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.** association 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 |
|       Infant age | (weeks prior to estimated | composition (estimated from |
|       Intra-cohort ethnicity | time of conception) | methylation data) |
|       Gambian - season of conception |  |       genotype (mQTL) |
|  |  |       EPIC array batch effects |
|  |  | (plate, slide, position, well) |
|  |  |  |
|  |  |  |
|  |  |  |
| *Maternal* | *Delivery* |  |
|  |
|       Age |       Live birth |  |
|       Weight / BMI |       Multiple birth |  |
|       Systolic BP |       Ballard  score for gestational age | |
|       Diastolic BP Systolic BP |  |  |
|  |  |  |

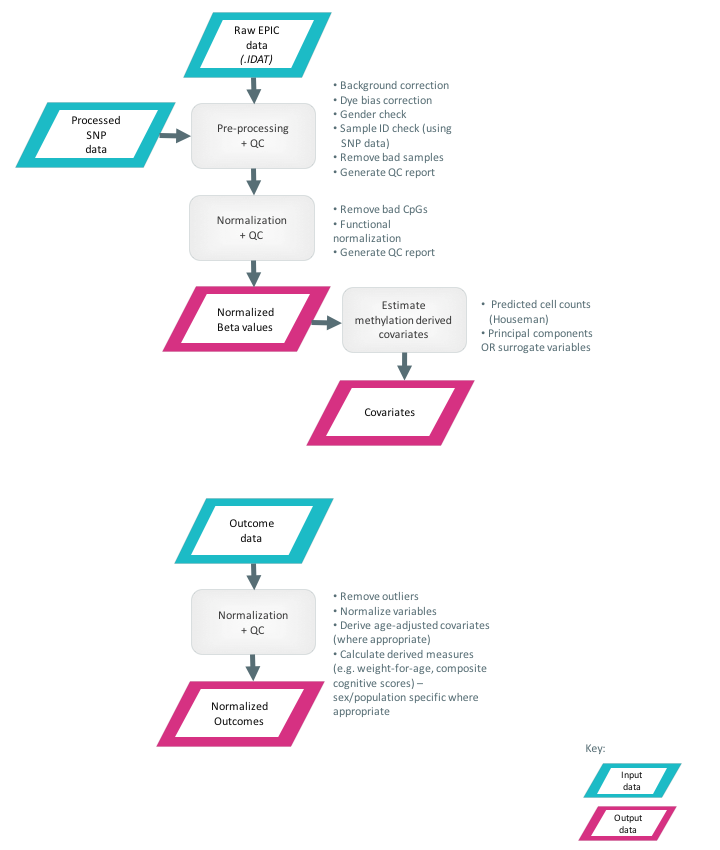
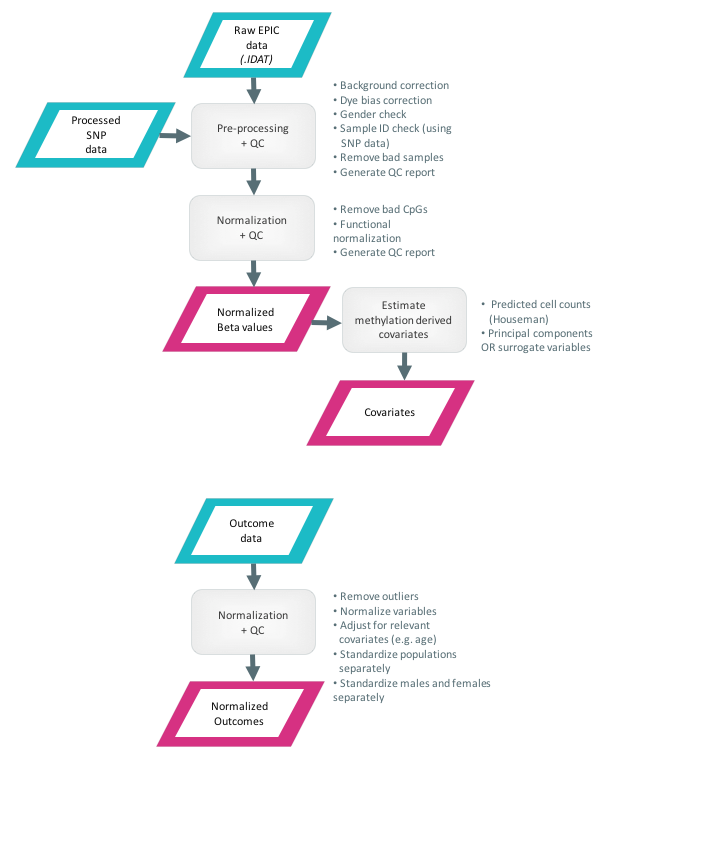
# Prerequisites

* 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* + *limma* + *sva* + *bumphunter* + *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. Pathways analysis
6. Cross-cohort comparison
7. Causal inference and mQTL analysis

# 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**

Default *meffil* sample exclusion criteria will be used (along with the *goDMC* thresholds):

* sex mismatch
* sex outliers – 5 SDs
* genotype mismatch (when genotype data available) – 0.8
* high detection *p*-value ( > 0.01) – 0.1
* low bead numbers ( < 3) – 0.1
* methylated vs unmethylated outlier – 3 SDs
* control probe outliers (bisulfite1, bisulfite2, & more?) – 5 SDs

**Remove bad probes**

Remove probes with:

* high detection *p*-value ( > 0.01) – 0.1
* low bead numbers ( < 3) – 0.1

**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*

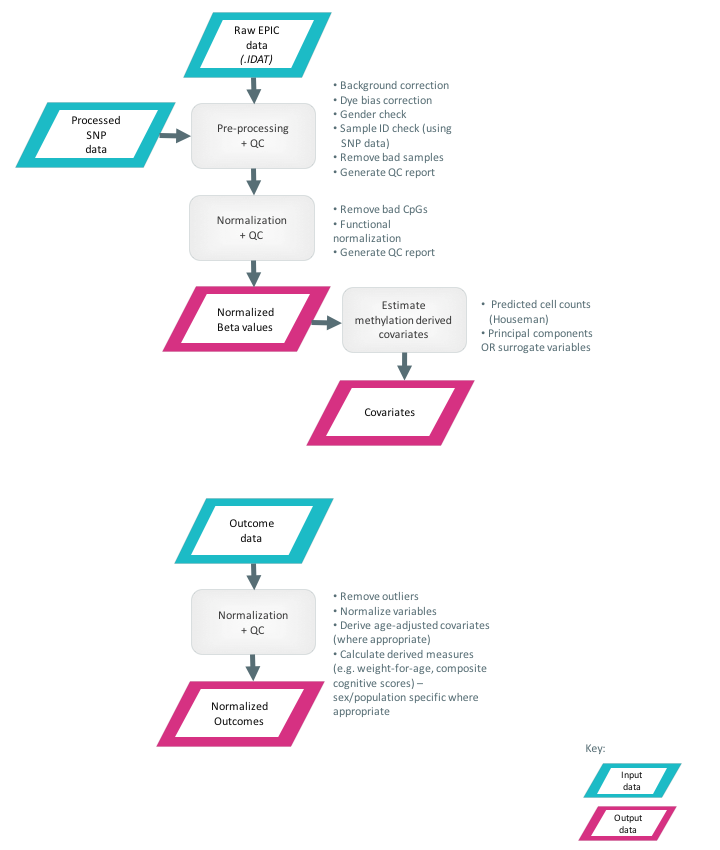
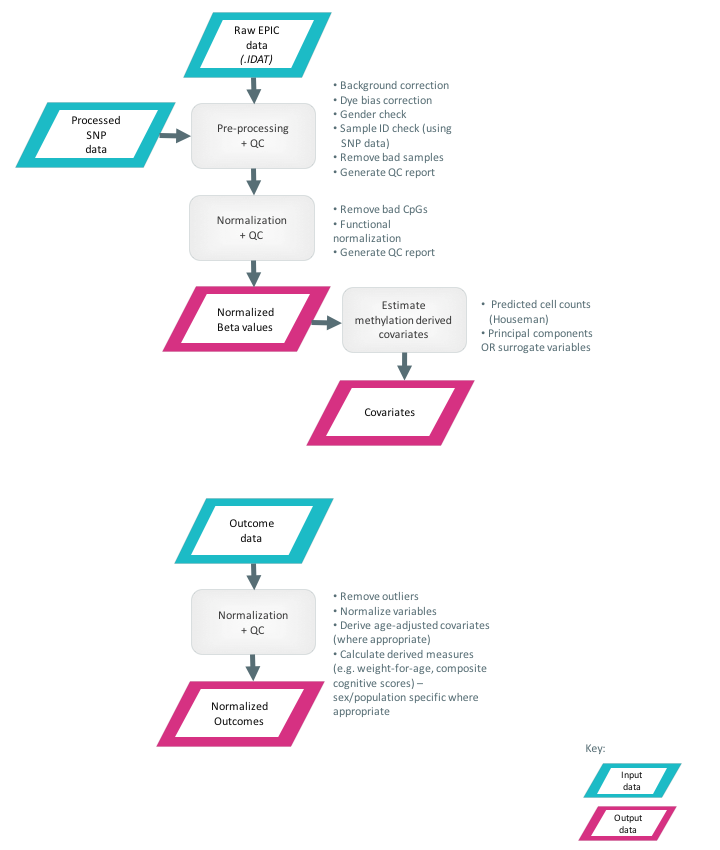
Use Houseman method to estimate blood cell counts from normalized Beta values. Provides estimates for the following cell types: Eos, CD8T, Bcell, Mono, NK, CD4T, Neu. Where applicable, compare with recorded cell count data.

*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. The precise strategy for adjusting for batch covariates, unmeasured sources of variation etc will be decided at the modelling stage by investigating test statistic distributions, correlations with known variables etc.

# 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 of processed outcome variables, for *i* subjects and *k* outcomes.

*Normalization + quality control*

**Remove outliers**

Remove extreme outliers, based on median absolute deviation: (below median - 5 x MAD or above median + 5 x MAD)

*Note: alternatively, we may wish to base this on interquartile range*

**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. This will be based on prior biological knowledge about which measures are appropriate to combine, or in some cases can be done in a data driven way (e.g. unsupervised dimension reduction of cognitive measures to produce one or more derived variables)(TBD)

**Adjust**

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

Where appropriate missing data will be imputed, otherwise subjects with missing data for a particular outcome will be removed. For imputation, we will use multiple imputation with chained equation (R *mice* package) to impute each variable using a regression model conditional on the others

**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.

**Exclusion criteria**

Main exclusion criteria were predetermined by the original study protocols (MMNP/PMMST).

*Compliance*

For MMNP compliance was defined as >= 3 snacks per week from 90d before last menstrual period. Non-compliant subjects are excluded from this study.

For PMMST compliance was measured as % of tablets apparently consumed. The estimate of compliance did not differ between treatment groups (MMN: 0.96±0.18; placebo: 0.92±0.19, P=0.15). No exclusions are reported in the paper.

*Glucose tolerance test*

PMMST: exclude GTT measurements for non-fasted subjects

MMNP: ?

# 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 will also look at variably methylated positions (VMPs) and variably methylated regions (VMRs).

*Modelling strategy*

**Identification of DMPs**

***Linear modelling***

To identify differentially methylated positions / CpGs (DMPs) associated with the intervention, the *limma* package in R will be used to fit the following model:

*M ~ intervention + SVs + sex + further\_covs*

Where

* M is the methylation signal for the site being tested (M-value = log2(beta/(1-beta));
* SVs are the surrogate variables intended to capture technical and other unmeasured variation not associated with intervention (e.g. batch effects). The precise method for determining surrogate variables (ISVA, SVA, PCA) will be decided after investigating correlations with known confounders and test statistic distributions after adjustments using QQ plots;
* futher\_covs are additional adjustment covariates that are a) associated with the intervention, or b) known to influence methylation (e.g. estimated cell composition, child’s BMI, maternal BMI, Gambian season of conception) that are not captured by (correlated with) included surrogate variables.

We may also perform a post-hoc sensitivity analysis where other covariates will be added to the model to see how these impact effect size estimates at significant loci.

*Note : an alternative strategy using Beta-values could be used which would fit the same model as above but using robust standard errors to account for non-constant variance (e.g. using the sandwich package in R)*

***Significance assessment***

We will account for multiple testing by controlling the false discovery rate (FDR) < 5%. We may also consider using permutation testing to estimate significance of associations.

**Identification of DMRs**

Region-level analysis will be performed using the *bumphunting* and *DMRcate* methods. These methods rest on different assumptions and results from each method will be compared. DMRs will be defined as regions identified with FDR < 5%.

**Identification of VMPs**

Loci showing differences in methylation variance (VMPs) will be identified using the diffVar method from the *missMethyl* package in R.

**Identification of VMRs**

Regional differences in methylation variance (VMRs) will be identified using the method by Ong and Holbrook described in <http://www.ncbi.nlm.nih.gov/pubmed/24112369>.

# b. Candidate loci

To be performed independently for each cohort.

*Overview*

**Identification of DMPs**

***Linear modelling***

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

*M ~ intervention + batch + sex + cell\_counts + further\_covs*

Where batch refers to the bisulfite conversion batch, and cell\_counts are those previously estimated for the EWAS.

***Significance assessment***

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

**Identification of DMPs**

Where a candidate locus contains multiple CpGs and these are correlated, a single outcome measure will be derived using mean methylation Z-scores (with logit transformation where necessary). Otherwise other methods will be considered, e.g. global test (combination of p-values) or taking the minimum p-value for the region.

# c. Technical validation

Technical validation of significant DMPs and DMRs from the EPIC association analysis will be performed by pyrosequencing a subset of samples spanning the range of observed methylation values and assessing correlations between platforms. Precise numbers for validation in each cohort will depend on the number of significant associations identified. A total of 700 pyrosequencing reactions have been budgeted for.

# d. Replication (Indian cohort only)

*Overview*

The technically validated sites from c. above will also have their methylation levels assessed by pyrosequencing in an independent replication sample of size n= 400.

# e. Cross-tissue correlation

*Overview*

Buccal cell samples will be pyrosequenced at technically validated significant loci from the EPIC EWAS and at candidate loci. The precise number of samples analysed will depend on the number of EWAS hits. 1,600 pyrosequencing reactions have been budgeted for . This will enable a comparison of methylation patterns between tissues derived from two different cell lineages to establish potential tissue-specificity of significant signals.

# 4. Methylation-outcome associations

*Overview*

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

1. Test for association between significant DMPs and DMRs associated with the intervention identifed in the Stage 1 EWAS, and the outcomes of interest.
2. Test for association between candidate (pyrosequenced) loci and outcomes of interest
3. Hypothesis-free EWAS to identify loci where methylation is associated with outcomes of interest

# a. Intervention-associated loci

An identical modelling strategy to that descrbed in 3a. above will be used to test associations between intervention-associated loci and all outcomes of interest, i.e.

*M ~ outcome + SVs + sex + further\_covs*

with outcomes of interest modelled as factors, continuous or ordinal variables as appropriate.

For nutrition-associated DMRs, regional methylation will be assessed either by taking mean logit-transformed methylation (for correlated loci) as a single outcome variable, or on a cpg-by-cpg basis.

b. Candidate loci

An identical modelling strategy to that descrbed in 3b. above will be used to test associations between candidate (pyrosequenced) loci and all outcomes of interest, i.e.

*M ~ outcome + batch + sex + cell\_counts + further\_covs*

Since all loci in a. and b. above are identified *a priori* as loci of interest, no correction for multiple testing is necessary.

# c. Hypothesis-free EWAS

An identical modelling strategy to that descrbed in 3a. above will be used, i.e.

*M ~ outcome + SVs + sex + further\_covs*

with outcomes of interest modelled as factors, continuous or ordinal variables as appropriate.

The same strategies for identifying DMRs, VMPs and VMRs will be employed as described in 3a.

# 5. Pathways analysis

Gene pathways analysis will be performed on both both intervention and outcome-associated EWAS. The *gometh()* gene ontology analysis function from *missMethyl* will be used as this accounts for potential biases due to large differences in the number of CpGs mapped to different genes on methylation arrays.

# 6. Cross cohort comparison

We will consider a number of methods for comparing results across Indian and Gambian cohorts. For example we will look for overlap between significant DMPs/DMRs/VMPs/VMRs at FDR < 10% between the two cohorts. We will also compare the results of pathway analyses.

# 7. Causal inference and mQTL analysis

The collection of genotype data using Illumina’s GSA array gives us a number of further options for exploring results from the analyses described above.

Depending on the results, we will consider options for performing causal analysis using Mendelian Randomisation (e.g. ‘two-sample MR’ using the *TwoSampleMR* package from MRC IEU). We can also consider potential SNP effects at significant loci by performing an mQTL analysis.