Paternal smoking PACE Analysis Plan: Phase One

Gemma Sharp

02/01/2018

Table of Contents

# Background and aims

There is increasing evidence that paternal lifestyle factors and environmental exposures in the prenatal period can influence offspring health, potentially via epigenetic changes in the germline. There is some evidence that paternal smoking is associated with variation in DNA methylation in sperm and offspring. Other potential explanations for associations between paternal smoking and offspring methylation at birth include 1) unobserved and/or residual confounding, including that introduced by active maternal smoking, and 2) an indirect intrauterine effect of paternal smoking due to maternal passive smoke exposure to fathers’ smoke.

The aims of this PACE project are to investigate:

1. Associations between paternal periconceptional smoking and offspring blood DNA methylation at birth, childhood and beyond.
2. The likelihood that such associations are causal, i.e. due to a direct effect of paternal smoking rather than a maternal (intrauterine) effect or confounding by genetic variation or other pre- or post-natal environmental factors.
3. Offspring phenotypes that might be mediated by paternal smoking-associated DNA methylation.

The project is split into two phases. **The first phase, described in this analysis plan, will focus on aim 1**.

Findings from the first phase will feed into the second phase, where we will contact cohorts again to complete additional analyses relating to aims 2 and 3. In aim 2, we will use three causal inference techniques, the maternal/paternal negative control design, cross-cultural comparisons and Mendelian randomization, in an attempt to infer causality in associations identified in aim 1. In aim 3, we will use mQTLs (SNPs robustly associated with methylation) and MR-Base (a database and analytical platform for Mendelian randomization) to conduct a pheWAS (phenome-wide association study) to identify offspring phenotypes associated with differential methylation at paternal smoking-associated CpGs.

All of the code necessary to run these analyses will be provided. We hope that this will lessen the burden for contributing cohorts and it will also ensure that all outputs are consistently formatted. This will significantly reduce the burden on the people conducting the meta-analysis. Therefore, **please use the supplied code**.

# Analyses relating to aim 1

## Exposures

* Paternal active smoking status around pregnancy: binary numeric variable (1=smoking during pregnancy or <=3 months before conception/0=no smoking during pregnancy or <= 3 months before conception). Smoking status can be self- or maternal-reported.
* In order to compare paternal and maternal effects, some models include maternal active smoking during pregnancy (as either the main exposure or a covariate). This should be coded as a binary numeric variable (1=smoking throughout pregnancy/0=no smoking in pregnancy or quitting smoking after the first trimester)
* In order to compare paternal and active and passive maternal effects, one model examines maternal passive smoking during pregnancy as the exposure variable. This should be coded as a binary numeric variable (1=passive smoking for at least an hour per day X times per week during pregnancy/0=no passive or active smoking during pregnancy). Please set as NA if the mother OR the father is an active smoker.

## Outcome

* Illumina Infinium 450k or EPIC BeadChip DNA methylation data in blood.
* We are interested in newborns (cord blood or neonatal blood spots collected at birth), young children (6 months-4 years old), children (5-11 years old) adolescents (12-17 years old) and/or adults (18 and above). If you have data for multiple time points, please analyse these separately.
* 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> 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

Perhaps moreso than in previous PACE analyses, it is very important that covariates are coded exactly as outlined below. The R code relies on these codings!

* **Paternal BMI:** continuous, numerical variable in kg/m2. Values can be either self-reported, maternal-reported or measured. Peri-conceptional BMI or BMI during pregnancy is preferred, but BMI in the perinatal period is also acceptable. Please double check BMI values >= +/- 5 SD from the mean in your dataset to make sure they are not data entry errors.
* **Maternal BMI:** continuous, numerical variable in kg/m2. Values can be either self-reported or measured. Pre-pregnancy BMI is preferred, but early pregnancy BMI is also acceptable. Please double check BMI values >= +/- 5 SD from the mean in your dataset to make sure they are not data entry errors.
* **Paternal social class:** binary numeric variable (1=high/0=low), please use your preferred classification, but note that our preference is for education level.
* **Maternal social class:** binary numeric variable (1=high/0=low), please use your preferred classification, but note that our preference is for education level.
* **Paternal age:** continuous numeric variable in years
* **Maternal age:** continuous numeric variable in years
* **Parity:** binary numeric variable (1=one or more previous children/0=no previous children)
* **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). Studies with newborn methylation should use the Bakulski et al. cord blood reference panel and *include all 7 cell types* generated: nRBC,CD8T, CD4T, NK, Bcell, Mono, Gran. Studies of older children and adults should use the Reinius adult reference panel and *include all 6 cell types* generated: CD8T, CD4T, NK, Bcell, Mono, Gran.
* **Selection factors:** Please include if relevant for your study, for example if your sample contains cases and controls for a condition, please include the case/control variable (coded as appropriate).
* **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:** Binary numeric variable. This will be used to stratify analyses (1=females,0=males).

## Exclusions

* Please exclude data for partners who are not the biological father of the child, according to either maternal or partner report. If there is any uncertainty about whether the partner is the biological father, please exclude. Please make a note of how many fathers are excluded for this reason and include this information in the Excel output.
* Please also exclude multiple pregnancies (e.g. twins) and siblings (i.e. each mother/father should appear only once in the dataset)

## EWAS models

### Minimally-adjusted: Paternal active smoking + Covariates + Cells (model name: minimal.pat)

Methylation ~ Paternal active smoking + SVs for batch + Paternal age + Paternal BMI + Paternal socioeconomic status + Maternal age + Maternal BMI + Maternal social class + Parity + Estimated cell counts

### Minimally-adjusted: Maternal active smoking + Covariates + Cells (model name: minimal.mat)

Methylation ~ Maternal active smoking + SVs for batch + Paternal age + Paternal BMI + Paternal socioeconomic status + Maternal age + Maternal BMI + Maternal social class + Parity + Estimated cell counts

### Mutually-adjusted: adjusted for other parent’s active smoking status, covariates and cells (model name: mutual)

*Note that the code extracts results for the effect of paternal AND maternal active smoking*

Methylation ~ Paternal smoking **+ Maternal active smoking status** + SVs for batch + Paternal age + Paternal BMI + Paternal socioeconomic status + Maternal age + Maternal BMI + Maternal social class + Parity + Estimated cell counts

### Mutually-adjusted, in boys only (model name: mutual.boys)

*Note that the code extracts results for the effect of paternal AND maternal active smoking*

*Please exclude any female offspring*

Methylation ~ Paternal active smoking + Maternal active smoking + SVs for batch + Paternal age + Paternal BMI + Paternal socioeconomic status + Maternal age + Maternal BMI + Maternal social class + Parity + Estimated cell counts

### Mutually-adjusted, in girls only (model name: mutual.girls)

*Note that the code extracts results for the effect of paternal AND maternal active smoking*

*Please exclude any male offspring*

Methylation ~ Paternal active smoking + Maternal active smoking + SVs for batch + Paternal age + Paternal BMI + Paternal socioeconomic status + Maternal age + Maternal BMI + Maternal social class + Parity + Estimated cell counts

### Paternal active smoking, only when mothers are not active smokers (model name: pat.only)

*Please exclude any mothers who are active smokers*

Methylation ~ Paternal active smoking + SVs for batch + Paternal age + Paternal BMI + Paternal socioeconomic status + Maternal age + Maternal BMI + Maternal social class + Parity + Estimated cell counts

### Maternal passive smoking in the absence of maternal or paternal active smoking (model name: passive)

*Please exclude any mothers who are active smokers themselves or have partners who are active smokers*

Methylation ~ Maternal passive smoking + SVs for batch + Paternal age + Paternal BMI + Paternal socioeconomic status + Maternal age + Maternal BMI + Maternal social class + Parity + Estimated cell counts

# Outputs

Please supply the following files in the specified formats:

1. EWAS results (For each time point, one Rdata file labelled as YOURSTUDY.patsmoking.ewasresults.timepoint.Rdata. This will contain all outputs from all EWAS for one time point. The code will generate this file for you. If you have multiple time points, you will have an Rdata file for each time point: birth, early.childhood, late.childhood, adolescence, or adult)
2. Extra cohort information (one Excel file labelled as YOURSTUDY.patsmoking.cohortinfo.xlsx. This will contain information for each cohort relating to things like normalisation method and number of fathers excluded because of non-paternity. Please download and use the template available at <https://github.com/ammegandchips/PACE_Paternal_Smoking/TEMPLATE.patsmoking.cohortinfo.xlsx>)
3. EWAS variables summary (For four EWAS models (mutual.pat, mutual.mat, pat.only, passive), a csv file labelled as YOURSTUDY.modelname.summary.timepoint.csv. This will contain summary statistics summarising predictor variables for each of these four EWAS models. The code will generate these files for you. If you have multiple time points, you will have a csv file for four EWAS within each time point: birth, early.childhood, late.childhood, adolescence, or adult)

Examples of these output files are provided at <https://github.com/ammegandchips/PACE_Paternal_Smoking/>

# Upload, timescale and contacts

* Deadline for upload of results is 1st March 2018.
* When you are ready to upload your results, please email <gemma.sharp@bristol.ac.uk> and I will provide you with a personal URL for upload.

# R code

Instructions and all the R code to perform these analyses is provided at: <https://github.com/ammegandchips/PACE_Paternal_Smoking/phase1.patsmoking.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>.

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 this analysis:

1. 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: “pat.active.smoking”, “mat.active.smoking”, “mat.passive.smoking”, “sex”,“pat.ses”, “mat.ses”, “pat.age”, “mat.age”, “pat.bmi”, “mat.bmi”, “parity”. If these columns are named differently in your dataset, please rename the columns accordingly.
2. 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. I will keep you updated on the progress of the project through the PACE calls. If we find associations in Phase One, I will be in touch regarding additional analyses for Phase Two :)