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

Introduction

- Important to use data out there to learn about studies! Loads of EWAS has been done so it's time we used the data to learn about EWAS!

Learning from the successes and mistakes we make is what drives forward development. Hundreds of epigenome-wide association studies (EWAS) have been conducted in the last 10-15 years (REF), yet few (OR NONE - CHECK THIS OUT), cross-EWAS studies, comparing results across a large group of EWAS results has been performed. By exploring the patterns of association across a large group of EWAS, one can discover potential explanations for the results found, that may shed light on failings in the literature as well as shared epigenetic architectures across traits.

- Problems with EWAS realised since its inception: 1. batch effects, 2. cell heterogeneity, 3. tissue specificity. Also typical problems with observational studies: confounding and reverse causation.
- However, EWAS has also uncovered some interesting findings (EXAMPLE) and there is still lots to learn about the epigenetic architecture of complex traits.
- With the recent advent of collated datasets, we can explore potential failings in the literature as well as characteristics of sites identified.
- In an attempt to understand what has been discovered so far by EWAS, we first describe the data present in the EWAS Catalog before exploring various explanations for the findings.

Methods

EWAS Data

X EWAS summary statistics were extracted from the EWAS Catalog [REF]. This includes X published EWAS, X EWAS performed in the Accessible Resource for Integrated Epigenomics Studies (ARIES) subsection of the Avon Longitudinal Study of Parents and Children (ALSPAC), which is described below and X EWAS performed using data from the gene expression omnibus (GEO) resource.

All EWAS have a sample size of greater than 100 and measured more than 100,000 DNA methylation sites genome-wide.

If there are multiple EWAS of the same trait, the EWAS with the largest N was selected for analyses when wishing to assess EWAS results of unique traits.

ARIES

Data used to derive estimates for DNA methylation variability and average methylation levels was taken from the ARIES subsection of the ALSPAC cohort. Pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were invited to take part in the study. The initial number of pregnancies enrolled is 14,541 (for these at least one questionnaire has been returned or a "Children in Focus" clinic had been attended by 19/07/1999). Of these initial pregnancies, there was a total of 14,676 foetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age. Full details of the cohort has been published previously (REFS). This study uses DNA methylation data from the mothers (N = 940), the fathers (N = X), and the children at age X (N = X) and at age X (N = X).

Code availability

Code used to run the analyses is available here: https://github.com/thomasbattram/something

All analyses were completed using R (version **X**).

Results

Description of the catalog

Before assessing what might be underlying various EWAS results, we present a brief summary of the data in the EWAS Catalog (**Table 1**).

Table 1: Description of data present in the EWAS Catalog

study-trait	value
Number of EWAS	594
Number of traits	531
Number of samples	389527
Median sample size (range)	536 (93 - 13474)
Number of associations	155946
Number of CpGs identified	129670
Number of genes identified	19305
Sex (Studies)	Both (313), Females (403), Males (17)
Ethnicities	EUR (74.6), Unclear (12.9), AFR (4.7), Other (3.7), ADM (1.6), EAS (1.5), SAS (1.0)
Age (Studies)	Adults (569), Geriatrics (91), Children (40), Infants (36)
Number of tissue types	42
Most common tissues (%)	whole blood (83.8), cord blood (4.4), cd4+ t-cells (2.7), placenta (1.3), saliva (1.0)

The number of traits each CpG associated with was fairly even across chromosomes (**Figure 1**). There were four CpGs that associated with more than ten traits, cg01940273 -, cg05575921 *AHRR*, cg00574958 *CPT1A*, cg06500161 *ABCG1*. cg06500161 *ABCG1* associated with more traits than any other site - 49 traits. These correspond mostly to metabolites, weight-related traits, and type two diabetes.

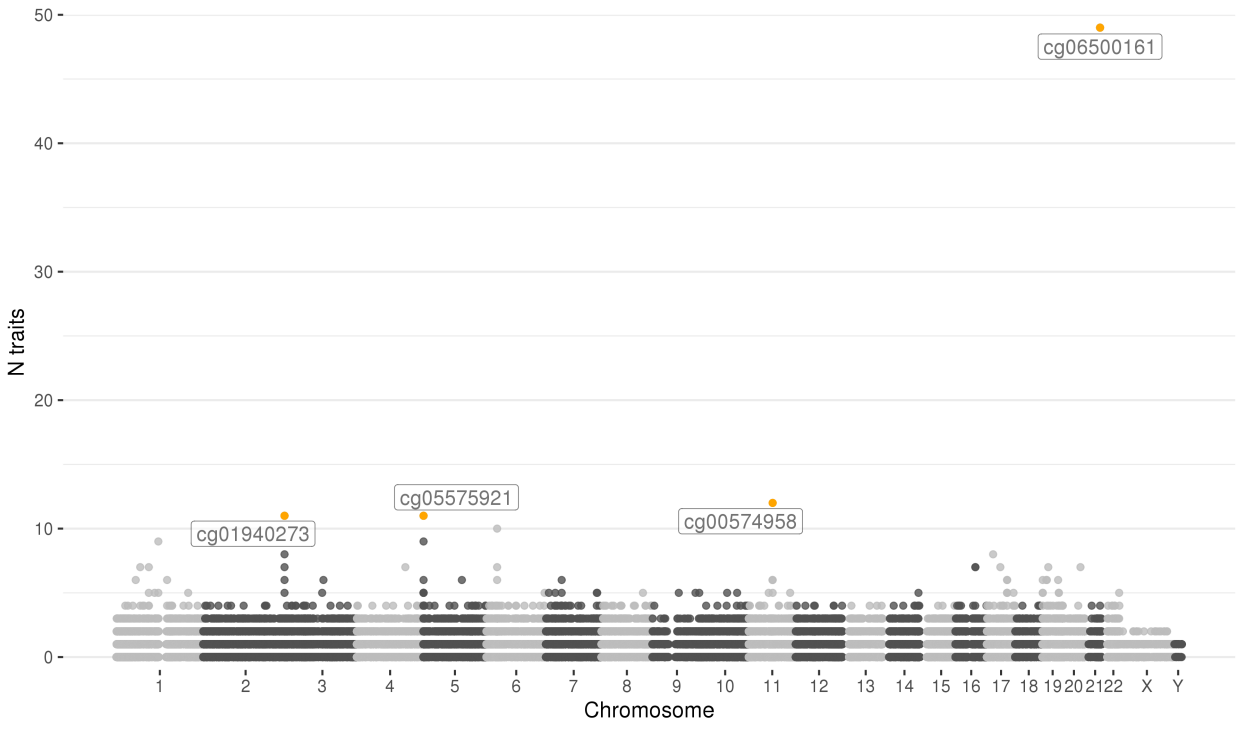

Figure 1: **Number of unique traits associated with each CpG**

The total trait variance correlated with DNA methylation (r^2) at each site varied from 0.0011 to 0.78 (**Figure 2**). There were `sum(rsq_sum_dat$rsq_sum > 1)` traits that had a sum of r^2 values greater than 1.

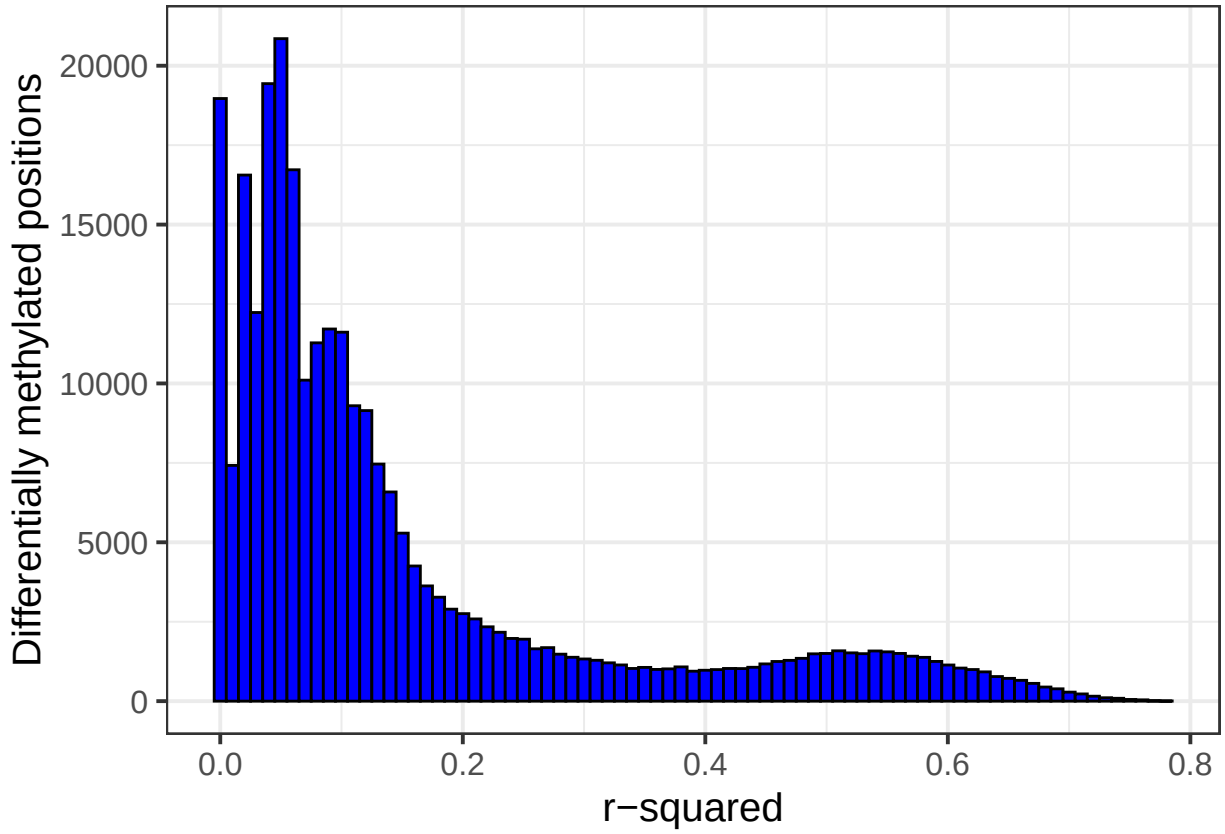

Figure 2: **Distribution of r-squared values across all CpG sites in the EWAS Catalog**

Robustness of results

559 studies adjusted for batch effects in at least one model. Of all DMPs identifed, 9.3% were measured by potentially faulty probes (REF) and an extra 0.64% were present on sex chromosomes (**Figure 3**).

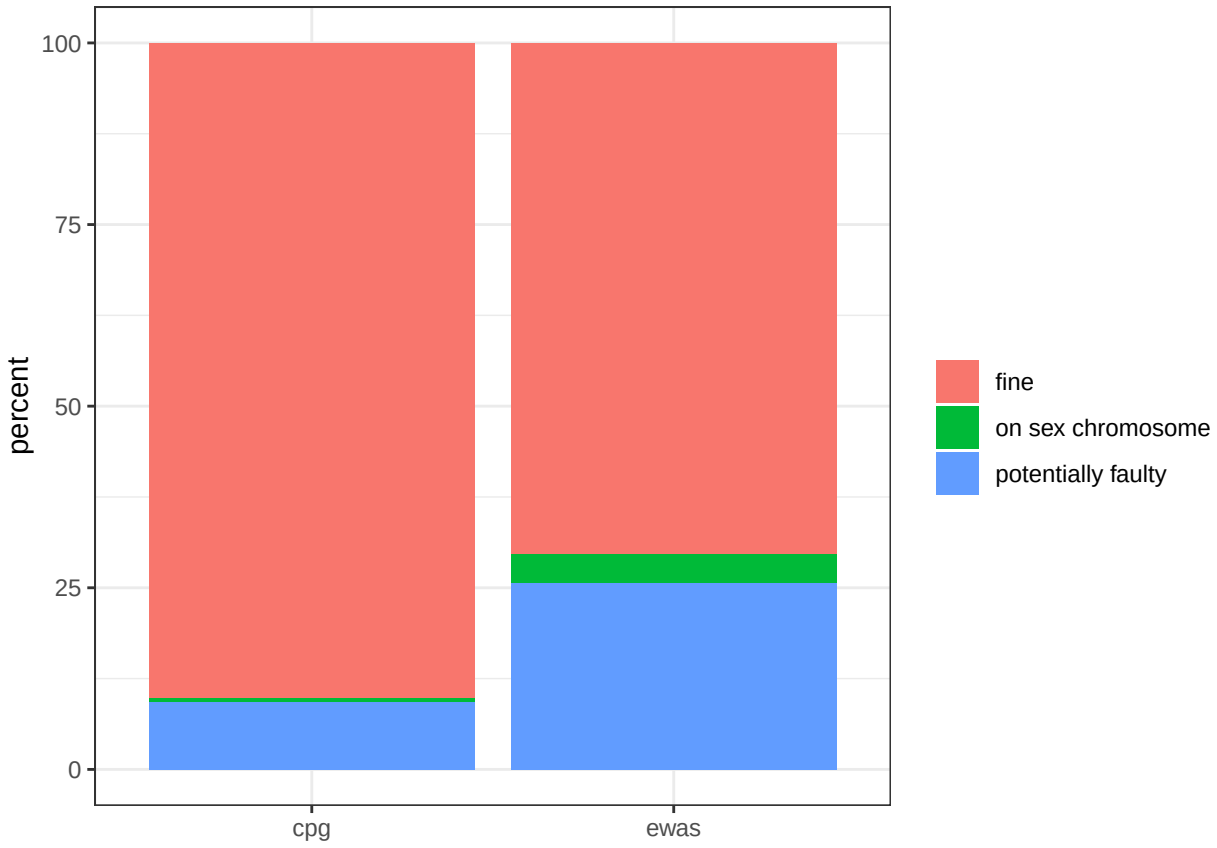

Figure 3: **Potentially unreliable differentially methylated positions identified**

There were 30 studies that performed a meta-analysis of discovery and replication samples. A further 48 studies performed a separate replication analysis. Together, this provides 1666 associations within the EWAS Catalog that have been replicated at P < $1x10^{-4}$. Using the Catalog data we further performed replication analyses. There were studies that shared a common phenotype of interest. Replication rate, judged as any reported CpG with P value < $1x10^{-4}$ in another study of the same trait, varied from 0 to 1 between studies (**Supplementary table X**).

Smoking is associated with large changes in DNA methylation across the genome (REF) and is associated with many different traits (REF). Thus, it may confound DNA methylation associations found in the catalog. If this was the case, one might expect smoking related CpGs to appear more in the catalog than expected by chance. The DMPs identified by EWAS of traits other than smoking were enriched for smoking related CpG sites (P = X).

Correlation across tissues

- There is correlation between DNA methylation sites across tissues, suggesting stability in DNAm, but would we expect to find associations at these positions across tissues?
- Replication of sites across tissues for same traits?
- Do sites that are highly correlated across tissues appear more than expected by chance in whole blood EWAS?
  - If yes then suggests the correlation might be due to decreased measurement error at those sites

  - If the opposite (correlated sites appear less than expected by chance), then it suggests that correlation occurs at positions that don't really matter (i.e. are just stable because of things like being at housekeeping genes)

- Can check housekeeping gene theory if needed

CpG characteristics

Faulty probes and bias may give rise to EWAS signal, but the characteristics of DNA methylation at CpG sites likely also contributes to trait-DNA methylation associations.

- Effect size and variation (**Figure 4**)

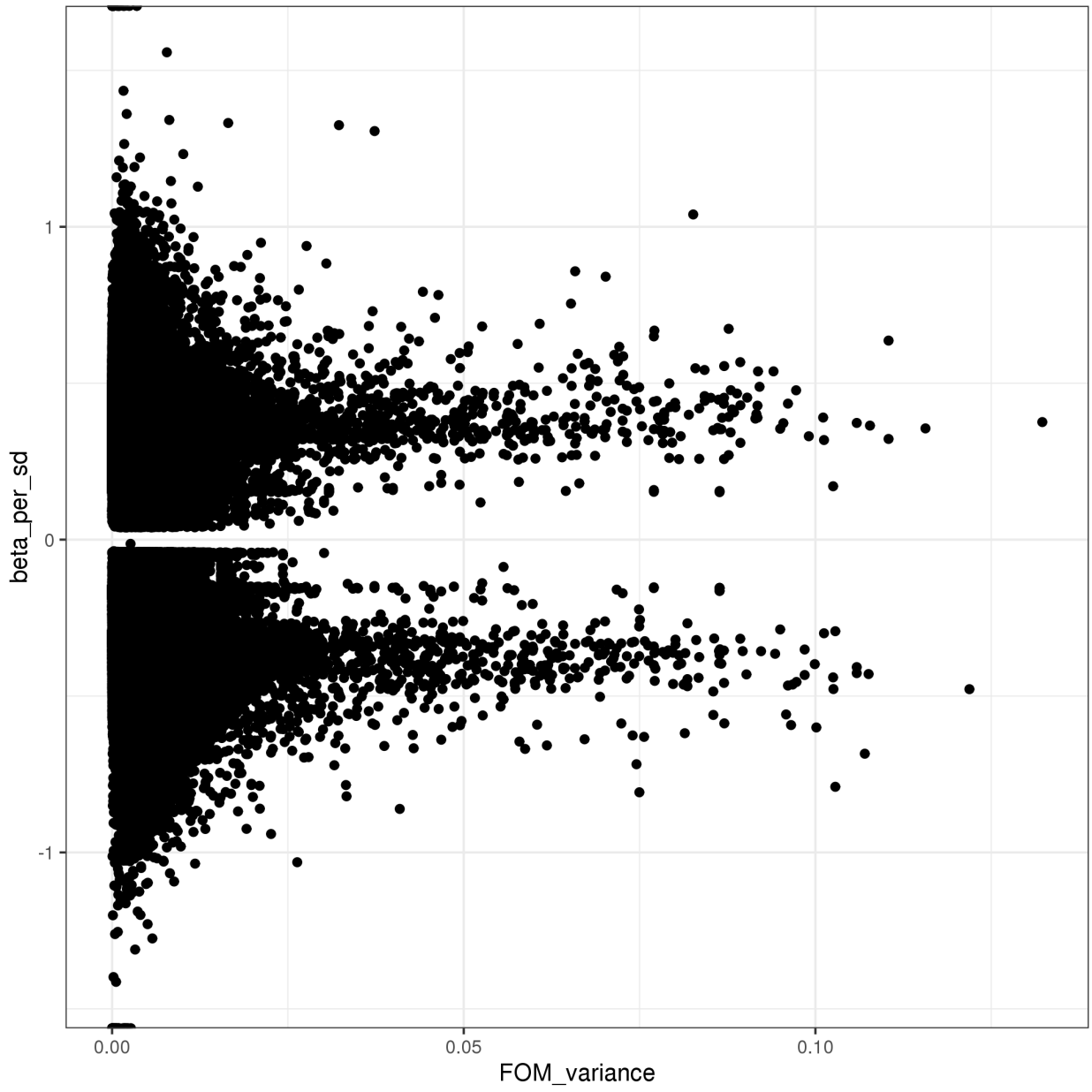

Figure 4: **Relationship between effect size of DNA methylation-trait associations and the variability of DNA methylation at CpGs**

- Effect size and avg methylation level (**Figure 5**)

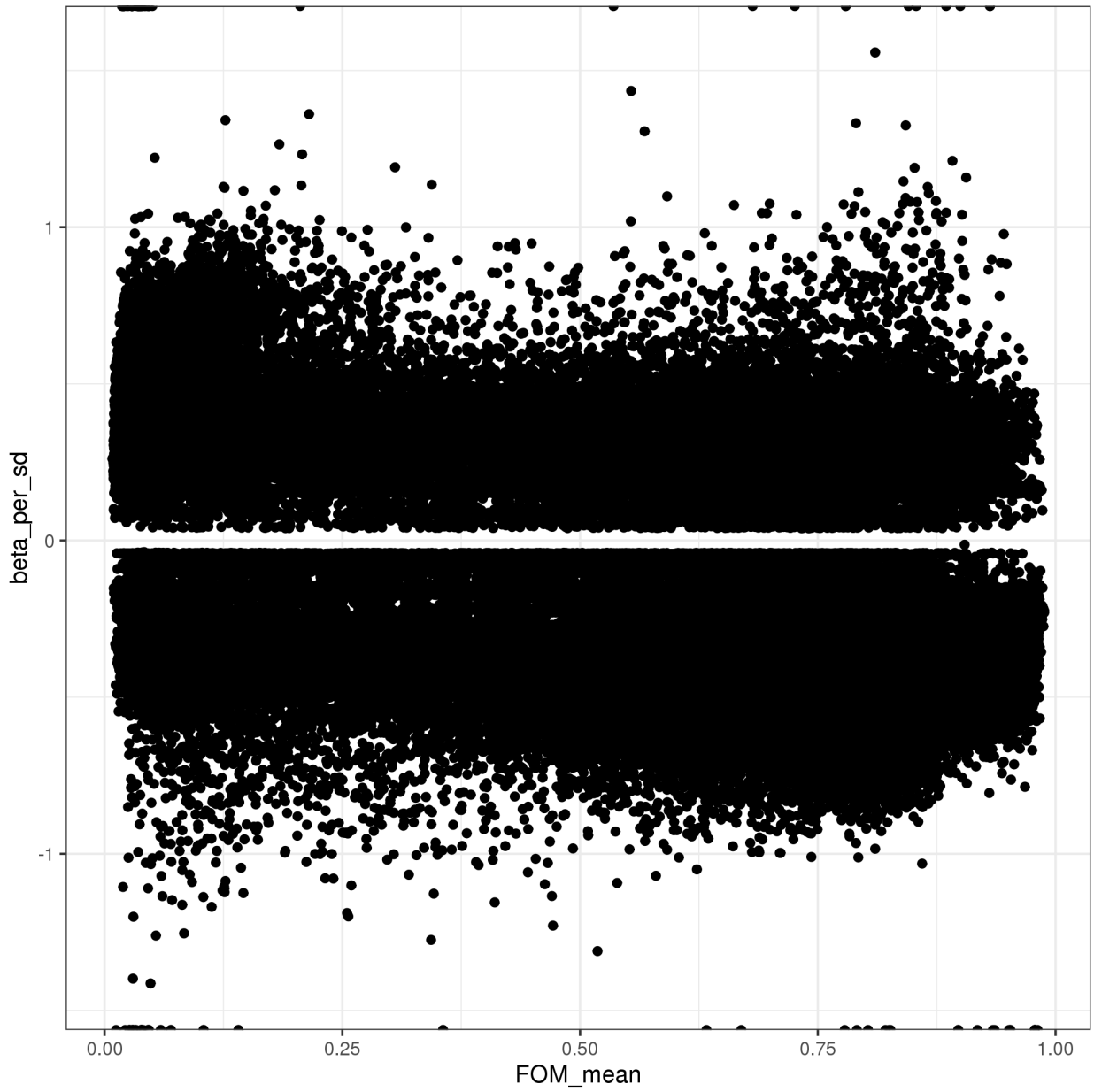

Figure 5: **Relationship between effect size of DNA methylation-trait associations and the average DNA methylation at CpGs**

- Effect size and heritability (**Figure X**)
- Enrichment of DMPs in various genomic regions etc.

Discussion

A discussion subheading

Limitations

Conclusion

Overall this paper is the best and should be accepted in Nature, Science, NEJM and JAMA all at the same time.

References