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).

# Phenotypes / outcomes of interest

|  |  |  |
| --- | --- | --- |
| *Birth measures* | *Anthropometry* | *Body composition* |
| *(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. |  |
|  |   Chest |  |
|  |   Hip |  |
|  |  |  |
| *Cardio-metabolic* | *Cognitive ability* | *Derived measures* |
| *risk markers* |
|  |  |  |
|   Systolic BP |   TODO |   Growth |
|   Diastolic BP |  |   Adiposity |
|   Pulse |  |   Cardio |
|  |  |   Cognitive |
|  |  |  |

# Adjustment covariates

|  |  |  |
| --- | --- | --- |
| *General* | *Intervention* | *Lab* |
|
|  |  |  |
|   Gender |   Time on intervention |   Full blood count |
|   Date of birth |  |   Sentrix plate |
|   Intra-cohort ethnicity |  |   Sentrix slide |
|   Gambian - season of |  |   Sentrix position |
| conception |  |   Sentrix well |
|  |  |  |
|  |  |  |
| *Maternal* | *Delivery* |  |
|  |  |  |
|   Age |   Live birth |  |
|
|   Weight |   Multiple birth |  |
|   Mid-upper arm circum. |   Ballard score for |  |
|   Systolic BP | gestational age |  |
|   Diastolic BP |   Congenital abnormality |  |
|  |   Placental weight |  |
|  |  |  |
|  |  |  |

# 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 purposes)
* 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. Phenotypic data processing
3. Intervention-methylation associations
   1. EWAS
   2. Candidate loci
   3. Technical validation
   4. Replication
   5. Cross-tissue validation
4. Methylation-phenotype associations
5. Cross-cohort comparison
6. Causal inference

# 1. Methylation data processing

*Overview*

AS_450K_pipeline_160509

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, normalization is used 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). Compare with sample sheet in order to make sure these have been named correctly (i.e. sample name/gender).

**Sanity 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.). Check gender labeling using MDS.

**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 - TBD)

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

**Remove outliers**

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

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

*Normalization + quality control*

**Perform normalization**

Use *meffil* to perform functional normalization (TBD). In *meffil* this also includes an adjustment for technical effects (using principle components), and removal of probes with low detection *p*-values.

Other options: 1. noob (background correction + dye-bias equalization) + functional normalization, 2. lumi (dye-bias + quantile-normalization)

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

*Estimate covariates*

**Batch effects**

There are a few different options for batch effect adjustment. The first is the default *meffil* approach using principal components associated with predefined technical effects. The second is the *sva* approach, whereby surrogate variables for unmeasured sources of variation are constructed, or alternatively, for known sources of variation e.g. Sentrix ID, array position (using *ComBat* function). These two approaches both produce batch-adjusted methylation values. A third alternative, would be to follow the *PACE* protocol and test known batch effect at the linear modeling stage.

**Blood cell counts**

Use Houseman method to estimate blood cell counts from normalized Beta values. Compare with blood cell proportions from FBC data, decide which provides optimal balance between variance explained vs number of dimensions (i.e. covariates) used (TBD).

# 2. Phenotypic data processing

*Overview*

AS_450K_pipeline_160509

The outcome data will first undergo processing to remove outliers, adjusted for known confounders, and 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 (TBD). The output generated here will be an *i* x *j* matrix processed outcomes, for *i* subjects and *j* outcomes.

*Normalization + quality control*

**Remove outliers**

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

**Separate by gender**

Perform the normalization and adjustment on males and females separately (if known gender differences).

**Normalize**

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

**Adjust**

Adjust for known confounders, e.g. adjust height for age, adjust BMI for age (TBD).

**Recombine data**

**Deal with missing data**

Either impute or remove (TBD).

**Create derived 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).

**Harmonize**

Check that comparable phenotypic measures are available for both cohorts and that they have been processed in the same way.

# 3. Intervention-methylation associations

# a. EWAS

*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**

Utilize a linear modeling strategy identify differentially methylated signals associated with the intervention (TBD). A number of approaches can be used, including robust linear regression (as in PACE), standard linear regression (as in GoDMC), linear regression with mixed effects etc. Important considerations here include: whether we want to use *Beta* or M values for methylation, whether we want to pre-adjust methylation values for batch effects or include these in the model, and which variables we wish to include in the model. A model selection approach could be used (similar to PACE) 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 complete model that might be used:

*beta ~ age + gender + cell counts + 1 | batch + time on intervention + intervention*

Here, methylation beta values are regressed on age, gender, 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.

Perform 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 (TBD).

TODO

**Identification of DMRs**

Region-level analysis is performed using the *bumphunting* method (TBD).

**Identification of VMPs**

TBD

**Identification of VMRs**

TBD

# b. Candidate loci

*Overview*

The analysis strategy outlined above will be also be used to identify methylation differences associated with nutritional intervention for the pyrosequenced candidate loci.

# c. Technical validation

*Overview*

For the top 20 (TBD) hits identified by EWAS, pyrosequencing data will be generated and analyzed taking the same approach as in **a.**

# 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 validation

*Overview*

For replicated loci and top hits, buccal cell samples will be pyrosequenced, and analyzed as before.

# 4. Methylation-phenotype associations

*Overview*

In the second stage of the inferential analysis, significant, and/or replicated/validated intervention-associated differentially methylated loci from the previous stage of the analysis will be tested for association with the outcomes.

*Strategy*

**DMPs**

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

*methylation (batch adjusted Beta-values) ~ age + gender + cell counts + time on intervention + outcome*

As before, permutation testing is then used to determine significance of the associations (TBD).

**DMRs**

Region-level analysis is performed using the *bumphunting* method (TBD).

# 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