GoDMC analysis plan

# Overall objective

Perform comprehensive analysis of genetics of methylation using data from multiple cohorts, and use this to link methylation with complex traits, including BMI, height, cell counts, smoking and ageing.

# Scope

## Genetics of methylation

Map cis- and trans- genetic variants underlying methylation variation; include SNPs and structural variants; perform conditional analysis to identify additional cis- effects; estimate the proportion of heritability that can be explained by mapped results; perform functional annotation

## Genetics of AAR

Estimate SNP-heritability and identify genetic associations for age acceleration residuals (AAR)

## Genetics of cell counts

Estimate SNP-heritability and identify genetic associations for cell counts

## Genetics of smoking

Estimate SNP-heritability and identify genetic associations for smoking

## Robustness of genetic signals

Investigate the robustness of genetic signals over time, between sexes, and across cell types

## Linking methylation to complex traits

Gauge the importance of methylation on complex traits, including estimation of the total predictive power of 450k methylation on height and BMI, make causal inference using Mendelian randomisation

# 

# Requirements

## Data for pre-pipeline

**Methylation:**

We wrote an R package “meffil” to preprocess raw idat files to a normalization matrix for big sample sizes. We strongly recommend using it to harmonize all cohorts. It comprises automated sample QC and normalization and prints out a pre- and post normalization report. It normalizes 5000 samples in ~2-3 hrs using 16 cores without requiring extremely large memory. The requirements are:

* Original idat intensities for methylation data using Illumina 450k array
* Optional: GWA genotypes to calculate sample discordance

**Genotypes:**

1000G imputed GWA data filtered on maf 1% and info score >0.8. Imputed data has been converted from dosage to best guess genotypes

We provide a 1000G imputation protocol for cohorts that haven’t done imputation. We also have a guide to convert imputed data to best guess genotypes.

## Data for pipeline

* **Datasets:** Include cohorts that have 450k methylation arrays and GWA data available in at least 100 individuals.
* **Ancestry:** Europeans only
* **Case-control cohorts:** We would like to analyse cases and controls separately and there should be at least 100 individuals in each group.
* **Tissue:** Whole blood, white cells/buffy coats, peripheral blood lymphocytes, cord blood, purified CD4/CD8, placenta

**Not included:** Buccal cells, LCLs, Carotid plaque tissue

* **Imputation:** 1000 genomes imputed SNP data (build 37)
* **Methylation normalized:** normalized beta values, preferable normalized with meffil
* **Optional: EWAS phenotypes:** height and BMI
* **Optional GWAS phenotypes:** cell counts
* **Covariates:** 
  + Age at measurement
  + Sex
  + Optional: smoking
  + Optional: Batch data (Slide, Plate etc.)

## Hardware

Unix cluster of some sort

# Analysis schedule

Many of the planned analyses will be contingent on other analyses having already been performed, so it will be necessary to divide the analysis into two rounds.

*i.e.* The analyst will run scripts to perform analyses in round one, send the results back to be processed, and then use these processed results to perform analyses in round two etc.

## Analysis 1

1. Data cleaning and harmonization
2. Cis- and trans-meQTL analysis
3. Cis- and trans-var-meQTL analysis
4. Structural variant analysis
5. Genetics of derived variables (cell counts, age, smoking)
6. EWAS on height and BMI

## Analysis 2

1. Obtain list of candidate meQTLs from all cohorts. Perform association tests on this list using full LMM.

## Meta analysis 1

1. Meta-analyse putative meQTLs from all cohorts
2. Meta-analyse common structural variant associations
3. Meta-analyse AAR, cell counts, smoking GWAS and GREML analysis
4. Meta-analyse EWAS on height and BMI
5. Meta-analyse whole methylome estimation on height and BMI
6. Conditional analysis on all significant meQTLs from meta-analysis
7. Functional annotation

## Analysis 3

1. meQTL GREML analysis of height and BMI
2. Test for genetic stability over age, between sex, across cell counts
3. Perform MR between probes and BMI / height
4. var-meQTL verification using DGLM

## Meta analysis 2

1. Meta-analysis of meQTL GREML analysis of height and BMI
2. Meta-analysis of interactions
3. Meta-analysis of MR analysis

# Analysis details

## Prepare data for pipeline

### Input:

* Best guess genotypes
* Normalized beta matrix
* Structural variants estimated from methylation idat files
* Batch variables
* Age at time of sampling
* Sex
* Height phenotype
* BMI phenotype

### The pipeline will do the following:

**Genotypes:**

* + Filter out SNPs with HWE P<1e-6, maf<1%, snp missingness 5%, sample missingness 5%
  + Make an unique SNP identifier: CHR:POS:{SNP/INDEL}
  + Calculate genetic PCs to adjust population stratification using HapMap3 SNPs without long range LD regions
  + Remove ethnic outliers with 5 SD from the mean of the first 10 principal components
  + Create genetic relationship matrix using HapMap3 SNPs with maf>0.01 in plink1.90
  + If family data, create pedigree matrix using 0.125 relatedness threshold
* If unrelated data, remove cryptic related individuals
  + Create R/matrixeqtl format data and calculate effect alleles

**CNVs:**

* Create R/matrixeqtl format data and calculate effect alleles

**Methylation:**

* 1. Rank transform methylation
* 2. Use (1) to adjust for known covariates, estimated cell counts, predicted smoking pedigree if necessary (take residuals after fitting pedigree matrix, i.e. GRAMMAR method)
* 3. Use (2) to estimate methylation PCs and remove any PCs that are associated with SNPs, BMI or height
* 4. Adjust for non-genetic, non-BMI and non-height associated methylation PCs
* 5. Convert methylation matrix to matrixeqtl format

**GWAS phenotypes**

* + **Cellcounts**
  + Use Houseman reference method to predict 7 cell counts (Bcells, CD4T, CD8T, Neutrophils, Eosinophils, Monocytes, Natural Killer cells)
* This follows the UK10K protocol:
* Examine differences in trait distribution by males and females. Where different, data will be handled in a sex specific manner from this point.
* Outliers greater than 5 SD from the mean are excluded
* Data will be transformed to obtain a normal distribution using an inverse normal rank transformation.
* Age and age+age^2 will be tested for association with phenotypes in the dataset. Adjustment for covariates will only be undertaken given evidence of association between covariate and phenotype.
* Traits were residualised on associated covariates to generate standardised residuals with a mean of zero and a SD of 1.
  + Where a sex-specific transformation was used, females and males were standardized separately before being combined.
* **Smoking**
  + Use Illig *et al.* smoking associations to predict smoking status to be used as covariate in further analysis
  + **Age prediction**
  + Create age predictor using method by Steve Horvath.

**EWAS phenotypes:**

* This follows the UK10K protocol:
* Examine differences in trait distribution by males and females. Where different, data will be handled in a sex specific manner from this point.
* Outliers greater than 5 SD from the mean are excluded
* Data will be transformed to obtain a normal distribution using an inverse normal rank transformation.
* Age and age+age^2 will be tested for association with phenotypes in the dataset. Adjustment for covariates will only be undertaken given evidence of association between covariate and phenotype.

Traits were residualised on associated covariates to generate standardised residuals with a mean of zero and a SD of 1.

* + Where a sex-specific transformation was used, females and males were standardized separately before being combined.

**Covariates**

* + Create covariate file (age, sex, predicted smoking, predicted cell counts, genetic PCs, methylation PCs.)

### Output:

* Filtered and QC’d 1000 genomes SNP data
* Genetic structural variant data
* QC’d methylation data
* Covariate and age and sex-adjusted EWAS and GWAS phenotype data

## Cis- and trans-meQTL analysis

### Background

A comprehensive catalogue of SNP-methylation associations can be used to understand the molecular determinants of methylation, and for downstream analyses in complex traits, such as Mendelian randomisation, functional annotations etc.

### Strategy

Performing hundreds of thousands GWASs will result in a huge computational burden if using standard tools, and will also require huge amounts of disk space to store all results. We would like to minimise computational time and data storage whilst also making accurate effect size estimates and reduce possibility of ‘publication bias’ in the meta analysis. We also need to account for heterogeneous study designs, some being from stratified but unrelated samples while others having family relatedness within the data.

The analysis will be performed in two stages.

**Round 1**

First, probes will be residualised for covariates and pedigree structure and R/matrixeqtl will be used to perform a fast, comprehensive analysis of all cis- and trans-associations on residualised probes. A list of all associations above a soft threshold **1e-4** will be sent to a central SFTP in Bristol.

**Round 2**

A unique list of all SNP-CPG pairs from **all cohorts** will be collated and sent back to each analyst where effect sizes for all putative associations will be recalculated in all cohorts using a **linear mixed model**, properly accounting for population stratification, family structure and covariates. Results from this step will be used for the final meta analyses and conditional analyses.

This is a slightly long winded approach but we think it is the best way to achieve a comprehensive analysis that minimises publication bias whilst remaining computationally tractable.

### Input:

* Rank transformed methylation data
* Pedigree adjusted methylation data if familial data for round 1
* GRM constructed using HapMap3 SNPs
* Filtered 1000 genomes imputed SNP data
* Covariate file

### Process:

* Use **R/matrixeqtl** software to perform trans-meQTL analysis of all probes against all SNPs. A genome-wide threshold of 1e-4 will be used to take forward to meta-analysis
* Use **FastLMM/GCTA** to perform follow up analysis on all putative associations. A lmm model will be used for all cohorts (related and unrelateds)
  + **Covariates** to be fitted:
    - age
    - cell counts (predicted)
    - sex
    - kinships as random effect
    - non-genetic, non-bmi and non-height associated methylation PCs

### Output:

* Round 1 : all associations at 10-4
* Round 2: all tested associations + all cis associations

## Cis- and trans-var-meQTL analysis

### Background

It is common to map the genetic variants that influence mean methylation, but many evolutionary mechanisms, such as hedge-betting and adaptive noise, depend on changes in variance. We can use these data to estimate the extent to which the variance in methylation can be influenced by SNPs. See Yang et al (2012).

### Strategy

The same strategy as for the meQTL analysis but with the following differences:

1. Use squared methylation residuals instead of standard normalised methylation data for R/matrixeqtl analysis
2. Use double generalised linear model to fit the full model in the second step to estimate the variance effect orthogonal to the main effect

### Input:

* Rank transformed methylation data
* Pedigree adjusted methylation data if familial data for round 1
* GRM constructed using HapMap3 SNPs
* Structural variant calls
* Covariate file

### Process:

* Use **R/matrixeqtl** software to perform trans-meQTL analysis of all probes against all SNPs. A genome-wide threshold of 1e-4 will be used to take forward to meta-analysis
* Use **FastLMM/GCTA** to perform follow up analysis on all putative associations incl cohorts of unrelateds
  + **Covariates** to be fitted:
    - age
    - cell counts (predicted)
    - sex
    - kinships as random effect
    - methylation PCs
    - non-genetic, non-bmi and non-height associated methylation PCs

### Output:

* Round 1: all associations at 10-4
* Round 2: all tested associations + all cis associations

## Structural variant analysis

### Background

Most genetic analyses focus on the association between SNP and trait, but with raw intensity data available from SNP arrays we can make inference on genomic structural changes, identify common features, and estimate the extent to which they influence methylation levels. See clonal mosaicism papers by Jacobs *et al* (2012) and Laurie *et al* (2012).

### Input:

* Rank transformed methylation data
* Pedigree adjusted methylation data if familial data for round 1
* GRM constructed using HapMap3 SNPs
* Structural variant calls
* Covariate file

### Process:

* Use **R/matrixeqtl** software to perform trans-meQTL analysis of all probes against all SNPs. A genome-wide threshold of 1e-4 will be used to take forward to meta-analysis.
* Use **FastLMM/GCTA** to perform follow up analysis on all putative associations incl cohorts of unrelateds

**Covariates** to be fitted:

* + - age
    - cell counts (predicted)
    - sex
    - kinships as random effect
    - non-genetic, non-bmi and non-height associated methylation PCs

### Output:

* Round 1 : all associations at 10-4
* Round 2: all tested associations + all cis associations

## Genetics of derived variables

### Background

Age can be predicted with high accuracy using a relatively small number of methylation probes. It has been hypothesised that divergence of methylation predicted age and chronological age may be associated with disease and other phenotypic outcomes, including mortality. Estimating the heritable component of age acceleration and mapping SNP associations may reveal insight into these findings. We can also do the same for cell type proportions and predicted smoking status.

### Input:

* 1000 genomes imputed SNP data
* Phenotype file with predicted smoking, predicted age, (predicted) cell counts
* Covariate files wit age, sex and predicted smoking

### Process:

* Use **FastLMM/GCTA** to perform LMM GWAS analysis. Use GCTA to perform GREML analysis. If there are relateds in the sample use Zaitlen et al (2013) method

**Phenotypes for GWAS AAR**

* predicted age use Horvath method

**Covariates for GWAS AAR**

* sex
* grm

**Phenotypes for GWAS smoking**

* There are a number of different ways to approach this, but it is a relatively imprecise variable (predicted smoking) that might have different interpretations in different cohorts. The proposed ways of analyzing the variable are:
* 1. continuous variable
* 2. below a score of 5 vs above a score of 5
* 3. below 10 vs above 10
* 4. Below 5 vs above 10
* 5. cluster smoking with k-means in two clusters

Analysis are done twice: on all subjects and subjects >25 years.

**Covariates for GWAS smoking**

* sex
* age
* grm

**Phenotypes for cell counts**

* + - * 1. gwas of each cell type proportion (i.e. 7 GWASs: Bcells, CD4T, CD8T, Neutrophils, Eosinophils, Monocytes, Natural Killer cells)
* 2. gwas for ‘cell diversity’ e.g. each individual is characterised as the shannon entropy of the row - each row represents the probabilities of the cell types, shannon entropy = -sum(p \* log(p)). This is an attempt to look for genetic markers that influence cell type variance.

**Covariates for GWAS cell counts:**

* sex
* age
* grm
* predicted smoking

### Output:

* SNP associations
* SNP heritability estimates

## EWAS on height and BMI

### Background

Linking methylation to complex traits is one of the broad goals in generating methylation data from large numbers of samples. We will use height and BMI as they are good models for complex traits, have interesting aetiologies, and will typically be available in all cohorts.

### Input:

* Phenotype file with age and sex-specific inverse normalized residuals
* Rank transformed methylation data
* Covariate files

### Process:

* Perform standard EWAS, and also use bumphunting and A-clustering methods
* Covariates to be fitted:
  + batch
  + cell counts (predicted)
  + smoking (predicted)
  + first 10 genetic principal components
  + non-genetic, non-bmi and non-height associated methylation PCs
  + GRM

### Output:

* Putative CpG-trait associations for BMI and height

**Generating candidate association list**

### Background

Each cohort will search for meQTLs by analyzing the ‘entire surface’ of SNPs vs methylation probes. However, saving all of these results is impractical. The issue is that the set of SNP-CpG pairs that reach significance at, say 1e-4 in cohort A will be a different set to cohort B, so performing meta analysis on only these initial results will lead to a loss of potential meQTLs. To avoid this problem the meQTL analysis will take on a second stage. Results from the first stage from each cohort will be uploaded to Bristol. We will then generate a list of all unique reported SNP-CpG pairs, creating a ‘candidate list’. The candidate list will be sent to each cohort and they will test each candidate for association. The advantages of this approach are:

* + - 1. We can get a much greater coverage of meQTLs with small effects in the meta analysis
      2. A full LMM can be used to provide final estimates of effect sizes because the candidate list will be much smaller (and therefore computationally tractable) than the entire surface
      3. We can specify putative cis-meQTLs for a particular probe to be used as covariates when analyzing trans-meQTLs, which should improve power.

### Input

* Results from meQTL analysis

### Process and output

* Identify unique list of SNP-CpGs from all cohorts

**Meta-analyse putative meQTLs from all cohorts**

### Background

There are several methods for meta-analysing across cohorts, for example see GWAMA software (Magi et al 2010). This analysis should identify signals with smaller effects that are not significant in any one study but are significant across all studies. We will only be able to keep a fraction of all SNP-CpG associations due to memory and transfer limitations (e.g. threshold cutoff 1e-4 should keep the transfer down to below 300MB), so it is possible that significant associations for which we do have power to detect will be missed (e.g. all studies show an effect with p = 1e-3).

### Input:

* meQTL summary statistics from all cohorts: SNV identifier, CHR, POS, EA, NEA, EAF, N, NO, N1, N2, CPG, HWE P, Call rate, BETA, SE, PVAL, IMPUTED, INFO TYPE, INFO.

### Process:

* Perform inverse variance fixed effects meta-analysis for all SNP-CpG pairs present in multiple cohorts
* We will use GWAMA to conduct meta-analysis

### Output:

List of nominally significant meQTLs to take forward for conditional analysis and replication

## Meta-analyse common structural variant associations

As described in “Meta-analyse putative meQTLs from all cohorts”

## Meta analyse GWAS of derived variables

Standard GWAS meta analysis procedures for AAR, predicted smoking status, predicted cell counts.

### Input:

* GWAS summary statistics from all cohorts: SNV identifier, CHR, POS, EA, NEA, EAF, N, NO, N1, N2, CPG, HWE P, Call rate, BETA, SE, PVAL, IMPUTED, INFO TYPE, INFO.

### Process:

* Perform inverse variance fixed effects meta-analysis for all GWAS present in multiple cohorts
* We will use GWAMA to conduct meta-analysis
* Sensitivity analysis (adjustment for smoking/cell counts)

## Conditional analysis on all significant meQTLs

### Background

Take the results from the cis- and trans-meQTL meta analysis and identify independent signals that may in close proximity to sentinel SNPs.

### Input:

* meQTL meta analysis results
* Reference genotype panel (e.g. ALSPAC)

### Process:

* Use GCTA to perform conditional analysis for each probe with a significant association

### Output:

* Independent associations for each methylation probe

## meQTL GREML analysis of height and BMI

### Background

GCTA-GREML analysis can be used to estimate the variance explained by all SNPs, and it can also be used to partition the genetic variance to different components. Here we can use the list of cis-meQTLs to construct one component and use HapMap3 SNPs to construct another component and fit them together to estimate the proportion of variance of complex traits that may be mediated through methylation

### Input:

* 1000 genomes data in each cohort
* List of significant meQTLs
* HM3 GRM

### Process:

* Construct GRM from meQTLs
* Fit 2 component GREML model

### Output:

* Estimate of each variance component from each cohort to be meta analysed when sent back

## Test for genetic stability over age, between sex, across cell counts

### Background

Genetic associations may change under certain conditions, for example there may be sex-specific, age-specific, or cell type-specific effects. We can test this by fitting interactions between age/sex/cell-type and genotype. For age and sex this is straightforward, and can be easily adapted to cell counts also. For example, with proportion estimates for 6 different cell types, measuring the interaction term between genotype and each cell-type estimate will capture whether a genotypic effect is specific to a particular cell type. e.g. sample R code:

# Create cell counts

x <- matrix(rnorm(5\*1000), 1000)

X <- t(apply(x, 1, function(x){ x <- x-min(x); return(x/sum(x))}))

# Create SNP

g <- rbinom(1000, 2, 0.5)

# Create methylation probe with effect only present in cell type 1

y <- g\*X[,1] + rnorm(1000)

# Test for interaction

summary(lm(y ~ g \* X))

### Input:

* SNPs and CpGs with significant mQTLs as detected from conditional analysis of meta-analysed meQTL results
* Covariate file

### Process:

* Use R functions to run statistical analysis of interactions between meQTLs and age/sex/cell-type proportions
* Include following covariates:
  + batch
  + age
  + cell counts (predicted)
  + smoking (predicted)
  + sex
  + first 10 genetic principal components

### Output:

* Interaction effects for each meQTL / mediator pair

## 

## MR between probes and BMI / height

### Background

One of the major applications for a catalogue of meQTLs is to use them as instruments to test for causality in associations between methylation and complex traits. For both height and BMI there exists a large number of associations (Wood *et al* 2014; Speliotes *et al* 2010) that could potentially be used as instruments, though genetic confounding may lead to erroneous causal inference and this should be tested.

### Input:

* 1000 genomes SNP data
* Normalised methylation data
* Covariates file
* List of CpG - trait associations from EWAS

### Process:

* Perform MR analysis for all CpG - trait associations.
* Perform bi-directional MR for all CpG - trait associations for which the CpG has a cis-acting meQTL

### Output:

* Causal inference estimates for CpG - BMI and CpG - height associations