# Properties of epigenome-wide association studies

Thomas Battram [1]*, et al.
[1] MRC Integrative Epidemiology Unit, University of Bristol

*Corresponding author: thomas.battram@bristol.ac.uk

## Rationale

Epigenome-wide association studies (EWAS) are the most commonly used study design in epigenetic epidemiology and have been around for over 10 years. There are now publically available databases providing published EWAS data which can be downloaded and examined. Analysing the data jointly allows the discovery of commonalities across methylome-trait associations and provides a platform to explore what is driving these commonalities. Further, pooling the data gives an opportunity to assess the robustness of published results by checking replication rate and whether sites measured by unreliable probes are prominent.

## Methods

### Data

All the data from the EWAS Catalog will be extracted. This means published data from EWAS with N > 100 and >100,000 DNA methylation sites measured. Associations are limited to $p < 1x10^{-4}$.

ARIES data will be used for some analyses when checking the properties of DNA methylation sites (see below for more detail).

### Data cleaning

When comparing traits and DNA methylation sites discovered, if there are multiple EWAS of the same trait, the trait with the largest N will be used.

When multiple models are present, the most basic model will be used for primary analyses and any conclusions made will be checked by replacing these with the "complete" models as a supplementary analysis.

When tissue type may influence results then whole blood will be used as it is the most common tissue type.

An example of how the data will initially be described is at the start of the results section.

### Exploring identified sites

### Robustness of results

- Percentage of CpGs identified in EWAS that may be problematic
  - Any sites listed in Zhou W. et al. 2017 as problematic will be extracted
  - EWAS that have excluded these sites will be removed
  - Will use Fisher's exact test to determine if a CpG site is more likely to be associated with a trait given it is a problematic CpG.

- What is the replication rate?
  - Extract traits that are present in multiple EWAS and check for overlap in CpG sites
  - Use ARIES data for replication analysis where possible. FOM, FOF and F7 to be used again. Will re-run the analysis adjusting for the covariates used in the original EWAS if possible as well as for cell counts, 20 SVs and 10 PCs.

### Architecture of EWAS results

- Do sites identified at $p < 1x10^{-7}$ tend to vary more?
  - GoDMC have CpG-variability data available
  - Will focus solely on whole blood here as variances likely to vary across tissues
  - Will assess association between effect size and variance – would expect to see effect size increase as variance decreases
  - Will assess association between effect size and average methylation level. It might be that changes in methylation may have more of an effect depending on where the starting point of methylation is. For example, a 5% change in methylation from 100% methylated to 95% may have a larger effect than 25% to 20%.
  - Average methylation level will be extracted from ARIES data across all time points with whole blood and averaged across them.

- Are sites identified at $p < 1x10^{-7}$ enriched for certain regions of the genome or are they enriched for any other epigenetic marks?
  - will do simple enrichment analyses using eFORGE (and maybe LOLA)

- Are sites identified at $p < 1x10^{-7}$ more heritable?
  - Want to see if effects might be in part driven by genetics
  - Extract twin estimates from Hannon et al. 2018
  - Limit to whole blood
  - `hit(y/n) ~ heritability`

**Clustered EWAS**

**THIS WON'T BE LOOKED AT AGAIN UNTIL AFTER THE CATALOG IS COMPLETED SO TECHNICAL DETAILS WILL NOT BE WORKED OUT FOR A WHILE. COULD BE SCRAPPED FROM THIS PROJECT**

Do traits clustered by EWAS results form clusters as expected? (i.e. form similar clusters to those just be assessing phenotype correlations)

- EWAS correlations
  - Matt is working on some all-v-all EWAS correlation stuff.
  - "We are currently exploring various clusterings of this all-against-all comparison to identify robust clusters of phenotypes and exposures with similar associations with DNA methylation"
  - One thing to look into here is how smoking/age EWAS correlate with other EWAS -> Are smoking and age driving associations in other EWAS?

## Results

### Catalog data

In total there are X studies available, studying X traits. These studies have uncovered X trait-DNA methylation site associations at $p < 1x10^{-7}$ which span all autosomes and tag X CpG sites in X genes. These data are summarised in **Tables 1** (an example study data table. On average the associations explained X amount of variance of the trait (**Figure 2**).

Table 1: Description of data present in the EWAS Catalog

| study-trait | value |
|---|---|
| Total EWAS | 600 |
| Total traits | 500 |
| Total N | 100000 |
| Total associations | 1000000 |
| Total CpGs | 100000 |
| Total genes | 10000 |
| Sample size median (range) | 200 (100-7000) |
| Sex (%females) | 50 |
| Populations | EUR=2000, AFR=100 |
| Mean age (range) | 47 (0-80) |
| Tissues | Whole blood, cord blood, CD4+ T cells, skin |

**Figure 2** = histogram of $r^2$ values.

X number of CpGs associated with more than 10 traits (**Figure 3**), with CpG1 (Gene1) associating with the highest number of traits (X).

**Figure 3** = manhattan plot with CpGs on x-axis and number of traits associated with at $p < 1x10^{-7}$ on the y-axis.

### Properties of identified sites

Analysis mentioned in methods.