PACE / NutriPROGRAM Analysis Plan
Maternal vegetarian/plant-based diets and cord blood DNA methylation

Date: August 2024
Version: v5.3
Authors: Peiyuan Huang, Maria Carolina Borges, Kate Northstone, Gemma Sharp
Contacts: Peiyuan Huang (peiyuan.huang@bristol.ac.uk), cc Gemma Sharp (g.c.sharp@exeter.ac.uk)

Background

Vegetarianism is becoming increasingly popular worldwide[1]. Following vegetarian/plant-based diets during pregnancy has been associated with lower birth weight and small for gestational age offspring[2-4]. Although inconclusive, some evidence also suggests an association with other neonatal outcomes, such as neural tube defects and hypospadias[2 5]. However, the molecular mechanisms underlying such associations are not well understood.

DNA methylation (DNAm) has been found to be related to fetal and neonatal health outcomes[6], and some key nutrient intakes (e.g., iron, folates, and omega-3 fatty acids) and other nutrition-related factors (e.g., body mass index [BMI]) associated with DNAm[6-10] tend to differ between vegetarians and non-vegetarians[11]. Therefore, we hypothesise that changes in DNAm might be a possible mechanism between maternal vegetarian/plant-based diets and offspring health.

Aim

This epigenome-wide association study (EWAS) aims to examine cord blood DNAm in relation to maternal adherence to vegetarian and plant-based diets during pregnancy. In addition to single cytosine-phosphate-guanine (CpG) site analyses, we will also analyse differentially methylated regions (DMRs), which can be more statistically powerful and potentially more biologically relevant.

Eligibility

Your study is eligible if it fits into BOTH criteria below:

- It has epigenome-wide DNAm data measured in cord blood.
- It has maternal dietary data collected during pregnancy (e.g., from food frequency questionnaires [FFQ], 24-hour dietary recall [24HR], or food diaries). Eligible studies will be further categorised into two tiers:
    - **Tier 1:** With dietary data that can be used to derive both weekly intake frequencies (as times/week) and daily intakes (as grams/day) of the 18 food groups listed in **Table 1** below (see detailed information about each food group in **Supplementary Table 1**). Tier 1 studies will participate in the analyses for both vegetarianism and plant-based diet indices (see sections below).
    - **Tier 2:** With dietary data that can be used to derive weekly intake frequencies (as times/week) of 4 key food groups: meat, fish/seafood, egg, and dairy (**Table 1**). Tier 2 studies will participate in the analyses for vegetarianism only.

Please note:

- The exposures of interest used in this project MUST be derived from dietary data; studies with only self-defined vegetarianism data (typically collected using a single question like "Are you a vegetarian/vegan?" or "Are you following a vegetarian/vegan diet?") are NOT eligible.
- Tier 2 analyses, with looser inclusion criteria, are for those studies using non-quantitative FFQs or with one or more of the 18 food groups missing. Studies will participate in EITHER tier 1 OR tier 2 analyses based on their eligibility.
- We assume that studies with dietary data in grams/day would have available data in times/week as well. Please get in contact if this is NOT the case in your study.
- For tier 2 studies, although the necessary food groups are only meat and fish/seafood because different subgroups of vegetarianism will be combined for the final analysis (see sections below), we still wish to derive detailed subgroups and describe their distribution. If your study has data on meat and fish/seafood intakes but not egg or dairy intakes, please get in touch with us as it may still be eligible.

Table 1. Food groups for defining vegetarian subgroups[3 12] and calculating plant-based diet indices[13].

Food groups (N = 18 in total)	Vegetarian subgroups				Plant-based diet indices		
	Non-vegetarian	Pesco-vegetarian	Full vegetarian				
			Lacto-ovo-vegetarian	Vegan	PDI	hPDI	uPDI
Healthy plant food groups (N = 7)							
Whole grains	No restriction	No restriction	No restriction	No restriction	Positive scores[d]	Positive scores	Reverse scores[e]
Fruits	No restriction	No restriction	No restriction	No restriction	Positive scores	Positive scores	Reverse scores
Vegetables	No restriction	No restriction	No restriction	No restriction	Positive scores	Positive scores	Reverse scores
Nuts	No restriction	No restriction	No restriction	No restriction	Positive scores	Positive scores	Reverse scores
Legumes	No restriction	No restriction	No restriction	No restriction	Positive scores	Positive scores	Reverse scores
Vegetable oils[a]	No restriction	No restriction	No restriction	No restriction	Positive scores	Positive scores	Reverse scores
Tea and coffee	No restriction	No restriction	No restriction	No restriction	Positive scores	Positive scores	Reverse scores
Less healthy plant food groups (N = 5)							
Fruit juices	No restriction	No restriction	No restriction	No restriction	Positive scores	Reverse scores	Positive scores
Refined grains	No restriction	No restriction	No restriction	No restriction	Positive scores	Reverse scores	Positive scores
Potatoes	No restriction	No restriction	No restriction	No restriction	Positive scores	Reverse scores	Positive scores
Sugar-sweetened beverages	No restriction	No restriction	No restriction	No restriction	Positive scores	Reverse scores	Positive scores
Sweets and desserts	No restriction	No restriction	No restriction	No restriction	Positive scores	Reverse scores	Positive scores
Animal food groups (N = 6)							
Animal fat[b]	-	-	-	-	Reverse scores	Reverse scores	Reverse scores
Dairy	No restriction	No restriction	≥1 time/mo	<1 time/mo	Reverse scores	Reverse scores	Reverse scores
Eggs	No restriction	No restriction			Reverse scores	Reverse scores	Reverse scores
Fish and seafood	No restriction	≥1 time/mo	<1 time/mo		Reverse scores	Reverse scores	Reverse scores
Meat	No restriction	<1 time/mo			Reverse scores	Reverse scores	Reverse scores
Miscellaneous animal-based foods[c]	-	-	-	-	Reverse scores	Reverse scores	Reverse scores

[a] Defined as oils from plant sources that are liquid in room temperature.

[b] Defined as fats from animal sources that are solid/semi-solid in room temperature; not considered in the definition of vegetarian subgroups.

[c] Defined as foods containing animal-based content that cannot be classified into other animal food groups; not considered in the definition of vegetarian subgroups.

[d] Scores 1, 2, and 3 are assigned to the 1st, 2nd, and 3rd tertile of intake, respectively.

[e] Scores 3, 2, and 1 are assigned to the 1st, 2nd, and 3rd tertile of intake, respectively.

PDI, overall plant-based diet index; hPDI, healthful plant-based diet index; uPDI, unhealthful plant-based diet index.

Exclusions

- Please exclude twins and siblings (i.e., conduct a singleton-only analysis and include one child per family in case of multiple siblings from the same family).
- We do NOT require excluding participants of ethnicities other than European ancestry. If your study is multi-ethnic, please analyse different ethnicities separately (the suggested minimum sample size per ethnic group is 50). Please get in contact with us if you think the number of cases is too low in any ethnic group.

Adherence to vegetarian/plant-based dies

Our exposures of interest are maternal adherence to vegetarian and plant-based dies during pregnancy, including the following five variables:

- Two binary variables for vegetarianism – "veggie1" (i.e., full vegetarians vs. non-vegetarians) and "veggie2" (i.e., full vegetarians + pesco-vegetarians vs. non-vegetarians)[3 12]
- Three continuous variables for the plant-based diet index (PDI) – overall PDI, healthful PDI (hPDI), and unhealthful PDI (uPDI)[13]. **Table 1** above summarises the definition of various vegetarian subgroups and the calculation of different versions of PDIs. More details can be found in **Supplementary Note 1**.

If your study has maternal dietary data measured at multiple time points throughout pregnancy, please get in touch.

Models

We will run linear regression models (from the "limma" R package) for our dietary exposures and neonatal DNAm outcomes.

Main analysis

There will be four models for each of the five exposures of interest:

- "veggie1": full vegetarians vs. non-vegetarians (binary, with non-vegetarians as the reference group)
- "veggie2": full vegetarians + pesco-vegetarians vs. non-vegetarians (binary, with non-vegetarians as the reference group)
- PDI (continuous, range 18 to 54)
- hPDI (continuous, range 18 to 54)
- uPDI (continuous, range 18 to 54)

Please note:

- Tier 2 studies will only participate in the analyses for "veggie1" and "veggie2".
- Please get in touch if the number of cases in "veggie1" or "veggie2" is too low (e.g., <10).

Table 2 below presents a summary of the modelling process. Please adjust for as many covariates specified below as possible in each model and record if any of them are not adjusted for as required and why. The main model of interest will be the fully adjusted one. The "no-cells" and minimally adjusted models are being run because they are common practice and can help understand the effect of adjustments. The rationale for setting up the additional model is that nutrition-related factors (e.g., BMI, energy intake, dietary supplement use) may mediate the effect of diet on health outcomes, and this model can provide insights into the potential role of these factors. To avoid potential collider bias, we will NOT consider gestational age at birth as a covariate because it is likely to be influenced by both maternal diet and other adjusted or unadjusted prenatal factors.

Table 2. Modelling for the association between maternal adherence to vegetarian/plant-based diets during pregnancy and cord blood DNA methylation.

	Maternal dietary exposures of interest				
	"veggie1"	**"veggie2"**	**PDI**	**hPDI**	**uPDI**
"No-cells" model Adjusted for child sex and batch (as surrogate variables) only	NoCellModel.veggie1	NoCellModel.veggie2	NoCellModel.PDI	NoCellModel.hPDI	NoCellModel.uPDI
Minimally adjusted model Additionally adjusted for blood cell types	MinModel.veggie1	MinModel.veggie2	MinModel.PDI	MinModel.hPDI	MinModel.uPDI
Fully adjusted model Additionally adjusted for maternal age, education attainment, parity, and maternal smoking during pregnancy	FullModel.veggie1	FullModel.veggie2	FullModel.PDI	FullModel.hPDI	FullModel.uPDI
Additional model Additionally adjusted for maternal pre-pregnancy BMI, total energy intake, and dietary supplementation during pregnancy.	AddModel.veggie1	AddModel.veggie2	AddModel.PDI	AddModel.hPDI	AddModel.uPDI

"veggie1" (full vegetarians vs. non-vegetarians) and "veggie2" (full vegetarians + pesco-vegetarians vs. non-vegetarians) are as binary variables, with non-vegetarians as the reference group.

PDI, hPDI, and uPDI are as continuous variables, ranging from 18 to 54.

PDI, overall plant-based diet index; hPDI, healthful plant-based diet index; uPDI, unhealthful plant-based diet index.

Sensitivity analyses
We will also conduct ethnic-specific analyses. We will re-run the meta-analysis in participants of European ancestry only. So please analyse different ethnicities separately if your study is multi-ethnic.

R Code

This analysis plan is accompanied by an R script (named as "**MatVegDiet_PACE_EWAS.r**"), which contains the functions necessary for preparing data and running the analyses. You will need to edit the code to tailor them to your study. In the script, sections that need to be edited are highlighted (see the "**[PLEASE CHANGE]**" notice). Before running the script, your phenotype data need to be set up correctly, with the correct variable column names, as explained in this analysis plan and the R script. The script also contains the code for deriving dietary exposure variables (i.e., vegetarian subgroups and PDIs).

If you have any questions or difficulties running the code or think you need to make extra changes to the script to suit your dataset, please do not hesitate to contact Peiyuan Huang (peiyuan.huang@bristol.ac.uk) and cc Gemma Sharp (g.c.sharp@exeter.ac.uk).

Inputs

In this study, maternal vegetarian/plant-based diets during pregnancy are always used as the exposure and DNAm in cord blood as the outcome. Before running the code, analysts need to prepare the following inputs:

- **A single phenotype file (including cell type data)** – containing a column called "sample.id" (used to match to the methylation matrix) and all phenotype data (including the dietary variables, covariates, and cell types). It is very important that variables in your phenotype data are named according to the guidelines set out in this document and R script. Please note that this will be a complete case analysis, so food intake tertiles for PDIs should be calculated in the final subset (i.e., those with non-missing data in exposure, outcome, and all covariates). The script will subset the samples to the complete cases before tertile calculation.
- **A methylation matrix** – with column names matched to "sample.id" in the phenotype file.

For the dietary variables, we will need the daily intake (in g/day) of each of the 18 food groups for deriving PDIs and the weekly intake frequency (in times/week) of each of the 4 key animal food groups (i.e., meat, fish/seafood, egg, and dairy) for classifying vegetarian subgroups.

Preparation of methylation data

This analysis uses methylation data (Illumina beta-values) from cord blood measured using either the EPIC or 450K Illumina arrays. Please prepare methylation matrices as follows:

- Normalise the beta-values using your preferred method and mention the method you chose in the attached "cohort info" Excel file.
- Please do NOT convert to M-values (i.e., logit transformation of beta-values).
- You can use your preferred study QC settings for probe filtering (e.g., removing probes with a high detection p-value, probes on sex chromosomes, and probes used as controls).

- However, if possible, please do NOT exclude probes identified as polymorphic sites based on the lists provided by Chen et al.[14] or Naeem et al.[15] (or similar). We will do this at the meta-analysis stage.
- The R script includes the code to remove outliers using the IQR*3 (Tukey) method. Please run these lines or use methylation data with outliers already removed using this method.
- The script includes the code to match your phenotype file to your methylation matrix, so you do NOT have to do this beforehand. However, the columns of your methylation matrix MUST be named according to a unique identifier, which is also used in your phenotype data under the column “sample.id”. If the methylation and the phenotype IDs in your study are not the same and you do not know how to set this up, please get in touch.

Batch
Please do NOT include a batch variable. The code provided calculates surrogate variables to adjust for technical variation. Please do NOT use ComBat or any other methods to adjust your beta matrix for batch before running the EWAS.

Cell types
Cord blood cell type estimation: Use the “Salas” reference set for cell type estimation in the ‘’FlowSorted.CordBlood.Combined.450K’’ or “FlowSorted.CordBlood.Combined.EPIC’’ Bioconductor package for cell type correction. These packages include the following cell types: “CD8T”, “CD4T”, “NK”, “Bcell”, “Mono”, “Gran”, and “nRBC”.

Covariates
A complete list of covariates and variable names (as used in the phenotype file and R script) is provided in **Table 3** below.

Table 3. List of covariates and their variable names.

Variables	Values	Variable name

Child sex	Male = 0
	Female = 1
		sex

Maternal age at delivery (or conception)	Continuous (years)	mat.age

Highest level of maternal education at the time of delivery (or shortly after) – Please use your preferred definition in your study.	Lower = 0
	Higher = 1
		mat.edu

Mother's parity (number of previous pregnancies)	No previous delivery = 0 At least 1 previous delivery = 1	mat.parity
Sustained maternal smoking during pregnancy	No smoking or quitted in the first trimester = 0 Sustained smoking throughout pregnancy = 1	mat.smoking
Maternal pre-pregnancy BMI (pre-pregnancy is preferred, but BMI at the early stage of pregnancy is also fine)	Continuous (kg/m^2)	mat.bmi
Maternal total energy intake during pregnancy[a]	Continuous (kcal/day)	mat.kcal
Maternal dietary supplementation during pregnancy[b]	No supplement use at any time points during pregnancy = 0 At least 1 type of supplement used at any time point during pregnancy (preferably occurring before or at the same time as dietary assessment) = 1	mat.suppl

[a] Estimated from maternal dietary data during pregnancy.

[b] Based on available information in each study; can be any type of supplement (e.g., in ALSPAC, this includes iron, zinc, calcium, folic acid, multivitamins, and "other"; omega-3 fatty acids may also be considered if available in your study).

Output format (how to share the results with us)
The code provided will generate the output files we want and name them accordingly. Please double-check that the initial parameters are set up as follows:

- [PROJECT]: Project name, which is **MatVegDiet** for this project
- [STUDY]: Your study name (e.g., **ALSPAC**)
- [EXPOSURE]: **veggie1** / **veggie2** / **PDI** / **hPDI** / **uPDI**
- [MODEL]: **NoCellModel** / **MinModel** / **FullModel** / **AddModel**
- [TIMEPOINT]: The time point at which methylation data were measured, which is **birth** for this project
- [DATE]: The date when the analysis is performed (in the format "YYYYMMDD", e.g., **20221130**); the R code will generate this automatically

Please supply the following output files in the specified formats (these are automatically created using the supplied R script):

- Rdata files (5 for tier 1 studies, 2 for tier 2 studies) for EWAS results, each containing outputs from all four EWAS models for one exposure variable, named as: "[PROJECT].[STUDY].[EXPOSURE].EWASres.[TIMEPOINT].[DATE].Rdata" (e.g., "MatVegDiet.ALSPAC.veggie1.EWASres.birth.20221130.Rdata").
- One Rdata files for the log of the IQR*3 (Tukey) method, named as: "[PROJECT].[STUDY].logIQR.[TIMEPOINT].[DATE].Rdata".
- One csv file showing the distribution of the 18 food groups, named as: "[PROJECT].[STUDY].food.group.distribute.[TIMEPOINT].[DATE].csv". Not applicable for tier 2 studies.
- One csv file showing the tertile cutoffs of the 18 food groups, named as: "[PROJECT].[STUDY].food.group.tertile.[TIMEPOINT].[DATE].csv". Not applicable for tier 2 studies.
- One csv file showing summary statistics of phenotype data, named as: "[PROJECT].[STUDY].tableone.[TIMEPOINT].[DATE].csv".
- One csv file showing the intake of each available food group by vegetarian subgroups, named as: "[PROJECT].[STUDY].ByVegDiet.[TIMEPOINT].[DATE].csv".
- One correlogram showing the correlation matrix between PDIs and intakes of the 18 food groups, named as: "[PROJECT].[STUDY].correlogram.[TIMEPOINT].[DATE].png". Not applicable for tier 2 studies.
- Three histograms for PDI, hPDI, and uPDI, named as: "[PROJECT].[STUDY].[EXPOSURE].histogram.[TIMEPOINT].[DATE].jpg". Not applicable for tier 2 studies.
- Q-Q plots (20 for tier 1 studies, 8 for tier 2 studies) – Separate plots will be created for each exposure variable and model, named as: "[PROJECT].[STUDY].[MODEL].[EXPOSURE].QQ.[TIMEPOINT].[DATE].jpg".
- One csv file showing lambdas for all EWAS models, named as: "[PROJECT].[STUDY].lambda.[TIMEPOINT].[DATE].csv".
- csv files (20 for tier 1 studies, 8 for tier 2 studies) showing the results from DMR analysis (based on the fully adjusted model) for each exposure variable, named as: "[PROJECT].[STUDY].[EXPOSURE].dmr.[TIMEPOINT].[DATE].csv".

In addition, please also complete the attached Excel file named as "MatVegDiet.[STUDY].cohortinfo.xlsx" and upload it along with other output files. Please note that there are two tabs to fill within the file. This file will ask for general information about your study and analyses. Example answers from ALSPAC are provided in the file.

When you are ready to upload your results, please zip them and email Peiyuan Huang (peiyuan.huang@bristol.ac.uk) and cc Gemma Sharp (g.c.sharp@exeter.ac.uk), and we will provide you with a personal URL for upload.

References

1. Leahy E, Lyons S, Tol R. An Estimate of the Number of Vegetarians in the World. Dublin: The Economic and Social Research Institute (ESRI), 2010.
2. Tan C, Zhao Y, Wang S. Is a vegetarian diet safe to follow during pregnancy? A systematic review and meta-analysis of observational studies. Critical Reviews in Food Science and Nutrition: Taylor and Francis Inc., 2019:2586-96.
3. Yisahak SF, Hinkle SN, Mumford SL, et al. Vegetarian diets during pregnancy, and maternal and neonatal outcomes. International Journal of Epidemiology: Oxford Academic, 2021:165-78.
4. Kesary Y, Avital K, Hiersch L. Maternal plant-based diet during gestation and pregnancy outcomes. Arch Gynecol Obstet: Springer, 2020:887-98.
5. Sebastiani G, Barbero AH, Borrás-Novell C, et al. The Effects of Vegetarian and Vegan Diet during Pregnancy on the Health of Mothers and Offspring. Nutrients 2019, Vol 11, Page 557: Multidisciplinary Digital Publishing Institute, 2019:557.
6. McGee M, Bainbridge S, Fontaine-Bisson B. A crucial role for maternal dietary methyl donor intake in epigenetic programming and fetal growth outcomes. Nutr Rev: Oxford Academic, 2018:469-78.
7. Taeubert MJ, de Prado-Bert P, Geurtsen ML, et al. Maternal iron status in early pregnancy and DNA methylation in offspring: an epigenome-wide meta-analysis. Clin Epigenetics: BioMed Central, 2022:1-12.
8. Joubert BR, den Dekker HT, Felix JF, et al. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nat Commun* 2016;7:10577. doi: 10.1038/ncomms10577 [published Online First: 20160210]
9. Tremblay BL, Guénard F, Rudkowska I, et al. Epigenetic changes in blood leukocytes following an omega-3 fatty acid supplementation. *Clin Epigenetics* 2017;9:43. doi: 10.1186/s13148-017-0345-3 [published Online First: 20170426]
10. Sharp GC, Lawlor DA, Richmond RC, et al. Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: Findings from the Avon Longitudinal Study of Parents and Children. International Journal of Epidemiology: Oxford University Press, 2015:1288-304.
11. Oussalah A, Levy J, Berthezène C, et al. Health outcomes associated with vegetarian diets: An umbrella review of systematic reviews and meta-analyses. Clin Nutr: Churchill Livingstone, 2020:3283-307.
12. Jaacks LM, Kapoor D, Singh K, et al. Vegetarianism and cardiometabolic disease risk factors: Differences between South Asian and US adults. *Nutrition* 2016;32(9):975-84. doi: 10.1016/j.nut.2016.02.011 [published Online First: 20160304]

13. Satija A, Bhupathiraju SN, Rimm EB, et al. Plant-Based Dietary Patterns and Incidence of Type 2 Diabetes in US Men and Women: Results from Three Prospective Cohort Studies. PLoS Med: Public Library of Science, 2016:e1002039.

14. Chen Y-A, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 203 Epigenetics, 2013:203-09.

15. Naeem H, Wong NC, Chatterton Z, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. BMC Genomics: BioMed Central Ltd., 2014.