EMPHASIS analysis plan

# Overall objectives

Analyze the DNA methylation data generated for our Indian and Gambian cohorts in order to: **1.** identify methylation differences associated with intervention, **2.** link these to the measured health outcomes, **3.** investigate outcome-associated methylation differences independent of intervention. We may also wish to explore: **4.** the associations between intervention and outcomes independent of methylation status (TBD).

# Outcomes of interest

|  |  |  |
| --- | --- | --- |
| *Birth*  *measures* | *Anthropometry* | |
| *(PMMST/MMNP)* |
|  |  |  |
|       Weight |       Weight |       BMI |
|       Length |       Waist |       Triceps skinfold |
|       Head circum. |       Height |       Bicep skinfold |
|  |       Sitting height |       Subscapular skinfold |
|  |       Head circum. |       Supra-iliac skinfold |
|  |       Mid-upper arm circum. |       DXA / body composition |
|  |       Chest |  |
|  |       Hip |  |
|  |  |  |
| *Cardio-metabolic* | *Cognitive ability* | *Derived measures* |
| *risk markers* |
|  |  |  |
|       Systolic BP |       short-term memory |       Growth |
|       Diastolic BP |       long-term mem. & retr. abil. |       Adiposity |
|       Pulse |       visuo-spatial ability |       Cardio |
| * Glucose metabolism |       language production |       Cognitive |
| * Triglyceride profile |       Intelligence quotient |  |

# Adjustment covariates

Final list to be determined by assessing association of covariates with i) intervention, ii) methylation principal components, iii) known effects on methylation (age, sex). Strategy for dealing with these will vary depending on stage of analysis - discussed further in the sections that follow.

|  |  |  |
| --- | --- | --- |
| *General* | *Intervention* | *Lab* |
|
|       Infant sex |       Time on intervention |       Full blood count / cell  composition (estimated from |
|       Infant age |  | methylation data) |
|       Intra-cohort ethnicity |  |       genotype (mQTL) |
|       Gambian - season of conception |  |       EPIC array batch effects |
|       Indian – season/month of measurement |  | (plate, slide, position, well) |
|  |  |  |
|  |  |  |
|  |  |  |
| *Maternal* | *Delivery* |  |
|  |
|       Age |  |  |
|       Weight / BMI |       Live birth |  |
|       Systolic BP |       Multiple birth |  |
|       Diastolic BP Systolic BP |       Ballard score for |  |
|  |       gestational age |  |

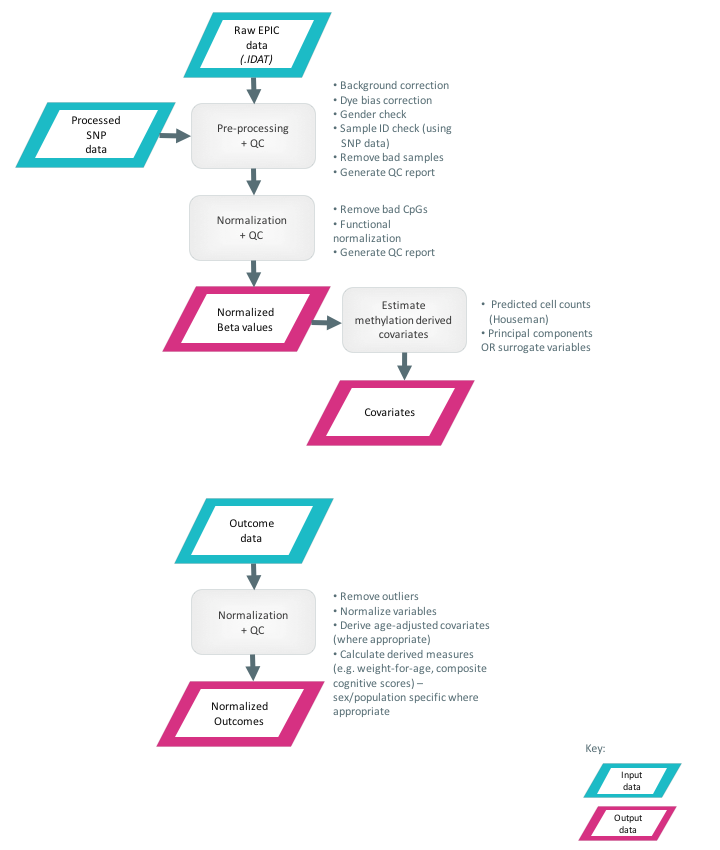
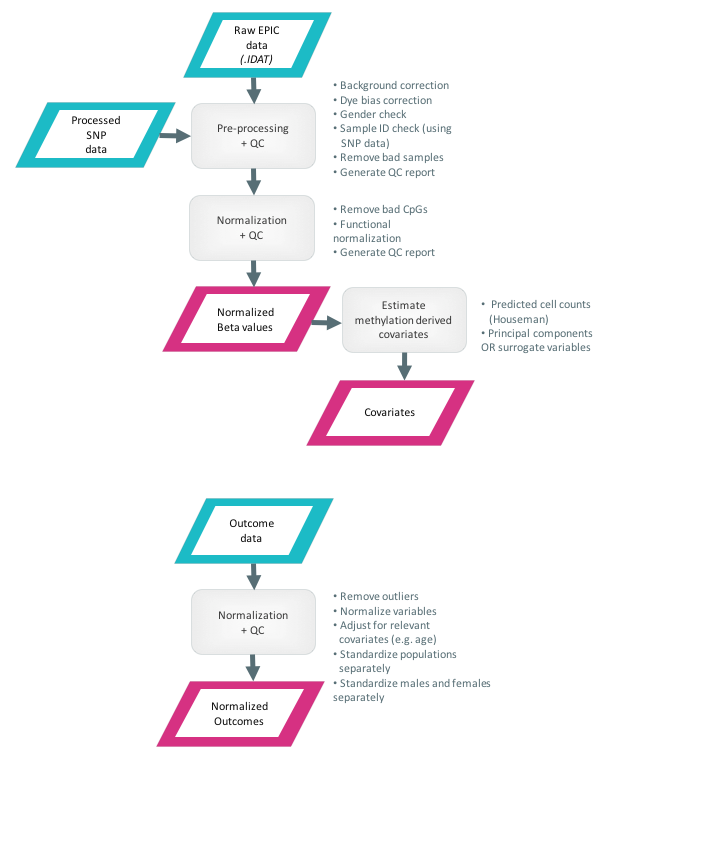
# Prerequisites

* Outcome data
* EPIC .IDAT data files - raw intensity measurements from the EPIC arrays
* Genotype data - processed genotypic data from the genotyping arrays (for QC, mQTL filtering / adjustment and MR)
* Bisulfite pyrosequencing data – processed quantitative methylation data for candidate and replicated loci
* Software: *R* + *bioconductor* + *meffil* + *sva* + *bumphunt* + *TwoSampleMR* + more (TBD)

# Analysis overview

1. Methylation data processing
2. Outcome data processing
3. Intervention-methylation associations
   1. EWAS
   2. Candidate loci
   3. Technical validation
   4. Replication
   5. Cross-tissue comparison
4. Methylation-outcome associations
5. Cross-cohort comparison
6. Causal inference

# 1. Methylation data processing

*Overview*

The raw intensity data from the EPIC arrays (.IDAT files) will first be subject to a series of pre-processing and quality control steps intended to ensure data integrity. Following this, the data is normalized to produce comparable measurements across the arrays and to adjust for technical factors. The output from this stage in the analysis will be an *m* x *n* matrix of normalized *Beta* values for *m* probes and *n* subjects, and an *i* x *j* matrix of covariates for *i* subjects and *j* covariates.

*Pre-processing + quality control*

**Check data integrity**

First, check that intact red and green .IDAT files are present for each sample (using MD5 checksum).

**Initial QC check**

Produce box plots of the raw intensities, manually inspect for outlying arrays (e.g. unexpectedly low/high intensity readings across the board, high variability etc.).

**Generate *meffil* objects**

Generate sample sheet and QC objects in *meffil*.

**Run QC procedure**

The *meffil* QC procedure includes background correction, dye bias correction, and sex prediction.

**Check for ID mismatches**

Extract set of control probes from genotype array data using *plink*, check that IDs match using genotypes from *meffil* (this might be done posthoc, depending on when the data becomes available)

**Generate and inspect *meffil* QC report**

**Remove sample outliers**

Remove bad samples based on QC report. Example criteria include: sex mismatch, genotype mismatch, low bead numbers, control probe issues (dye bias, bisulfite conversion).

**Remove bad probes**

Remove probes with low detection *p*-values (det P <0.01) using *meffil*

**Generate and inspect *meffil* post-QC report**

*Normalization + quality control*

**Perform normalization**

Perform functional normalization using *meffil*

**Generate and inspect *meffil* post-normalization QC report**

*Cell composition estimation*

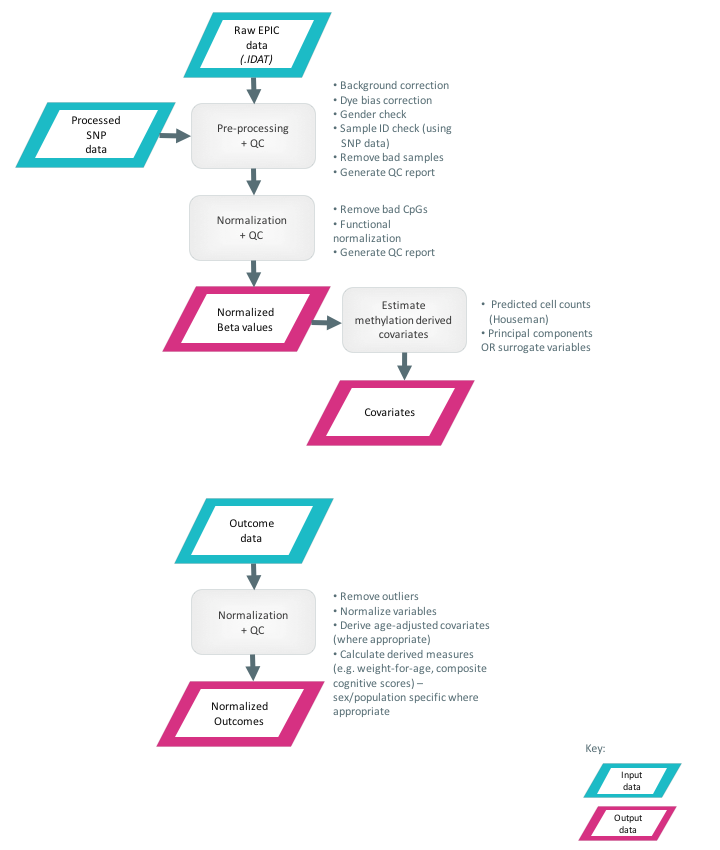
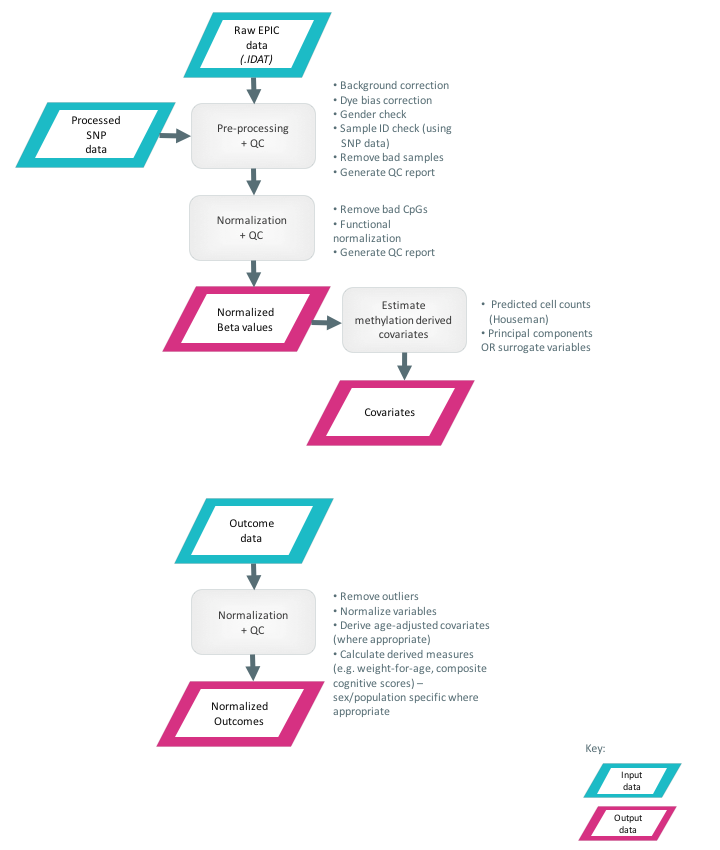
Use Houseman method to estimate blood cell counts from normalized Beta values. Compare with blood cell proportions from FBC data. Decide on final combination of cell composition adjustment covariates (FBC + Houseman)

*Batch / principal component adjustment*

Generate surrogate variables – either independent surrogate variables (ISVs) or principal components (PCs) to be used as adjustment covariates at linear modelling stage

# 2. Outcome data processing

*Overview*



The outcome data will first undergo processing to remove outliers and adjust for known confounders. Where appropriate data will be transformed to approximate normal distributions. Any missing data also needs to be flagged at this stage and a decision taken on whether to impute or remove. The output generated here will be an *i* x *k* matrix processed outcomes, for *i* subjects and *k* outcomes.

*Normalization + quality control*

**Remove outliers**

Remove extreme outliers, e.g. 5 SD outliers (TBD).

**Create derived/composite measures**

Given the number of measures available to test, we may wish to derive composite or latent measures, e.g. of anthropometry, body composition, etc. (TBD). This will be based on prior biological knowledge about which measures are appropriate to combine, or can be done in a data driven way (through unsupervised dimension reduction)(TBD)

**Adjust**

Where appropriate adjust for known confounders, e.g. height for age, BMI for age.

**Sex adjustment**

Where appropriate, perform normalization and adjustment on males and females separately.

**Normalize**

Normalize data, e.g. inverse normal transformation for height and BMI.

**Missing data**

Either impute or remove.

**Harmonize**

Compare phenotypic distributions across cohorts and consider whether further steps need to be taken to ensure cross-cohort comparability of downstream cohort-specific statistical analyses.

# 3. Intervention-methylation associations

# a. EWAS

To be performed independently for each cohort.

*Overview*

The first stage of the inferential analysis will attempt to identify genome-wide methylation differences associated with nutritional intervention, in the form of differentially methylated positions (DMPs) and differentially methylated regions (DMRs). We may also look at variably methylated positions (VMPs) and variably methylated regions (VMRs) (TBD).

*Strategy*

**Identification of DMPs**

***Linear regression***

Utilize a linear modeling strategy to identify differentially methylated positions / CpGs (DMPs) associated with the intervention. A number of approaches can be used, two are described here: 1. robust linear regression and model selection, as used in PACE, 2. standard linear regression using estimated covariates, as used in GoDMC. Important considerations here include: whether we want to use *Beta* or M values for methylation, whether we should pre-adjust methylation values for batch effects or include these in the model, and which variables we wish to include in the model.

1. A model selection approach could be used where a number of different possible models are tested, and the one providing the best fit or closest to the expected null distribution of *p*-values chosen.

An example of a model that might be used:

*Beta ~ sex + cell counts + 1 | batch + time on intervention + intervention*

Here, methylation beta values are regressed on sex, cell counts, batch (as a random effect), time on intervention and intervention (the variable of interest). Various reduced forms of the above model might be fitted, and the results compared to select the optimal one.

1. An alternative strategy would be to use covariates estimated from the data, which we would expect to capture the variation from technical sources, and use a model such as:

*Beta ~ SV1 + SV2 + sex + cell counts + batch + time on intervention + intervention*

Where surrogate variables are included to model variation that is not associated with intervention. The selection of SVs will depend on correlation with batch variables and potentially some of the other outcome variables (TBD).

***Significance assessment***

We will account for multiple testing by controlling the false discovery rate (FDR) at 10%. Alternatively, we will consider using permutation testing to estimate significance of associations. This involves the repeated random shuffling of phenotype labels to generate a null distribution, which can then be used to assess the significance of the observed findings.

TODO

**Identification of DMRs**

Region-level analysis will be performed using the *bumphunting* and *DMRcate* methods. DMRs will be definted as regions identified with FDR<10% using both methods.

**Identification of VMPs**

TBD

**Identification of VMRs**

TBD

# b. Candidate loci

To be performed independently for each cohort.

*Overview*

**Identification of DMPs**

***Linear regression***

A similar modelling strategy to the one outlined above will be also be used to identify methylation differences associated with nutritional intervention for the pyrosequenced candidate loci.

An example of a model that might be used:

*Beta ~ sex + cell counts (estimated from array data) + time on intervention + intervention*

***Significance assessment***

Since these candidates are identified *a priori* as loci of interest, no correction for multiple testing is necessary.

**Identification of DMPs**

Where CpGs within a candidate locus are correlated, a single outcome measure can be derived, eg mean methylation value / Z-scores. Otherwise other methods can be used, e.g. performing a global test (combination of p-values) or taking the minimum p-value for the region.

# c. Technical validation

To be performed independently for each cohort

*Overview*

For the top 20 (TBD) hits identified by EWAS, pyrosequencing data will be generated and the correlation between the methylation values from both platforms assessed to confirm that these agree.

# d. Replication (Indian cohort only)

*Overview*

The top 20 (TBD) hits identified by EWAS will also be pyrosequenced in the replication sample, and this data once again analyzed as described in **a.** above.

# e. Cross-tissue correlation

*Overview*

For replicated loci and top hits, buccal cell samples will be pyrosequenced, and analyzed as before. This will enable a comparison of methylation status between tissues derived from two different cell lineages, which will allow significant sites to be separated into those showing evidence of metastability, and those that are tissue specific.

# 4. Methylation-outcome associations

*Overview*

The second stage of the inferential analysis concerns associations between methylation loci and selected outcomes. Two separate analyses will be conducted:

1. EWAS to identify loci where methylation is associated with outcomes of interest
2. Test for association between any significant/replicated differentially methylated loci from the Stage 1 EWAS and the outcomes of interest.

# a. EWAS

*Strategy*

**Identification of DMPs**

***Linear regression***

For each locus and for all the outcomes, the following linear model is used (TBD):

*Beta ~ sex + cell counts + 1 | batch + time on intervention + outcome + other adjustment cov*

Other adjustment covariates will be selected depending on which are associated with outcome.

***Significance assessment***

We will account for multiple testing by controlling the false discovery rate (FDR) at 10%. Alternatively, we will consider using permutation testing to estimate significance of associations. This involves the repeated random shuffling of phenotype labels to generate a null distribution, which can then be used to assess the significance of the observed findings.

**DMRs**

Region-level analysis will be performed using the *bumphunting* and *DMRcate* methods. DMRs will be definted as regions identified with FDR<10% using both methods.

# b. Intervention-associated loci

*Overview*

Intervention-associated candidate loci from the previous stage of the analysis will also be tested for association with the outcomes.

*Strategy*

**DMPs**

For each locus and for all the outcomes, the following linear model is used (TBD):

*Beta-value ~ sex + cell counts + time on intervention + outcome + adjustment covariates*

Other adjustment covariates will be selected depending on which are associated with outcome.

Since these candidates are identified *a priori* as loci of interest, no correction for multiple testing is necessary.

**DMRs**

For each locus and for all the outcomes, the following linear model is used (TBD):

*Beta-value or derived DMR summary measure ~*

*sex + cell counts + time on intervention + outcome + adjustment covariates*

Other adjustment covariates will be selected depending on which are associated with outcome.

Since these candidates are identified *a priori* as loci of interest, no correction for multiple testing is necessary.

Where CpGs within a candidate DMR are correlated, a single outcome measure can be derived, eg mean methylation value / Z-scores. Otherwise other methods can be used, e.g. performing a global test (combination of p-values) or taking the minimum p-value for the region.

# 5. Cross cohort comparison

TODO

# 6. Causal inference

Two step, two sample Mendelian randomization performed using the *TwoSampleMR* package from *MRCIEU* (TBD).

*Overview*

TODO

*Strategy*

TODO