

Enrichment of bimodal and trimodal patterns

To test the enrichment of probes with bimodal and trimodal patterns in the CHILD set of HVPs and probes in the 2,375 overlapping VML between CHILD and the PREDO data sets, we conducted a Fisher's test in R. All probes in the CHILD DNAme data set (785,689) were used as background.

Enrichment of mQTLs in selected SNP and best models

To test if (A) the SNPs selected in the variable selection step and (B) the SNPs present in the best model per VML (if G was a component of the selected model) were more likely to be mQTLs, we conducted a chi-squared test. For the over enrichment analysis in A

we used the selected *cis* SNPs as target and all the SNPs in *cis* of a VML as background. For the enrichment analysis in B, we used the *cis* SNPs in all the winning models as target and the selected *cis* SNPs as background.

Enrichment of CpG categories in VML and best models

We used the CpG manifest provided by Illumina to annotate the CpG probes regarding gene annotation and island context using the *UCSC_RefGene_Group* and *Relation_to_UCSC_CpG_Island* columns. Additionally, we used the core 15-state model of cord blood T cells (epigenome E033) generated by the Roadmap epigenomics project [93] (downloaded from https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html) to annotate the CpG probes regarding chromatin state. We then conducted an over-representation analysis of each CpG category by doing a one side Fisher's test with the *clusterProfiler* R package [94]. For the over representation analysis in VML we used all the probes in VML as target and the probes in the EPIC array as background. For the over representation analysis in best models, we annotated each VML according to its probes (e.g., if a VML was composed of two probes that were annotated to Tx and Enh respectively, we annotated said VML as Tx and Enh); we then used the VML in each best model group as targets and all VML as background.

Information imbalance analysis

To explore the effect of the genome-exposome information imbalance in the contribution to DNA methylome variability analysis, from the set of *cis* SNPs we randomly sampled 10, 100, 500, and 1000 SNPs for each VML. In the cases where a given VML had less *cis* SNPs than the required number to be sampled, no random sampling was conducted. Next, the entire G&E contribution modeling phase of the analysis was conducted (Fig. 1A3) to get the proportion of best models in each model category. This procedure was repeated 5 times to estimate the variability of the results.

Variable imbalance analysis

To dissect the effect of the genome-exposome variable imbalance in the contribution to DNA methylome variability analysis, we chose to explore this in a null hypothesis scenario. In a context where the genome and exposome have no effect on DNAme variability (i.e., no information imbalance in the variables since none of them are associated to DNAme), we could expect the variable number imbalance to be the main driver of the differences in G and E best models' proportion, therefore exploring the sensitivity of the method itself. We shuffled the G and E variables together in the data set and conducted the G and E contribution part of the analysis (Fig. 1A3.1–3.3) with the original number of *cis* SNPs (i.e., a higher amount of G variables than E).

Model and factor agreement

Overlapping VML were defined as regions with at least one overlapping genomic position and were identified with the *plyranges* R package v 1.18.0 [95]. Model agreement in overlapping VML was evaluated in two modalities: (A) overall model and (B) factors. To improve comparability, VML with B models best explaining their DNAme were excluded from all comparisons, since B models can increase the non-agreement percentages due to confounding factors such as differences in information content between cohorts,

sample size, and methodology (specifically regarding Czamara-2019 [33] where the B category does not exist). For (A), we treated each best model as an independent category and labelled overlapping VML from two datasets as concordant if they belonged to the same group (e.g., best model being G + E in both data sets). To account for the fact that models are not independent categories, but rather reduced forms of a full interaction model, we next evaluated the individual factor agreement. For (B), overlapping VML from two datasets were labelled as agreeing regarding G, E, or GxE if said factor was both present or both absent in the best model. For example, if one model were E (model: DNAme ~ E + covars) and the other G + E (DNAme ~ E + G + covars), that VML would be labelled as agreeing regarding E, since E is present in both models, agreeing regarding GxE, since the model agrees that an interaction does not best explain the DNAme, and not agreeing regarding G since G is not involved in the first scenario but included in the second.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03864-4.

Additional file 1. Supplementary figures, tables and methods [100–115].

Additional file 2. Statistics and details of CHILD exposome variables.

Additional file 3. Results of the genome-exposome contribution analysis to methylome variability in CHILD cord blood samples.

Additional file 4. Results of the genome-exposome contribution analysis to methylome variability in PREDO I cord blood samples.

Additional file 5. Results of the genome-exposome contribution analysis to methylome variability in PREDO II cord blood samples.

Additional file 6. Significant cis mQTLs in CHILD cord blood samples (chr1-6; *p*-value < 1x10-5).

Additional file 7. Significant cis mQTLs in CHILD cord blood samples (chr7-22; *p*-value < 1x10-5).

Acknowledgements

We thank the families that took part in the CHILD and PREDO cohort studies for their dedication and commitment to advancing health research, as well as the team involved in the study design, sample collection, and data availability, such as interviewers, and clinical and administration staff. We are grateful with Dr. Angela Devlin and Dr. Alejandra Wiedeman for providing valuable input for processing the nutrition variables, Dr. Meingold Chan for providing guidance in the processing and scoring of the psychological questionnaires, Dorothy Lin and Carlos Cortés-Quiñones for helping with the creation of the RAMEN logo, and Marcia Jude for conducting the global genetic ancestry analysis in the CHILD samples.

Peer review information

Hamish King and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

E.I.N.D., K.K. and M.S.K. conceptualized the project. E.I.N.D. conceived the analyses, developed the RAMEN package, conducted the exposome pre-processing, genotype imputation and genome-exposome contribution analysis in CHILD and wrote the manuscript. D.C. conducted the data pre-processing and genome-exposome contribution analysis using RAMEN in PREDO. K.E. conducted the CHILD DNAme pre-processing. E.I.N.D and K.E. conducted the CHILD genotype pre-processing. E.I.N.D, K.K, M.S.K., M.P.F., S.M.M., and C.K. made intellectual contributions to the project analysis and results interpretation. J.L.M. and D.T.S.L. ran the DNAme and genotyping arrays in CHILD. K.K. supervised the statistical analyses. P.M, T.J.M, E.S., P.S. and S.E.T. conceptualized and planned the CHILD study and data collection. K.R. conceptualized and planned the PREDO study and data collection. K.R. and J.L. were involved in the cord blood DNAme profiling in the PREDO study. All authors contributed to and approved the final version of the manuscript.

Funding

This work was supported by BC Children's Hospital Research Institute Establishment Award (to K.K.). E.I.N.D. was funded by the Graduate Globalink Fellowship from MITACS, the Society to Cell Clyde Hertzman Memorial Fellowship from the Social Exposome Cluster, the Gertrude Langridge Graduate Scholarship in Medical Sciences, the Patrick David Campbell Graduate Fellowship, the Bank of Montreal Graduate Fellowship and the 4-Year PhD Fellowship from the University of British Columbia. The CHILD Cohort study was supported by core funding from The Allergy, Genes and Environment Network of Centres of Excellence (AllerGen NCE) and the Canadian Institutes of Health Research (CIHR); for further information visit childstudy.ca. CHILD epigenetic analysis at the University of British Columbia was funded by CIHR,

AllerGen NCE and Genome Canada and Genome British Columbia ([274CHI]). S.E.T. holds a Tier 1 Canada Research Chair in Pediatric Precision Health and the Aubrey J. Tingle Professor of Pediatric Immunology. M.S.K. is the Canada Research Chair in Social Epigenetics and the Edwin S.H. Leong Chair in Healthy Aging - a UBC President's Excellence Chair. The PREDO study was supported by funding from the Academy of Finland, European Union's Horizon Europe research and innovation program under grant agreement No 101057390 (HappyMums), Helsinki Institute of Life Sciences (HiLife) Fellows Programme 2023–2024, and HUS VTR (A special Finnish state subsidy for health science research).

Data availability

The participant data (genotype, prenatal exposome, and DNA methylation) is not publicly available for the protection of participants of the CHILD and PREDO studies. Access to this data can be obtained by request; if interested, contact the PREDO (predo.study@helsinki.fi) or CHILD (child@mcmaster.ca) study boards. For more information, please see https://childstudy.ca. The result datasets supporting the conclusions of this article are included within the article (and its additional files). Summary statistics of prenatal exposures and results are available in Additional files 2–5. The RAMEN package, a platform independent tool in R with a GLP-3.0 license, and the code used to conduct the analyses in this article can be found in GitHub repositories [96, 97].

Declarations

Ethics approval and consent to participate

The CHILD study adheres to all applicable ethical guidelines and was developed and authorized by the University of British Columbia, University of Manitoba, University of Toronto, McMaster University, BC Children's Hospital, The Hospital for Sick Children, and Simon Fraser University; the Research Ethics Board (#H07-03120) examined and authorized this study in accordance with the Tri-Council Policy Statement: Ethical Conduct for Human Research (TCPS2, 2018). The PREDO study protocol was approved by the Ethics Committee of Obstetrics and Gynaecology and Women, Children and Psychiatry of the Helsinki and Uusimaa Hospital District and by the participating hospitals; all participants provided written informed consent. Consent of participating children was provided by parent(s)/guardian(s).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Edwin S. H. Leong Centre for Healthy Aging, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[2]British Columbia Children's Hospital Research Institute, Vancouver, BC, Canada
[3]Centre for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver, BC, Canada
[4]Department of Genes and Environment, Max-Planck-Institute of Psychiatry, Munich 80804, Germany
[5]Department of Psychology, University of Massachusetts Lowell, MALowell, United States
[6]University of Alberta, Edmonton, AB, Canada
[7]UCSI university, Kuala Lumpur, Malaysia
[8]Department of Pediatrics & Child Health, University of Manitoba, Winnipeg, MB, Canada
[9]Translational Medicine Program, The Hospital for Sick Children, Toronto, ON, Canada
[10]Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[11]Folkhäsan Research Centre, Helsinki, Finland
[12]Department of Psychology and Institute for Policy Research, Northwestern University, Evanston, IL, United States
[13]Department of Psychology, Faculty of Medicine, University of Helsinki, Helsinki 00014, Finland
[14]Department of Obstetrics and Gynecology, Helsinki University Hospital (HUS), Helsinki 00290, Finland
[15]Department of Pediatrics, Faculty of Medicine, University of British Columbia, Vancouver, Canada
[16] Department of Statistics, University of British Columbia, Vancouver, Canada
[17]Department of Medical Genetics, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

Received: 3 June 2025 / Accepted: 7 November 2025
Published online: 10 December 2025

References

1. Hoffman DJ, Reynolds RM, Hardy DB. Developmental origins of health and disease: current knowledge and potential mechanisms. Nutr Rev. 2017;75:951–70.
2. Padmanabhan V, Cardoso RC, Puttabyatappa M. Developmental Programming, a pathway to disease. Endocrinology. 2016;157:1328–40.
3. Consortium EGG, et al. Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. Nat Genet. 2019;51:804–14.
4. CHARGE Consortium Hematology Working Group. Genome-wide associations for birth weight and correlations with adult disease. Nature. 2016;538:248–52.
5. Felix JF, Cecil CAM. Population DNA methylation studies in the developmental origins of health and disease (DOHaD) framework. J Dev Orig Health Dis. 2019;10:306–13.
6. Bianco-Miotto T, Craig JM, Gasser YP, van Dijk SJ, Ozanne SE. Epigenetics and dohad: from basics to birth and beyond. J Dev Orig Health Dis. 2017;8:513–9.
7. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. Lancet. 2018;392:777–86.

8. Greenberg MVC. Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. 2019;20:590–607.
9. Flanagan JM, et al. Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study. Cancer Epidemiol Biomarkers Prev. 2015;24:221–9.
10. Apsley AT, et al. Biological stability of DNA methylation measurements over varying intervals of time and in the presence of acute stress. Epigenetics. 2023;18:2230686.
11. Mulder RH, et al. Epigenome-wide change and variation in DNA methylation in childhood: trajectories from birth to late adolescence. Hum Mol Genet. 2021;30:119–34.
12. Boyce WT, Kobor MS. Development and the epigenome: the 'synapse' of gene–environment interplay. Dev Sci. 2015;18:1–23.
13. Villicaña S, Bell JT. Genetic impacts on DNA methylation: research findings and future perspectives. Genome Biol. 2021;22:127.
14. Martin EM, Fry RC. Environmental influences on the epigenome: Exposure- associated DNA methylation in human populations. Annu Rev Public Health. 2018;39:309–33.
15. Gaunt TR, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17:61.
16. Brunst K. *Epigenome-Wide Meta-Analysis of Prenatal Maternal Stressful Life Events and Newborn DNA Methylation*. https://www.researchsquare.com/article/rs-1906930/v1 (2022) https://doi.org/10.21203/rs.3.rs-1906930/v1
17. Heijmans BT, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc Natl Acad Sci U S A. 2008;105:17046–9.
18. Argos M, et al. Screening for gene–environment (G×E) interaction using omics data from exposed individuals: an application to gene-arsenic interaction. Mamm Genome. 2018;29:101–11.
19. Yousefi M, et al. The methylation of the LEPR/LEPROT genotype at the promoter and body regions influence concentrations of leptin in girls and BMI at age 18 years if their mother smoked during pregnancy. Int J Mol Epidemiol Genet. 2013;4:86–100.
20. Klengel T, et al. Allele-specific FKBP5 DNA demethylation mediates gene–childhood trauma interactions. Nat Neurosci. 2013;16:33–41.
21. Bergstedt J, et al. The immune factors driving DNA methylation variation in human blood. Nat Commun. 2022;13:5895.
22. Schaffner SL, et al. Genetic variation and pesticide exposure influence blood DNA methylation signatures in females with early-stage parkinson's disease. Npj Parkinsons Dis. 2024;10:98.
23. Motsinger-Reif AA, et al. Gene-environment interactions within a precision environmental health framework. Cell Genomics. 2024;4:100591.
24. Virolainen SJ, VonHandorf A, Viel KCMF, Weirauch MT, Kottyan LC. Gene–environment interactions and their impact on human health. Genes Immun. 2022;24:1–11.
25. Zyphur MJ, Zhang Z, Barsky AP, Li W-D. An ACE in the hole: twin family models for applied behavioral genetics research. Leadersh Q. 2013;24:572–94.
26. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era — concepts and misconceptions. Nat Rev Genet. 2008;9:255–66.
27. Gordon L, et al. Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. Genome Res. 2012;22:1395–406.
28. Bell JT, et al. Epigenome-Wide scans identify differentially methylated regions for age and age-Related phenotypes in a healthy ageing population. PLoS Genet. 2012;8:e1002629.
29. Grundberg E, et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to Disease-Associated variants in distal regulatory elements. Am J Hum Genet. 2013;93:876–90.
30. van Dongen J, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat Commun. 2016;7:11115.
31. Hannon E, et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. PLoS Genet. 2018;14:e1007544.
32. Bell JT, Saffery R. The value of twins in epigenetic epidemiology. Int J Epidemiol. 2012;41:140–50.
33. Czamara D, et al. Integrated analysis of environmental and genetic influences on cord blood DNA methylation in newborns. Nat Commun. 2019;10:2548.
34. Teh AL, et al. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. Genome Res. 2014;24:1064–74.
35. Chatterjee S, Ouidir M, Tekola-Ayele F. Genetic and *in utero* environmental contributions to DNA methylation variation in placenta. Hum Mol Genet. 2021;30:1968–76.
36. Takaro TK, et al. The Canadian healthy infant longitudinal development (CHILD) birth cohort study: assessment of environmental exposures. J Expo Sci Environ Epidemiol. 2015;25:580–92.
37. Moraes TJ, et al. The Canadian healthy infant longitudinal development birth cohort study: biological samples and biobanking: the CHILD study: biological samples. Paediatr Perinat Epidemiol. 2015;29:84–92.
38. Subbarao P, et al. The Canadian healthy infant longitudinal development (CHILD) study: examining developmental origins of allergy and asthma: table 1. Thorax. 2015;70:998–1000.
39. Barker M, et al. Introducing the FAIR principles for research software. Sci Data. 2022;9:622.
40. Eckhardt F, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38:1378–85.
41. Hui T, et al. High-Resolution Single-Cell DNA methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. Stem Cell Rep. 2018;11:578–92.
42. Ong M-L, Holbrook JD. Novel region discovery method for infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways. Aging Cell. 2014;13:142–55.
43. Pidsley R, et al. Critical evaluation of the illumina methylationepic BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17:208.
44. Derakhshan M, et al. Tissue- and ethnicity-independent hypervariable DNA methylation States show evidence of establishment in the early human embryo. Nucleic Acids Res. 2022;50:6735–52.

45. Andrews SV, Ladd-Acosta C, Feinberg AP, Hansen KD, Fallin MD. Gap hunting to characterize clustered probe signals in illumina methylation array data. Epigenetics Chromatin. 2016;9:56.
46. Sugden K, et al. Patterns of reliability: assessing the reproducibility and integrity of DNA methylation measurement. Patterns. 2020;1:100014.
47. Welsh H, et al. A systematic evaluation of normalization methods and probe replicability using infinium EPIC methylation data. Clin Epigenet. 2023;15:41.
48. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30:2811–2.
49. Bühlmann P, van de Geer S. Statistics for High-Dimensional data: Methods, theory and applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. https://doi.org/10.1007/978-3-642-20192-9.
50. Du P, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.
51. Garg P, Joshi RS, Watson C, Sharp AJ. A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. PLoS Genet. 2018;14:e1007707.
52. Min JL, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat Genet. 2021;53:1311–21.
53. McRae AF, et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol. 2014;15:R73.
54. Nustad HE, Page CM, Reiner AH, Zucknick M, LeBlanc M. A bayesian mixed modeling approach for estimating heritability. BMC Proc. 2018;12:31.
55. Huan T, et al. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Nat Commun. 2019;10:4267.
56. Day K, et al. Heritable DNA methylation in CD4 + Cells among complex families displays genetic and Non-Genetic effects. PLoS ONE. 2016;11:e0165488.
57. Tarrade A, Panchenko P, Junien C, Gabory A. Placental contribution to nutritional programming of health and diseases: epigenetics and sexual dimorphism. J Exp Biol. 2015;218:50–8.
58. Lapehn S, Paquette AG. The placental epigenome as a molecular link between prenatal exposures and fetal health outcomes through the dohad hypothesis. Curr Envir Health Rpt. 2022;9:490–501.
59. Wild CP. The exposome: from concept to utility. Int J Epidemiol. 2012;41:24–32.
60. Siroux V, Agier L, Slama R. The exposome concept: a challenge and a potential driver for environmental health research. Eur Respir Rev. 2016;25:124–9.
61. Vineis P, et al. What is new in the exposome? Environ Int. 2020;143:105887.
62. Santos S, et al. Applying the exposome concept in birth cohort research: a review of statistical approaches. Eur J Epidemiol. 2020;35:193–204.
63. Mansell G, et al. Guidance for DNA methylation studies: statistical insights from the illumina EPIC array. BMC Genomics. 2019;20:366.
64. Biwer C, Kawam B, Chapelle V, Silvestre F. The role of stochasticity in the origin of epigenetic variation in animal populations. Integr Comp Biol. 2020;60:1544–57.
65. McAdams TA, Cheesman R, Ahmadzadeh YI. Annual research review: towards a deeper Understanding of nature and nurture: combining family-based quasi-experimental methods with genomic data. Child Psychol Psychiatry. 2023;64:693–707.
66. Heyn H, et al. DNA methylation contributes to natural human variation. Genome Res. 2013;23:1363–72.
67. Pappalardo XG, Barra V. Losing DNA methylation at repetitive elements and breaking bad. Epigenetics Chromatin. 2021;14:25.
68. Law P-P, Holland ML. DNA methylation at the crossroads of gene and environment interactions. Essays Biochem. 2019;63:717–26.
69. R Core Team. R: A Language and environment for statistical computing. R Foundation for Statistical Computing; 2021.
70. Wickham H, et al. Welcome to the tidyverse. JOSS. 2019;4:1686.
71. Girchenko P, et al. Prediction and prevention of preeclampsia and intrauterine growth restriction (PREDO) study. Int J Epidemiol. 2016;dyw154. https://doi.org/10.1093/ije/dyw154.
72. Heiss JA, Just AC. Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. Clin Epigenet. 2018;10:73.
73. Aryee MJ, et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.
74. Morin AM, et al. Maternal blood contamination of collected cord blood can be identified using DNA methylation at three CpGs. Clin Epigenet. 2017;9:75.
75. Du P, Kibbe WA, Lin S. M. *lumi* : a pipeline for processing illumina microarray. Bioinformatics. 2008;24:1547–8.
76. Pidsley R, et al. A data-driven approach to preprocessing illumina 450K methylation array data. BMC Genomics. 2013;14:293.
77. Haftorn KL, et al. An EPIC predictor of gestational age and its application to newborns conceived by assisted reproductive technologies. Clin Epigenet. 2021;13:82.
78. Salas LA, et al. An optimized library for reference-based Deconvolution of whole-blood biospecimens assayed using the illumina humanmethylationepic BeadArray. Genome Biol. 2018;19:64.
79. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The Sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3.
80. Fortin J-P, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014;15:503.
81. Illumina Inc. Infinium® Genotyping Data Analysis. (2014).
82. Das S, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.
83. Danecek P, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008. https://doi.org/10.1093/gigascience/giab008.
84. Chang CC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaSci. 2015;4:7.
85. Buuren SV, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. J Stat Soft. 2011;43(3):1-67.

86. Filzmoser P, Hron K, Reimann C. Principal component analysis for compositional data with outliers. Environmetrics. 2009;20:621–32.
87. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
88. Gatev E, Gladish N, Mostafavi S, Kobor MS. CoMeBack: DNA methylation array data analysis for co-methylated regions. Bioinformatics. 2020;36:2675–83.
89. Chakrabarti A, Ghosh JK, AIC. BIC and recent advances in model selection. Philosophy of statistics. Elsevier; 2011. pp. 583–605. https://doi.org/10.1016/B978-0-444-51862-0.50018-6.
90. Grömping U. Relative importance for linear regression in R: the package relaimpo. J Stat Soft. 2006;17(1):1-27.
91. Lindeman RH, Merenda PF, Gold RZ. Introduction to bivariate and multivariate analysis. Glenview, IL: Scott, Foresman & Co; 1980.
92. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353–8.
93. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
94. Wu T, et al. ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov. 2021;2:100141.
95. Lee S, Cook D, Lawrence M. Plyranges: a grammar of genomic data transformation. Genome Biol. 2019;20:4.
96. Navarro-Delgado EI. RAMEN: regional association of DNA methylome variability with exposome and genome. Zenodo. 2025. https://doi.org/10.5281/ZENODO.15171547.
97. Navarro-Delgado EI. NavarroDelgado_2025_VMRs. 2025. github.com/ErickNavarroD/NavarroDelgado_2025_VMRs.
98. Zheng X, et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28:3326–8.
99. The Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) Consortium. ENIGMA 1KGP_p3v5 Cookbook. http://enigma.ini.usc.edu/genetics-protocols/ (2017).
100. Radloff LS, The CES-D, Scale. A Self-Report depression scale for research in the general population. Appl Psychol Meas. 1977;1:385–401.
101. Cohen S, Williamson G. Perceived stress in a probability sample of the United States. in *The Social Psychology of Health* 31–67 (Sage, Newbury Park, CA).
102. Wheaton B. Sampling the stress universe. In: Avison WR, Gotlib IH, editors. Stress and mental health. Boston, MA: Springer US; 1994. pp. 77–114. https://doi.org/10.1007/978-1-4899-1106-3_4.
103. Kurdek LA. Dimensionality of the dyadic adjustment scale: evidence from heterosexual and homosexual couples. J Fam Psychol. 1992;6:22–35.
104. Cohen S, Mermelstein R, Kamarck T, Hoberman HM. Measuring the functional components of social support. Social support: theory, research and applications. The Hage, Netherlands: Springer; 1985.
105. Turner RJ, Wheaton B, Lloyd DA. The epidemiology of social stress. Am Sociol Rev. 1995;60:104.
106. Diemer MA, Mistry RS, Wadsworth ME, López I, Reimers F. Best practices in conceptualizing and measuring social class in psychological research: social class measurement. Analyses Social Issues Public Policy. 2013;13:77–113.
107. Arntzen A, Magnus P, Bakketeig LS. Different effects of maternal and paternal education on early mortality in Norway. Paediatr Perinat Epidemiol. 1993;7:376–86.
108. Jeong J, Kim R, Subramanian SV. How consistent are associations between maternal and paternal education and child growth and development outcomes across 39 low-income and middle-income countries? J Epidemiol Community Health. 2018;72:434–41.
109. Rammohan A, Awofeso N, Fernandez RC. Paternal education status significantly influences infants' measles vaccination uptake, independent of maternal education status. BMC Public Health. 2012;12:336.
110. Fred Hutchinson Cancer Research Center. Food frequency questionnaires (FFQ). (2014).
111. Schakel SF. Maintaining a nutrient database in a changing marketplace: keeping Pace with changing food Products—A research perspective. J Food Compos Anal. 2001;14:315–22.
112. Guenther PM, et al. Update of the healthy eating index: HEI-2010. J Acad Nutr Dietetics. 2013;113:569–80.
113. for the NutriGen Alliance Investigators. Harmonization of Food-Frequency questionnaires and dietary pattern analysis in 4 ethnically diverse birth cohorts. J Nutr. 2016;146:2343–50.
114. Gulacha P, de Souza RJ. Maternal diet and asthma in children. Ontario, Canada: McMaster University; 2021.
115. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Soft. 2010;33(1):1-22.

Publisher's Note