**Maternal Caffeine EWAS Analysis Plan**

**Stage 1: Identify cord blood CpGs associated with maternal caffeine intake during pregnancy**

**Participating cohorts:** MoBa, ALSPAC, BiB, GOYA

**Rationale:** Animal studies have provided some evidence that maternal caffeine consumption can influence offspring DNA methylation (PMIDs: 22970234, 24475304, 25354728, 25868845, 25868845), but what about humans?

**Aim:** To carry out an exploratory EWAS meta-analysis to identify associations between maternal caffeine intake in pregnancy and cord blood DNA methylation.

**Follow-on work:** If we find interesting positive associations, we can investigate these further. For example, we can look at the correlation in methylation in cord blood and other tissues, and enrichment for regions associated with other maternal exposures like smoking. We can also look at the influence of timing (before, early, mid, late or after pregnancy) and source (tea, coffee, fizzy drinks, chocolate, etc) of exposure, the stability of associations over time (in longitudinal data from ALSPAC) and whether the associations are indicative of causal effects. To infer causality, we can compare effect estimates for maternal prenatal, postnatal and paternal caffeine. We can also use SNPs identified in a recent caffeine GWAS to infer causal relationships using Mendelian Randomisation.

**EWAS analysis protocol**

**Exposure**

**mat.caff:** Maternal caffeine intake during pregnancy in mg/day. Please double check caffeine values >= +/- 5 SD from the mean in your dataset to make sure they are not data entry errors. If you have multiple time points, please use the time point closest to 22 weeks (the time point available for MoBa).

**Outcome**

Illumina Infinium 450k or EPIC BeadChip DNA methylation data in cord blood or neonatal blood spots collected at birth.

The methylation data should be normalised beta values on a scale of 0 to 1 with no transformation (e.g. not M values). You can use your preferred method to normalise the data, but our preference is for Functional Normalisation. Please contact [gemma.sharp@bristol.ac.uk](https://github.com/ammegandchips/PACE_Paternal_BMI/blob/master/gemma.sharp@bristol.ac.uk) if you would like R code to conduct functional normalisation on your data.

Outliers should be trimmed using the IQR3 (Tukey) method. The code for doing this is provided.

Please use your preferred study QC settings for probe filtering. However, please do not exclude probes just because they are on a published list of possibly problematic probes (e.g. Chen or Naeem) and please do not exclude probes on the sex chromosomes. If in any doubt, please include rather than exclude probes at this stage.

**Other variables**

It is very important that covariates are coded exactly as outlined below. The R code relies on these codings!

* **Maternal social class (mat.ses):** 4 category ordinal variable treated as numeric (i.e. not categorical/factor) please use your preferred classification but note that our preference is for education level.
* **Maternal age (mat.age):** continuous numeric variable in years
* **Maternal BMI (mat.bmi):** continuous in kg/m2
* **Maternal smoking status around pregnancy (mat.smoking):** binary numeric variable (1=smoking throughout pregnancy or 0=no smoking during pregnancy OR smoking in the first trimester then giving up)
* **Parity (parity):** binary numeric variable (1=one or more previous children/0=no previous children)
* **Gestational age at birth (gest.age):** Continuous numeric variable in days
* **Surrogate variables (SVs) to adjust for batch:** please do NOT include a known batch variable in your models or adjust for batch using another method such as ComBat. The code for calculating surrogate variables is encorporated in the EWAS code provided. We hope (with some support for this hope from the literature and personal experience) that using this approach in all cohorts will reduce heterogeneity and lambdas.
* **Estimated cell proportions:** Cell proportions are estimated using the Houseman method (e.g. by using the estimateCellCounts() function in minfi) and the Gervin et al. (2016) cord blood cell count reference.
* **Ethnicity:** If your study has more than one major ethnic group (for example, European ancestry, Latino, African Ancestry, Asian), please analyse them separately.
* **Child's sex (sex):** Binary numeric variable. This will be used to stratify analyses (1=females,0=males).

**Exclusions**

Please also exclude multiple pregnancies (e.g. twins) and siblings (i.e. each mother/father should appear only once in the dataset)

**EWAS models**

* **Crude model: Maternal caffeine (model name: crude)**
  + Methylation ~ Maternal caffeine + SVs for batch
* **Cell-adjusted: Maternal caffeine + Cells (model name: cells)**
  + Methylation ~ Maternal caffeine + SVs for batch + Estimated cell counts
* **Covariate-adjusted: Maternal caffeine + Covariates (model name: covs)**
  + Methylation ~ Maternal caffeine + SVs for batch + mat.ses + mat.age + mat.smoking + parity + mat.bmi
* **Fully-adjusted: Maternal caffeine + Covariates + Cells (model name: full)**
  + Methylation ~ Maternal caffeine + SVs for batch + Estimated cell counts + mat.ses + mat.age + mat.smoking + parity + mat.bmi
* **Gestational age-adjusted: Maternal caffeine + gestational age + covariates + Cells (model name: gest)**
  + Methylation ~ Maternal caffeine + SVs for batch + Estimated cell counts + mat.ses + mat.age + mat.smoking + parity + mat.bmi + gest.age
* **Males only: Maternal caffeine + Covariates + Cells (model name: males)**
  + Methylation ~ Maternal caffeine + SVs for batch + Estimated cell counts + mat.ses + mat.age + mat.smoking + parity + mat.bmi
  + *The R code will remove female offspring*
* **Females only: Maternal caffeine + Covariates + Cells (model name: females)**
  + Methylation ~ Maternal caffeine + SVs for batch + Estimated cell counts + mat.ses + mat.age + mat.smoking + parity + mat.bmi
  + *The R code will remove male offspring*

**Outputs**

Please supply the following files in the specified formats:

* EWAS results: one Rdata file labelled as YOURSTUDY.matcaff.ewasresults.timepoint.Rdata.
* Extra cohort information: one Excel file labelled as YOURSTUDY.matcaff.cohortinfo.xlsx. This will contain information for each cohort relating to things like normalisation method. Please download and use the template available at: [https://github.com/LauraSchellhas/mat\_caffeine\_EWAS.git](https://github.com/LauraSchellhas/mat_caffeine_EWAS)
* EWAS variables summary: a csv file labelled as YOURSTUDY.matcaff.modelname.summary.timepoint.csv. This will contain summary statistics summarising predictor variables. The code will generate these files for you.

**Upload of results**

When you are ready to upload your results, please email [laura.schellhas@bristol.ac.uk](mailto:laura.schellhas@bristol.ac.uk) and I will provide you with a personal URL for upload.

**R code**

All the R code to perform these analyses is provided at: <https://github.com/ammegandchips/Prenatal_Caffeine/blob/master/Rcode.R>

**Please** use this code! If you have any questions about the analysis and/or are struggling to get the code to run, please email [gemma.sharp@bristol.ac.uk](https://github.com/ammegandchips/PACE_Paternal_BMI/blob/master/gemma.sharp@bristol.ac.uk).

If you have insufficient data to complete one or more of the EWAS, you can just skip those models. The code also produces .csv files summarising the variables included in the EWASs. You shouldn't have to rewrite or add to the code, unless otherwise stated.

There are just two inputs required for these analyses:

pheno: a dataframe containing all the "phenotype" data needed for this project. Each row is a sample(individual) and each column is a different variable. Necessary variable names are: "mat.caff", mat.smoking ", "sex","mat.ses", "mat.age", "parity", “mat.bmi”, "gest.age". If these columns are named differently in your dataset, please rename the columns accordingly.

meth: a matrix of methylation illumina beta values. Each column is a sample and each row is a probe on the array (450k or EPIC). Column names must correspond to the sample.id column in pheno.

**Thank you!**

Finally, thank you VERY much for taking the time to run these analyses. We will keep you updated on the progress of the project.