Chapter 3

The EWAS Catalog: a database of epigenome-wide association studies

3.1 Abstract

In recent years, the increase in availability of DNA methylation measures in population-based cohorts and case-control studies has resulted in a dramatic increase in the number of EWAS being performed and published. To make this rich source of molecular data more accessible, a manually curated database has been made containing CpG-trait associations (at P < 1x10^{-4}) from published EWAS, each assaying over 100,000 CpGs in at least 100 individuals. The database currently contains these associations from 178 published EWAS as well as full summary statistics for over 180 million association tests of 427 EWAS in the Avon Longitudinal Study of Parents and Children (ALSPAC)

and the Gene Expression Omnibus (GEO). It is accompanied by a web-based tool and R package that allow these associations to be easily queried. This database provides a platform for analyses in Chapter 4 and 6. Further, it will give other researchers the opportunity to quickly and easily query EWAS associations to gain insight into the molecular underpinnings of disease as well as the impact of traits and exposures on the DNA methylome. The EWAS Catalog is available at: `http://www.ewascatalog.org`.

3.2 Introduction

Epigenome-wide association studies (EWAS) aim to assess the associations between phenotypes of interest and DNA methylation across the genome (49,66,125). These associations may then be used for disease diagnosis or prediction (49,66,125). Also, unlike genetic variants, changes in DNA methylation are responsive to the environment and so may be targeted for treatment. EWAS of smoking (68), body mass index (BMI) (70) and aging (75) have shown that various exposures are related to large perturbations in DNA methylation across the genome. Furthermore, a paper recently estimated that over 60% of the total proportion of BMI variation was captured by DNA methylation at about 150 CpG sites (126). In recent years, there has been a dramatic increase in the number of EWAS being performed and published due to technological advancements making it possible to measure DNA methylation at hundreds of thousands of CpG sites cheaply and effectively. Giving researchers easy access to EWAS outputs will help them gain insight into the molecular underpinnings of disease as well as the impact of traits and exposures on the DNA methylome. Furthermore, current collections of summary statistics have already proven useful to various fields, for example the GWAS Catalog (94) has been cited over 2000 times in papers contributing to new methods and exploring the genetic architecture of a plethora of traits.

At the time of making the database, to our knowledge, there were no databases that had collected well-curated EWAS on all traits (not just diseases) in an online database accessible to researchers. During production one database fulfilled those metrics: EWAS Atlas (97). Other databases are available but are limited to certain diseases (e.g. MethHC (127)). The EWAS Atlas provides a simple-to-use website with annotated CpG sites and information on traits. Ideally a database of EWAS results will provide summary statistics, including betas, standard errors and p-values where provided from publications, in an easily accessible manner, this enables researchers to explore various aspects of the published data without having to retrieve the published article. For example, researchers might compare effect estimates between studies in the database or check to see if their results are replicated in another published study. At the time of writing, the EWAS atlas platform did not enable users to download effect estimates and standard errors. Another caveat is that there is currently only published data on the platform, not full summary statistics from EWAS.

The EWAS Catalog aims to improve upon current databases to 1) allow easy and programmatic access to summary statistics for downstream analyses by researchers and 2) provide full summary statistics from a range of EWAS conducted in multiple cohorts. To this end we have produced The EWAS Catalog, a manually curated database of currently published EWAS, 387 EWAS performed in the Avon Longitudinal Study of Parents and Children (ALSPAC) (128,129) and 40 EWAS performed from data from the Gene

Expression Omnibus (GEO) database. The process and data inclusion are summarised in **Figure 3.1**.

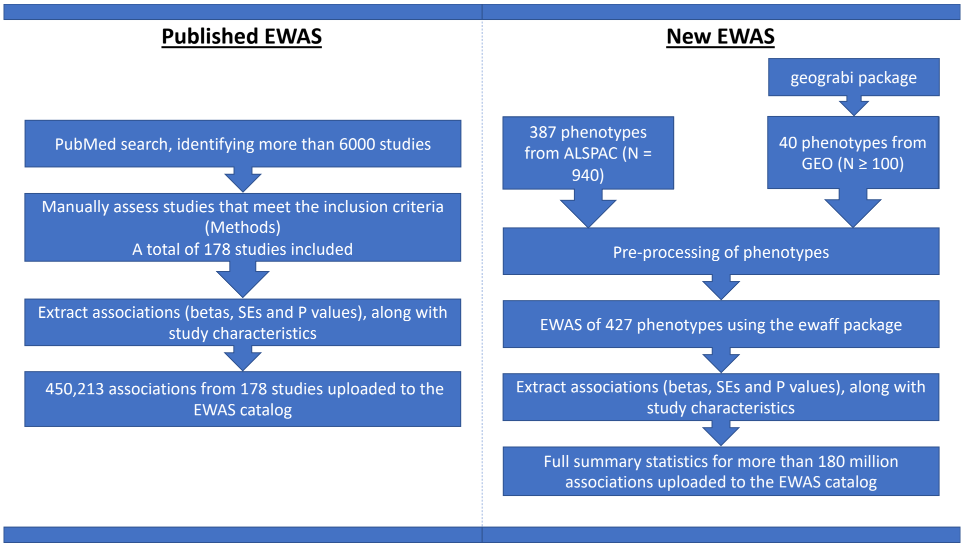

Published EWAS
PubMed search, identifying more than 6000 studies
Manually assess studies that meet the inclusion criteria (Methods)
A total of 178 studies included
Extract associations (betas, SEs and P values), along with study characteristics
450,213 associations from 178 studies uploaded to the EWAS catalog
New EWAS
geograbi package
387 phenotypes from ALSPAC (N = 940)
40 phenotypes from GEO (N ≥ 100)
Pre-processing of phenotypes
EWAS of 427 phenotypes using the ewaff package
Extract associations (betas, SEs and P values), along with study characteristics
Full summary statistics for more than 180 million associations uploaded to the EWAS catalog

Figure 3.1: **EWAS Catalog project flowchart**. On the left is a brief description of how the CpG-phenotype associations were assembled from published works and on the right is a brief description of the EWAS performed using individual level data.

I am not responsible for all the work presented in this chapter. Dr James Staley built the original website. Dr Matthew Suderman has been key in development, and maintenance of the website. Dr Paul Yousefi extracted data from GEO. All three also provided (and continue to provide) expert knowledge to adapt the database to optimise user experience. There was also a team to help gather and input the data. I helped develop and maintain the website, gather and input the data, ran all the EWAS using data from the ALSPAC cohort and the GEO database. The team, led by myself, is continuing to develop and maintain the database. Full acknowledgements to the team can be found on the website: `http://www.ewascatalog.org/about/`.

3.3 Methods

3.3.1 Implementation

The EWAS Catalog web app was built using the Django Python package (`https://djangoproject.com`). The data is stored in a combination of MySQL databases and fast random access files (67) and can be queried via the web app or the R package (www.github.com/ewascatalog/ewascatalog-r/).

3.3.2 Overview of publication data extraction

To identify publications, periodic literature searches are performed in PubMed using the search terms: "epigenome-wide" OR "epigenome wide" OR "EWAS" OR "genome-wide AND methylation" OR "genome wide AND methylation".

Our criteria for inclusion of a study into The EWAS Catalog are as follows:

1. The EWAS performed must contain over 100 humans
2. The analysis must contain over 100,000 CpG sites
3. The DNA methylation data must be genome-wide
4. The study must include previously unpublished EWAS summary statistics

CpG-phenotype associations are extracted from studies at $P < 1x10^{-4}$. All these criteria along with the variables extracted are documented on the website (www.ewascatalog.org/documentation). Experimental factor ontology (EFO) terms were mapped to traits to unify representation of these traits. These EFO terms were manually entered after looking up the trait in the

European Bioinformatics Institute database (www.ebi.ac.uk/efo).

Based on these criteria, from 2020-10-19, The EWAS Catalog contained 450213 associations from 605 studies.

3.3.3 Overview of GEO data extraction

To recruit additional datasets suitable for new EWAS analysis, the geograbi R package (`https://github.com/yousefi138/geograbi`) was used to both query GEO for experiments matching The EWAS Catalog inclusion criteria (described above) and extract relevant DNA methylation and phenotype information. The query was performed by Dr Paul Yousefi on 20 March 2019 and identified 148 such experiments with 32,845 samples where DNA methylation and phenotype information could be successfully extracted. From these, the aim was to repeat the analyses performed in the publications linked by PubMed IDs to each GEO record. Thus, I looked up the corresponding full texts for each dataset and identified the main variables of interest. Of the 148 putative GEO studies, only 34 (23%) contained sufficient information to replicate the original analysis.

3.3.4 EWAS methods

Avon Longitudinal Study of Parents and Children (ALSPAC)

EWAS were conducted for 387 continuous and binary traits in peripheral blood DNA methylation of ALSPAC mothers in middle age (N = 940), generated as part of the Accessible Resource for Integrated Epigenomics Studies (ARIES) project (130).

ALSPAC recruited pregnant women in the Bristol and Avon area, United

Kingdom, with an expected delivery date between April 1991 and December 1992 (`http://www.bris.ac.uk/alspac/`). Over 14,000 pregnancies have been followed up (both children and parents) throughout the life-course. Full details of the cohort have been published previously (128,129). The EWAS performed for the EWAS catalog were done so using phenotypic and DNA methylation data from the mothers (N = 940). All continuous and binary phenotypes were extracted from the same timepoint that blood was drawn for DNA methylation assays.

Ethical approval for ALSPAC was obtained from the ALSPAC Ethics and Law Committee and from the UK National Health Service Local Research Ethics Committees. Written informed consent was obtained from both the parent/guardian and, after the age of 16, children provided written assent. The study website contains details of all the data that is available through a fully searchable data dictionary (`http://www.bristol.ac.uk/alspac/researchers/access/`).

Study data were collected and managed using REDCap electronic data capture tools hosted at ALSPAC (131,132). REDCap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies, providing 1) an intuitive interface for validated data capture; 2) audit trails for tracking data manipulation and export procedures; 3) automated export procedures for seamless data downloads to common statistical packages; and 4) procedures for data integration and interoperability with external sources.

Ancestry principal components were generated within ALSPAC mothers using PLINK (v1.9). Before analysis, genetic data went through quality control and were imputed as follows.

Mothers were genotyped using the Illumina human660W-quad genome-wide single nucleotide polymorphism (SNP) genotyping platform (Illumina Inc., San Diego, CA, USA) at the Centre National de Génotypage (CNG; Paris, France). SNPs were removed if they displayed more than 5% missingness or a Hardy-Weinberg equilibrium P value of less than 1.0e-06. Additionally, SNPs with a minor allele frequency of less than 1% were removed. Samples were excluded if they displayed more than 5% missingness, had indeterminate X chromosome heterozygosity or extreme autosomal heterozygosity. Samples showing evidence of population stratification were identified by multidimensional scaling of genome-wide identity by state pairwise distances using the four HapMap populations as a reference, and then excluded. Cryptic relatedness was assessed using a IBD estimate of more than 0.125 which is expected to correspond to roughly 12.5% alleles shared IBD or a relatedness at the first cousin level. Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation.

Imputation of mother's genotype data in ALSPAC was done with ALSPAC children's data. So, genotypes in common between the sample of mothers and sample of children were combined. SNPs with genotype missingness above 1% due to poor quality were removed along with subjects due to potential ID mismatches. We estimated haplotypes using ShapeIT (v2.r644) which utilises relatedness during phasing. We obtained a phased version of the 1000 genomes reference panel (Phase 1, Version 3) from the Impute2 reference data repository (phased using ShapeIt v2.r644, haplotype release date Dec 2013). Imputation of the target data was performed using Impute V2.2.2 against the reference panel (all polymorphic SNPs excluding singletons), using all 2186 reference haplotypes (including non-Europeans).

After quality control and imputation, independent SNPs ($r^2 < 0.01$) were used to calculate the top 10 ancestry principal components.

For all traits, linear regression models were fitted with DNA methylation at each site as the outcome and the phenotype as the exposure. DNA methylation was coded as beta values between 0 and 1. For a particular site, a beta value of 0 represents no methylation being detected in all cells measured and a value of 1 represents all cells being methylated at that site. Covariates included age, the top 10 ancestry principal components, and 20 surrogate variables.

Transforming phenotypic data

Values of continuous phenotypes were defined as outliers and set to missing, then all phenotypes with 100 or more non-missing values were kept for further analysis. To ensure all phenotypes were approximately normal, each of their distributions were examined and then transformed. If a variable was deemed right-skewed, it was log-transformed then its distribution was re-assessed by eye. Square-roots and cube-roots were used to try and approximate normality if log-transformation did not work. If a variable was deemed left-skewed, it was squared, and the distribution re-assessed by eye.

EWAS statistical models

For all traits, linear regression models were fitted with DNA methylation as the outcome and the phenotype as the exposure as for the ARIES data. Twenty surrogate variables were included as covariates. Other covariates were considered, but surrogate variables only were used for two reasons: 1) to help automate the process and 2) because covariates used in the original

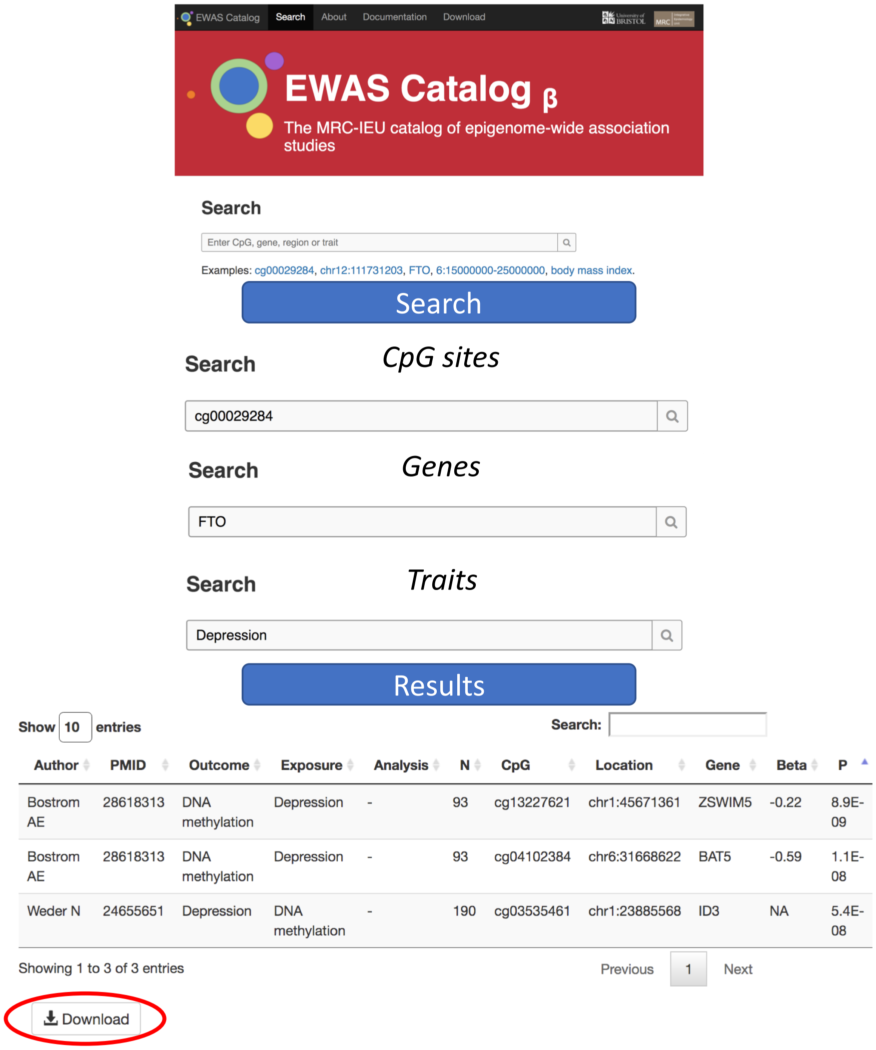
EWAS were not included with many of the GEO datasets.

Statistical analyses were conducted in R (Version 3.3.3). The smartsva package (133) was used to create surrogate variables and the ewaff R package (`https://github.com/perishky/ewaff`) was used to conduct the EWAS, all p-values are two-sided.

3.4 Results

3.4.1 Database interface and use

There are two ways to access this large, curated database: through the main website www.ewascatalog.org or by using the R package “ewascatalog”. The website provides a simple user interface, which resembles that of the GWAS catalog (94), whereby there is a single search bar to explore the database and links to tabs that contain documentation on the contents and how to cite its use (Figure 1). Users may enter a CpG, gene, genome position or trait into the search bar and it will rapidly return detail for relevant EWAS associations, including CpG, trait, sample size, publication and association (effect and P value) (Figure 1). This information along with additional information such as ancestry, outcome, exposure units, and tissue analysed are available for download as a tab-separated value (tsv) text file. Unlike other EWAS databases, we provide the option of downloading summary results for both the user’s search and for the entire database.



Search

Enter CpG, gene, region or trait

Examples: cg00029284, chr12:111731203, FTO, 6:15000000-25000000, body mass index.

Search — *CpG sites*

cg00029284

Search — *Genes*

FTO

Search — *Traits*

Depression

Show 10 entries

Search:

Author	PMID	Outcome	Exposure	Analysis	N	CpG	Location	Gene	Beta	P
Bostrom AE	28618313	DNA methylation	Depression	-	93	cg13227621	chr1:45671361	ZSWIM5	-0.22	8.9E-09
Bostrom AE	28618313	DNA methylation	Depression	-	93	cg04102384	chr6:31668622	BAT5	-0.59	1.1E-08
Weder N	24655651	Depression	DNA methylation	-	190	cg03535461	chr1:23885568	ID3	NA	5.4E-08

Showing 1 to 3 of 3 entries

Figure 3.2: **Using the EWAS catalog**. At the top of the figures is the home page URL, ewascatalog.org. Below that are examples of three types of searches possible: 1. CpG sites, 2. genes and 3. traits. Finally, the results are displayed after searching the catalog for “Depression”. Circled in red is the download button, this button enables the user to download the results of their search as a tab-separated value file. This file will contain the information shown on the website as well as additional analysis information."

The R package, along with installation instructions and examples are available at `https://github.com/ewascatalog/ewascatalog-r/`. Once installed, the database can be queried directly in R using the "ewascatalog()" function similar to the website: simply supply the function with a CpG site, gene, genome position or trait and the function returns the same output as is downloadable from the website.

3.5 Discussion

In this chapter, a database of previously published EWAS and the full summary statistics of 427 newly performed EWAS within ALSPAC and GEO has been established. This is freely available for all researchers to use and provides a platform to explore what information has been gained from EWAS as well as a platform that can be used to pool all existing data to gain new insights into both the EWAS study itself and how DNA methylation associates with traits. Despite the fact The EWAS Atlas has similar aims to The EWAS Catalog, the latter provides full summary statistics, extra information and a user-friendly platform to enable more downstream analyses.

The EWAS catalog team will continue to collate and upload newly published EWAS and further increase the number of full summary statistics on the website by performing additional EWAS on available datasets and by inviting EWAS authors to provide full summary statistics. Currently work is ongoing to include additional functionality to allow users to easily and systematically compare their EWAS findings to EWAS in the database. With this full summary data, it is possible to make greater strides into discovering the epigenetic architecture of traits.

Therefore, despite the fact no extra information about EWAS was presented in this chapter, a platform has been made that enables us to explore 1) what information has been gained from EWAS and 2) what could explain EWAS associations. This will be explored in the next chapter.

Chapter 4

Properties of epigenome-wide association studies

4.1 Abstract

Understanding the nature of EWAS associations is imperative for biological inference from these studies. This understanding may also impact future study design. Of the data in the EWAS Catalog, 9.9% of reported associations are from CpGs measured by unreliable probes and 21% of studies did not account for both batch effects and cellular composition. Suggesting, some associations may be false positives. However, characteristics of DNA methylation also likely partly explain some EWAS associations - heritability and variability of DNA methylation explained 0.084% of the variance of effect EWAS effect sizes. Differentially methylated positions (DMPs) were found

to be present in actively transcribed promoter regions, enhancer regions and in over 100 transcription factor binding sites more than expected by chance, suggesting targeting these sites for measurement of DNA methylation may be more likely to yield results in future EWAS. This study also identifies associations at sites common to multiple traits. cg06500161 *ABCG1* associated with 71 traits, which were all traits relating to weight, metabolites or type-2-diabetes. This highlights the potential to use the data collected for the EWAS Catalog in **Chapter 3** to generate new hypotheses and connect DNA methylation changes to the broad range of potential phenotypic changes.

4.2 Introduction

Learning from successes and mistakes helps drive forward development. Hundreds of epigenome-wide association studies (EWAS) have been conducted in the last 10-15 years, yet no 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.

Since the inception of EWAS, it has become clear that batch effects and cellular heterogeneity can generate false positives and bias effect sizes (2,79,82). However, there are examples of replication amongst EWAS results, (103,134–139) and further, use of triangulation can be used to bolster evidence that changes in DNA methylation estimated are unlikely due to bias. By way of an example, changes in DNA methylation at *AHRR* have been replicated across multiple smoking EWAS (68,103,104,140) and as functional reaseach has implicated this gene in handling toxic substances found in tobacco smoke (141), it seems unlikely these findings are chance occurances.

The characteristics of the DNA methylome may also explain some EWAS findings. Heritability varies across DNA methylation sites (142,143), and so if genetic effects are driving associations, either through confounding or with DNA methylation as a mediator, one would expect heritable sites to be commonly identified in EWAS. Variance is also heterogenous across sites (144). Technical effects are more likely to influence DNA methylation variation at sites for which measured variation is low. Thus, some studies have advocated removing these sites to prevent reporting generating false positives and to reduce the multiple testing burden (145,146). However, it is unclear how variance in DNA methylation relates to the magnitude of effect estimates. Experimental studies have shown DNA methylation changes at different locations of the genome correlate with different regulatory functions. For example, an increase in DNA methylation at transcriptional start sites is correlated with a decrease in gene expression (18,33,34), but an increase in DNA methylation within a gene body shows the opposite association (35,36). Thus, genomic location of DNA methylation sites is likely to influence their likelihood of association with a trait.

Understanding underlying factors that drive EWAS results is essential for future study design. This may come in the form of proper consideration of potential biasing factors, or by selecting certain DNA methylation sites based on their specific characteristics. Further, understanding the characteristics of DNA methylation-trait associations can inform the design of future technologies aimed at measuring DNA methylation for EWAS.

Also, by examining the commonalities of EWAS results, one has the potential to uncover links between traits that have not previously been made or to identify new potential mediating factors between traits.

In this study we first describe the data present in the EWAS Catalog before exploring various explanations for the findings.

4.3 Methods

4.3.1 Epigenome-wide association studies data

All the data for the analysis were extracted from The EWAS Catalog (**Chapter 3**). This includes 178 published studies, 387 EWAS from the ARIES subsection of ALSPAC (**Section 3.3.4**) (128–130) and 40 EWAS performed using data from the gene expression omnibus (GEO) resource. See **Chapter 3** for more details.

Description of catalog data

Associations between DNA methylation and traits, unless otherwise stated, were extracted at $P < 1x10^{-7}$. Each of the CpGs in the Catalog are annotated to genes, using data from the meffil R package .

T-statistics (t) were calculated using P-values, sample sizes (n) and the qt() function in R. r^2 values were calculated from t-statistics as follows

$$r^2 = \frac{t^2}{t^2 + n - 1} \tag{4.1}$$

We identified traits for which r^2 values might be inflated. For each EWAS the estimated r^2 values were summed and these values were transformed to approximate a normal distribution. Then a z-test was performed to assess which sum of r^2 values were greater than the mean sum of r^2 values. From the z-test, those with a FDR-corrected P-value of less than 0.05 were labelled as having inflated r^2 values.

4.3.2 Identifying faulty probes

By far the most common method to measure DNA methylation across the studies in The EWAS Catalog is using the Illumina Infinium HumanMethylation450 Beadchip. Since its development, the array has been extensively characterised (2,79,82,83) and it was found that not all probes map just to the CpG they were designed to bind to. Some probes map to SNPs, others are non-specific and some are prone to cross-hybridisation. We assigned probes to be 'potentially faulty' if they were characterised as such by Zhou et al. (83).

4.3.3 Replication

An association (at $P<1x10^{-7}$) was deemed to be replicated if it had been identified by another study at $P < 1x10^{-4}$. We assessed replicability of EWAS within the database in two separate ways. Firstly, replication within studies is recorded in the EWAS Catalog, thus we simply performed a lookup for any studies that performed a replication or meta-analysed discovery and replication datasets. Secondly, we performed a lookup of results for any traits for which multiple EWAS had been conducted.

The Catalog also contains results from studies that have uploaded their data to GEO as well as results from the re-analysis of that data performed by The EWAS Catalog team. These re-analyses adjusted for 20 surrogate variables only as many studies did not provide a complete set of covariates to GEO. We performed a lookup of results found in the original EWAS in the re-analysed data.

4.3.4 Selecting data to assess DNA methylation characteristics

Before further analyses, all potentially faulty probes and probes that mapped to sex chromomsomes were removed, studies with likely inflated r^2 values were excluded, and studies for which re-analysis of the data replicated less than 10% of the findings were removed.

4.3.5 DNA methylation characteristics

The relationship between heritability, variability and average level of DNA methylation at each CpG site and EWAS effect size was assessed. To allow this across traits, we standardised beta coefficients, $\beta_{standard}$, like so,

$$\beta_{standard} = \frac{\beta\sigma(x)}{\sigma(y)} \tag{4.2}$$

As individual participant data were not available to us, the variance in DNA methylation sites was approximated by the variance in DNA methylation at sites as supplied by the GoDMC (147) and the trait variance was estimated by rearranging equation (4.3) depending on whether DNA methylation was the independent (x) or dependent (y) variable in the model.

$$r^2 = \frac{\beta^2\sigma^2(x)}{\sigma^2(y)} \tag{4.3}$$

GoDMC (147) also provided the mean levels of DNA methylation at each site. Heritability of DNA methylation at each site has been previously estimated by McRae et al. 2014 (142) and Van Dongen et al. 2016 (148), these values were kindly made publically available by the authors of those studies and were used in this study.

Relationships between each characteristic and effect size were assessed using linear regression, fitting the standardised effect size as the dependent variable and the characteristic as the independent variable. The standardised effect sizes were rank normalised to ensure normality and remove the impact of outliers.

4.3.6 Enrichment tests

Locus Overlap Analysis (LOLA) (149) was used to assess whether DMPs identified in the EWAS Catalog were enriched for 25 chromatin states and 167 transcription factor binding sites in 127 different cell types comprising 30 distinct tissues. These data were generated by the Roadmap Epigenomics Project (150) and ENCODE (151).

Five different groups of DMPs were defined for the enrichment analyses. Group A comprised all sites associated with any complex trait at the conventional P-value threshold used in EWAS, $P < 1x10^{-7}$. As multiple EWAS were conducted, DMPs in group B were defined as being associated with any complex trait at a stricter P-value threshold, defined as the conventional threshold divided by the number of EWAS included in the analyses, $P < 1.3e\text{-}10$. DMPs replicated at $P < 1x10^{-4}$ in any other EWAS of the same trait comprised group C. Group D and E were equivalent to groups A and B, except were restricted to DMPs identified in whole blood.

To assess enrichment, LOLA performs Fisher's exact test and generates an odds ratio that can be interpreted as the odds of the DMPs being within an annotation divided by the odds of the DMPs not being within an annotation. Genomic annotations may differ by CG content and thus a differential CG content of regions containing the DMPs of interest and the background group of CpG sites might bias enrichment estimates. Thus, background sites were matched on CG content before the analysis.

All analyses were completed using R (version 3.6.2).

4.4 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 4.1**).

Table 4.1: Description of data present in the EWAS Catalog

study-trait	value
Number of EWAS	614
Number of traits	556
Number of samples	389527
Median sample size (range)	536 (93 - 13474)
Number of associations	155976
Number of CpGs identified	129670
Number of genes identified	19305
Sex (%)	Both (38.6), Females (52.0), Males (2.1)
Ethnicities	EUR (75.3), Unclear (12.5), AFR (4.6), Other (3.6), ADM (1.6), EAS (1.4), SAS (1.0)
Age (%)	Adults (72.5), Geriatrics (11.2), Children (4.9), Infants (4.4)
Number of tissue types	42
Most common tissues (%)	whole blood (84.14), cord blood (4.34), cd4+ t-cells (2.60), placenta (1.24), saliva (0.99)

Identified associations were defined as those P < 1x10^{-7}

It may be that certain regulatory mechanisms are more important to phenotypic differences between individuals. By analysing datasets such as the EWAS Catalog, it might be possible to identify which regions may be more important and further, it could be used to identify novel mediating factors between traits.

The number of traits each CpG associated with was fairly even across chromosomes (**Figure 4.1**). There were five CpGs that associated with more than ten traits, cg01940273 -, cg05575921 *AHRR*, cg00574958 *CPT1A*, cg17901584 *DHCR24*, cg06500161 *ABCG1*. cg06500161 *ABCG1* was associated with more traits than any other site - 71 traits. These correspond mostly to metabolites, weight-related traits, and type two diabetes.

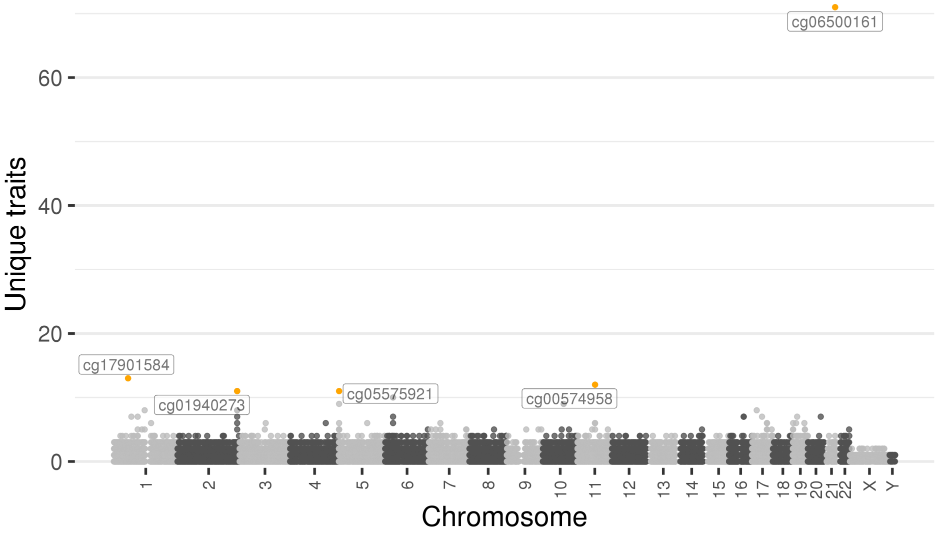


Figure 4.1: **Number of unique traits associated with DNA methylation at each CpG**. Sites associated with more than 10 unique traits are highlighted in orange and labelled.

Next we estimated the variance (see equation (4.1)) captured by each association to gauge the level of covariation between complex traits and DNA methylation.

The total trait variance correlated with DNA methylation (r^2) at each site varied from 0.0011 to 0.97 with a median of 0.093 (**Figure 4.2**). The sum of r^2 values ranged greatly from 0.0055 to 23,879 (**Figure 4.3**), with a median of 1.2. There was evidence that eight studies had a total sum of r^2 values greater than the mean (FDR < 0.05) and r^2 values from individual associations from these studies made up the majority of r^2 values greater than 0.1 (**Figure 4.2**). When excluding those studies from the results, the median r^2 value at individual sites was 0.025.

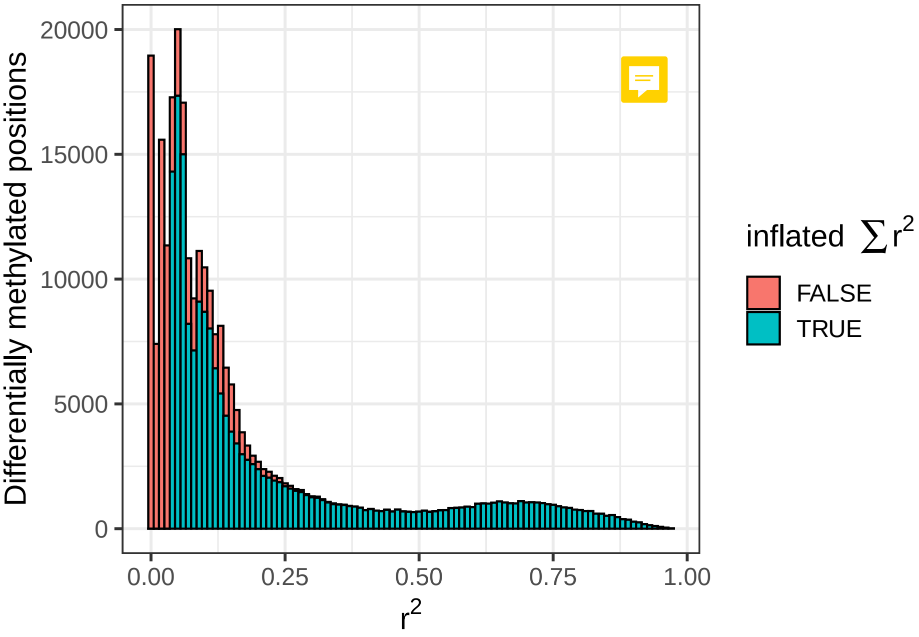


Figure 4.2: **Distribution of $\mathbf{r^2}$ values across all CpG sites in The EWAS Catalog**. Each EWAS can identify multiple differentially methylated positions, each of which will capture some variance of the trait of interest for that EWAS (r^2). $\sum r^2$ is the sum of r^2 values, the distribution of which is shown in **Figure 4.3**. Eight studies were identified for which there was strong evidence that the sum of r^2 values were greater than the mean across all studies. All of the differentially methylated positions identified by those studies are highlighted in blue on the plot.

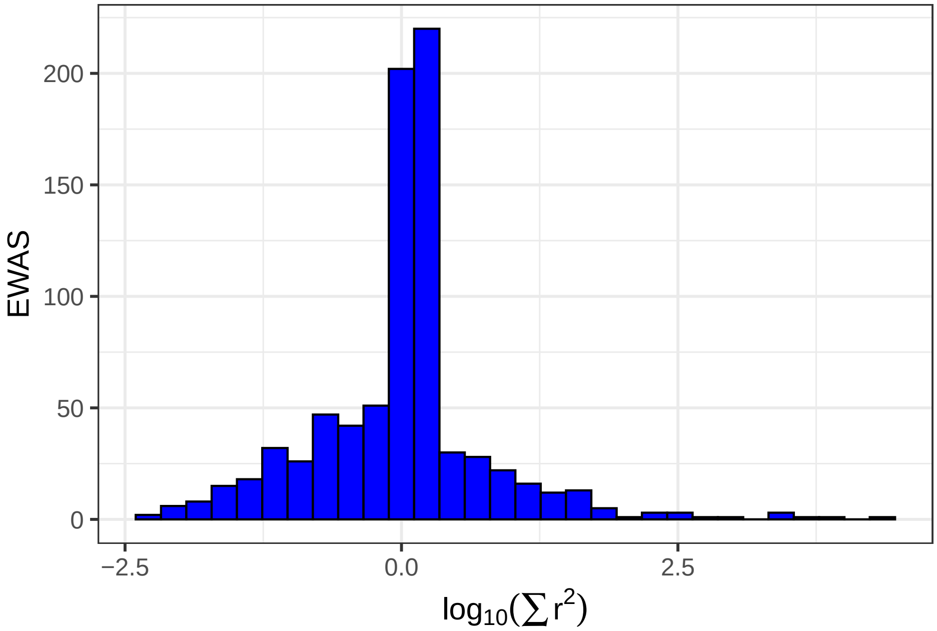


Figure 4.3: **Distribution of the sum of r^2 values across each study in The EWAS Catalog**.

These results suggest that some associations within the database are likely to be inflated, yet for most traits, variation at individual DNA methylation sites captures little trait variance. Summing the r^2 values indicates a substantial proportion of trait variance can be captured by multiple DNA methylation sites for some traits, but this can only be estimated by jointly modelling the contribution of all sites to trait variance. This is explored in **Chapter 5**. Here, the sum of r^2 values is used to indicate whether the results of a study are likely inflated and thus unlikely to be robust.

4.4.1 Robustness of results

As discussed, cellular heterogeneity, batch effects and inclusion of faulty probes can lead to false positives in EWAS. The extent to which this might be the case within EWAS included within The EWAS Catalog was explored.

Each study may have reported results across multiple EWAS models, adjusting for different covariates. In at least one model, 579 studies adjusted for batch effects, 518 studies adjusted for cell composition, and 489 adjusted for both. Of all DMPs identifed, 9.3% were measured by potentially faulty probes and an extra 0.64% were present on sex chromosomes (**Figure 4.4**).

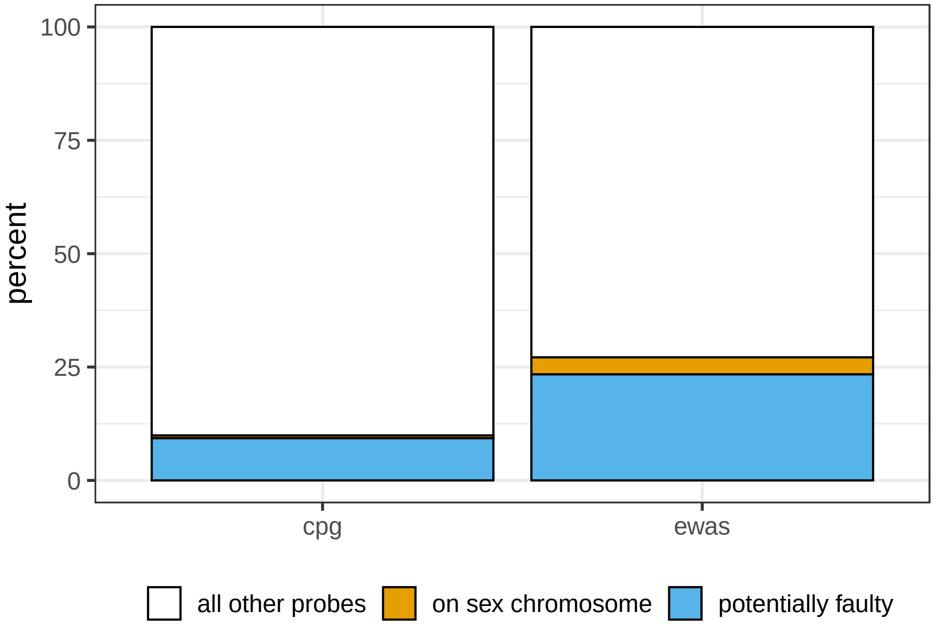


Figure 4.4: **The percentage of differentially methylated positions that may have been identified by faulty probes and the percentage of EWAS that reported identifying at least one of these probes**. Some CpGs are both on a sex chromosome and were identified as faulty by Zhou et al. They were labelled as 'potentially faulty'.

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 $P < 1x10^{-4}$.

From the studies that put their data on GEO, we re-analysed the association between DNA methylation and the phenotype of interest from the original study, including 20 surrogate variables as covariates. Both the original study results and the results from the re-analysis of the phenotype of interest are in The EWAS Catalog database for 10 studies. Across the studies, between 0% and 96.875% of DMPs were replicated at P < 1x10^{-4} (**Table 4.2**).

Table 4.2: GEO re-analysis replication

Trait	N-DMPs	N-replicated	Percent-replicated
Age at menarche	1	0	0.00
Arsenic exposure	11	0	0.00
Arsenic exposure	1	0	0.00
Fetal alcohol spectrum disorder	19	1	5.26
Inflammatory bowel disease	14	13	92.86
Nevus count	1	0	0.00
Psoriasis	16	0	0.00
Rheumatoid arthritis	47,875	116	0.24
Smoking	30	12	40.00
Smoking	32	31	96.88

N-DMPs = number of differentially methylated positions identified at P < 1x10^{-7}

N-replicated = number of DMPs replicated in the GEO re-analysis at P < P < 1x10^{-4}

Using the Catalog data I looked up whether CpG sites identified in relation to a trait in one study at P < 1x10^{-7} were also associated with that same trait in another study at P < 1x10^{-4}. There were 72 studies that shared a common phenotype of interest. Replication rate, judged as the percentage of CpGs also present in any other study of the same trait with P value < 1x10^{-4}, varied from 0 to 100 between studies (___Table 4.3, **Table 4.4**, **Table 4.5**).

Table 4.3: Replication rate

Trait	N-DMPs	N-replicated	N-replication-studies	Prop-replicated
glucose	4	1	1	0.25000
insulin	3	1	2	0.33333
insulin	1	0	2	0.00000
alzheimers	21	5	1	0.23810
alzheimers	25	7	1	0.28000
Birth weight	27	0	4	0.00000
Birth weight	1	0	4	0.00000
Birth weight	2	0	4	0.00000
Triglycerides	4	2	3	0.50000
Triglycerides	11	6	3	0.54545
Triglycerides	33	26	3	0.78788
Triglycerides	1	1	3	1.00000
Waist circumference	172	6	2	0.03488
Waist circumference	11	3	2	0.27273
Waist circumference	2	1	2	0.50000
Type II diabetes	11	2	2	0.18182
Type II diabetes	6	0	2	0.00000
Type II diabetes	1	1	2	1.00000
HOMA-IR	1	1	1	1.00000
HOMA-IR	5	1	1	0.20000
Schizophrenia	3	0	2	0.00000
Schizophrenia	163	0	2	0.00000
C-reactive protein	3	3	1	1.00000
C-reactive protein	226	17	1	0.07522
High-density lipoprotein cholesterol	2	2	1	1.00000
High-density lipoprotein cholesterol	63	5	1	0.07937
Serum high-density lipoprotein cholesterol	22	17	1	0.77273
Serum high-density lipoprotein cholesterol	213	11	1	0.05164
Serum low-density lipoprotein cholesterol	61	0	1	0.00000
Serum total cholesterol	1	0	2	0.00000
Serum total cholesterol	111	0	2	0.00000
Serum total cholesterol	1	0	2	0.00000
Serum triglycerides	46	38	1	0.82609
Serum triglycerides	99	33	1	0.33333
Rheumatoid arthritis	47,875	8	1	0.00017
Rheumatoid arthritis	6	0	1	0.00000
Depression	1	0	1	0.00000
Depression	2	0	1	0.00000

N-DMPs = number of differentially methylated positions identified at $P<1x10^{-7}$

N-replicated = number of DMPs replicated in the GEO re-analysis at $P<1x10^{-4}$

N-replication-studies = number of studies for which replication was examined

Prop-replicated = proportion of DMPs replicated.

Table 4.4: Replication rate in EWAS of smoking

Trait	N-DMPs	N-replicated	N-replication-studies	Prop-replicated
smoking	1	1	21	1.00000
smoking	1	1	21	1.00000
smoking	10	10	21	1.00000
smoking	1,065	862	21	0.80939
smoking	1	1	21	1.00000
smoking	22	20	21	0.90909
smoking	30	9	21	0.30000
smoking	44	42	21	0.95455
smoking	32	31	21	0.96875
smoking	450	417	21	0.92667
smoking	37	28	21	0.75676
smoking	3	3	21	1.00000
smoking	60	57	21	0.95000
smoking	171	171	21	1.00000
smoking	258	257	21	0.99612
smoking	20	1	21	0.05000
smoking	2,780	1,117	21	0.40180
smoking	524	424	21	0.80916
smoking	192	0	21	0.00000
smoking	177	172	21	0.97175
Maternal smoking in pregnancy	19	19	2	1.00000
Maternal smoking in pregnancy	24	24	2	1.00000
Maternal smoking in pregnancy	1,591	93	2	0.05845
Maternal smoking during pregnancy	121	18	1	0.14876
Maternal smoking during pregnancy	4	4	1	1.00000

N-DMPs = number of differentially methylated positions identified at $P<1x10^{-7}$

N-replicated = number of DMPs replicated in the GEO re-analysis at $P<1x10^{-4}$

N-replication-studies = number of studies for which replication was examined

Prop-replicated = proportion of DMPs replicated.

Table 4.5: Replication rate in EWAS of body mass index

Trait	N-DMPs	N-replicated	N-replication-studies	Prop-replicated
Body mass index	2	2	8	1.00000
Body mass index	133	83	8	0.62406
Body mass index	13	8	8	0.61538
Body mass index	14	12	8	0.85714
Body mass index	3	1	8	0.33333
Body mass index	5	3	8	0.60000
Body mass index	821	306	8	0.37272
Body mass index	182	113	8	0.62088
Body mass index	1	1	8	1.00000

N-DMPs = number of differentially methylated positions identified at $P<1x10^{-7}$

N-replicated = number of DMPs replicated in the GEO re-analysis at $P<1x10^{-4}$

N-replication-studies = number of studies for which replication was examined

Prop-replicated = proportion of DMPs replicated.

Before continuing to assess what CpG characteristics might, in part, explain some associations found in EWAS, we removed sites that were identified by potentially faulty probes and were on either of the sex chromosomes. Further, we removed the eight studies that had an inflated sum of r^2 values and studies for which fewer than 10% of sites identified in the original analyses were identified in a re-analysis using the data provided via GEO. Overall, this left 789 EWAS and 77127 associations (at $P < 1x10^{-4}$).

4.4.2 CpG characteristics

Using the selected EWAS results, we investigated whether the characteristics of DNA methylation at CpG sites explained associations found in EWAS.

It has previously been suggested that sites at which DNA methylation variability is low should be removed (145,146). The rationale being if total variation is low then the ratio of variation due to technical effects to variation due to biological effects will be greater and thus any association with a complex trait is more likely to be due to technical artefacts. However, selection would dictate that phenotypes (including DNA methylation) that have a large effect would be selected for (if they had a positive impact on fitness) or against (if they had a negative impact on fitness) (152,153). Therefore, it would be expected that DNA methylation at sites that have large impacts on cellular function would end up stabilising over time, and so the largest effects may be missed by removing CpG sites with low variances.

There was strong evidence of an inverse association between variance at a CpG site and effect size (P = 1e-99, **Table 4.6**), suggesting that removal of sites with little variation may reduce the chances of discovering changes in DNA methylation that have larger effects.

DNA methylation is a binary measurement, a CpG site can either be methylated or not. However, when measuring methylation across multiple DNA molecules, the proportion of those molecules methylated at a given site will be between 0 and 1. If DNA methylation at a given site is important for specific regulatory functions within a group of cells, one might expect that site to be methylated (or unmethylated) in the majority of the cells. Thus, changes in methylation away from an extreme, might have more of impact on cellular function.

There was strong evidence of an association between mean DNA methylation levels and negative effect sizes (P = 5.1e-86, **Table 4.6**) and an inverse association between mean methylation levels and positive effect sizes (P = 1e-99, **Table 4.6**).

DNA methylation changes are heritable (142,148), and DNA methylation could mediate the effects of genotype on complex traits or genotype might confound the association between DNA methylation and complex traits.

There was evidence that effect sizes tended to be greater in more heritable sites (P = 5.9e-05, **Table 4.6**).

The combined variance in effect size estimates explained by DNA methylation variability and heritability was 0.084 (**Table 4.6**).

Table 4.6: Association between CpG chars and associations in EWAS

characteristic	beta	r^2	p	auc
avg-meth (beta>0)	-6.7e-01	0.08377	1.0e-99	NA
avg-meth (beta<0)	4.3e-01	0.09138	5.1e-86	NA
variance	-1.2e+03	0.01806	1.0e-99	0.62
h^2	8.1e-02	0.00037	5.9e-05	0.78
variance + h^2	NA	0.08395	NA	0.78

avg-meth = average methylation level
beta > 0 = DNA methylation hypermethylated with respect to the trait
beta < 0 = DNA methylation hypomethylated with respect to the trait
auc = area under the curve

As the position of DNA methylation relative to genes is pertinent to its association with gene expression (**Section 1.2.3**) (18,33–36), the enrichment of DMPs identified in The EWAS Catalog across genomic regions and chromatin states were assessed (**Figure 4.5**). Across all tissues, there was a trend for sites to be enriched for promoter regions (OR > 1). Evidence of enrichment across different enhancer types was mixed and there was a trend towards depletion of sites within heterochromatic regions, poised and bivalent promoters, regions repressed by polycomb proteins and quiesscant regions (**Figure 4.5**, OR < 1).

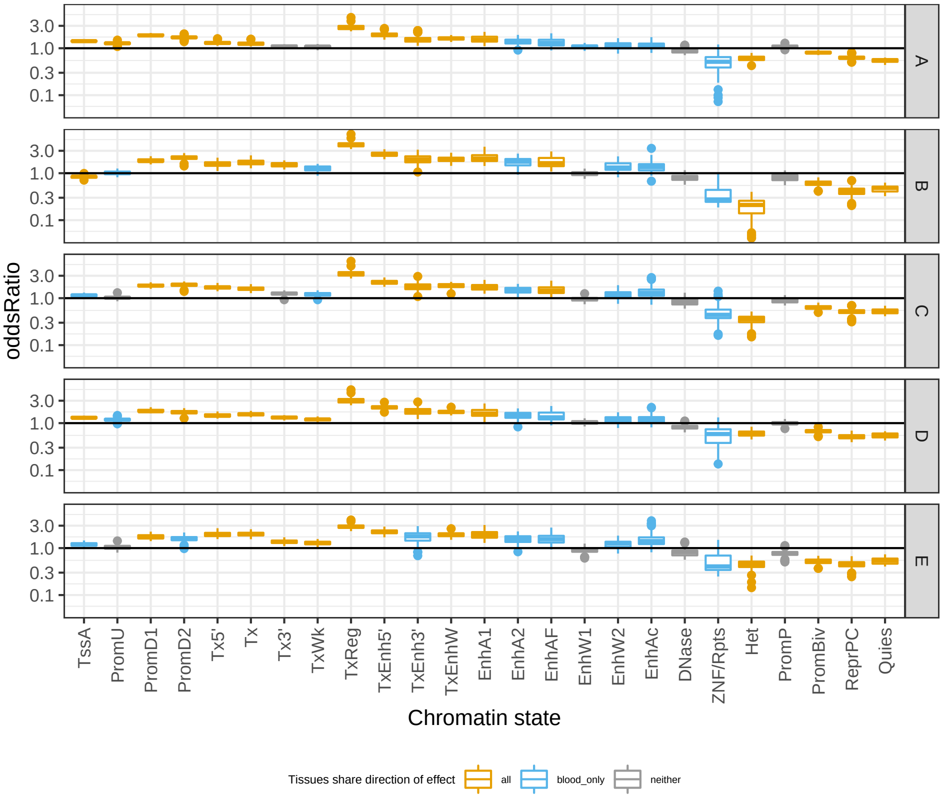


Figure 4.5: **Enrichment of DMPs for 25 chromatin states**. Chromatin states across the genome of 127 cell types comprising 30 distinct tissues were available from the Roadmap Epigenomics Project. Using LOLA, the enrichment of DMPs from across all data in The EWAS Catalog for chromatin states were assessed. DMPs were divided into five categories **A**: CpG sites associated with any complex trait at P<1e-7, **B**: sites from A that replicated in another study at P<1e-4, **C**: CpG sites associated with any complex trait at P<1.3e-10 (P < conventional threshold divided by total number of EWAS), **D**: sites from A that were measured in blood, **E**: sites from C that were measured in blood. The x-axis show the 25 chromatin states: TssA, Active TSS; PromU, Promoter Upstream TSS; PromD1, Promoter Downstream TSS with DNase; PromD2, Promoter Downstream TSS; Tx5', Transcription 5'; Tx, Transcription; Tx3', Transcription 3'; TxWk, Weak transcription; TxReg, Transcription Regulatory; TxEnh5', Transcription 5' Enhancer; TxEnh3', Transcription 3' Enhancer; TxEnhW, Transcription Weak Enhancer; EnhA1, Active Enhancer 1; EnhA2, Active Enhancer 2; EnhAF, Active Enhancer Flank; EnhW1, Weak Enhancer 1; EnhW2, Weak Enhancer 2; EnhAc, Enhancer Acetylation Only; DNase, DNase only; ZNF/Rpts, ZNF genes & repeats; Het, Heterochromatin; PromP, Poised Promoter; PromBiv, Bivalent Promoter; ReprPC, Repressed PolyComb, Quies, Quiescent/Low.

The sites identified by EWAS were also enriched for transcription factor binding sites (**Figure 4.6**). Of the 167 transcription factor binding sites tested, there was evidence that identified DMPs (P<1e-7) were enriched in XXX of them across all tissue types (P<XXX).

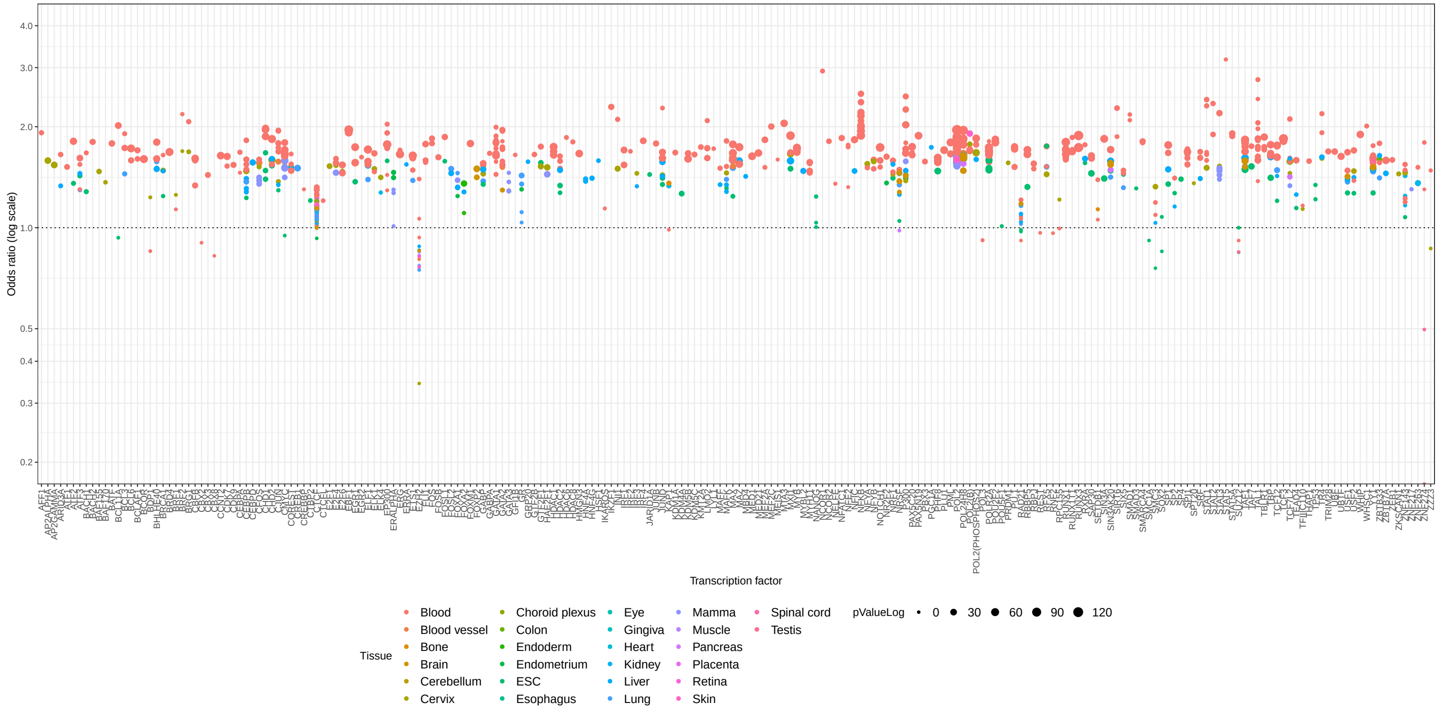


Figure 4.6: **Enrichment of DMPs for 167 transcription factor binding sites.**

4.5 Discussion

Understanding the nature of EWAS associations is imperative for biological inference. Using data from the EWAS Catalog we show that many CpGs associate with multiple different unique traits and the magnitude of these associations are partly explained by the characteristics of DNA methylation levels. False positives may also explain a proportion of EWAS associations. Roughly 10% of the differentially methylated positions identified were measured by potentially faulty probes and the median percentage of CpGs that could be replicated across studies was 50%.

4.5.1 Identifying mediators

Identifying modifiable molecular traits that mediate the effect of complex traits on disease is something that motivates a substantial portion of molecular epidemiology research (49,115). Having a database of associations between DNA methylation and various traits and diseases may enable easy identification of potential mediators that warrant follow-up. Overall, DNA methylation at 126,673 CpGs are associated with multiple traits. The CpG that was identified in the most EWAS, cg06500161 *ABCG1*, had evidence from multiple studies that methylation at that site associated with weight-related traits such as body mass index (60,70,154,155) and waist circumference (155), roughly 60 metabolites (137,138,156–158) and with type 2 diabetes (159). Some studies have explored these associations further, for example, two studies used Mendelian randomization (MR) to provide evidence that body mass index caused changes in methylation at this site (70,154). However, full characterisation and assessment of whether methylation at that site mediates the effect of adverse adiposity on any diseases has not been undertaken and could be followed-up.

4.5.2 Biased results

The potential biases in EWAS have been well documented (1) and were discussed at length in **Section 1.3.4**. It is encouraging that the majority of studies include batch effects and cell composition in at least one of their models (79.1262136%). However, there are still some studies including probes that have been characterised as faulty.

Differences in cell composition, sample ethnicity, covariates used and other differential biases between studies might explain the low replication rate. However, studies only tend to report associations below the conventional EWAS P-value threshold, $P < 1e\text{-}7$, so differences in study power could also be a major factor.

4.5.3 Understanding CpG charactersitcs

Characteristics of DNA methylation discovered in experimental studies, such as its association with gene expression, were used to select sites to measure DNA methylation (65). Further, studies have suggested selecting from those sites commonly measured, CpGs that have certain characteristics such as high variance (145,146).

Our results suggest removing CpG sites with low variances may make it more likely to remove sites with greater effects. Variance had a modest ability to predict whether or not a CpG site was likely to be identified in an EWAS, and it did not add to the predictive ability of heritability, despite explaining a higher proportion of variance in effect estimates. This may be explained by two things. Firstly, having a lower variance in the independent or dependent variable increases the standard error of the beta coefficient in a linear regression. Secondly, heritability will in part determine variance of DNA methylation.

4.5.4 Choosing sites to measure

As discussed in **Section 1.3.3**, the HM450 array was designed to capture DNA methylation in various regions of the genome. The probes of the array target over 99% of protein coding genes and predominantly target the promoter regions of these genes (REF). The newer HMEPIC array captures much of what the HM450 does, and further covers 58% of FANTOM5 enhancers (160).

The trend for DMPs to be enriched for promoter regions suggests there may have been some justification for the chosen sites. However, not all promoter regions were enriched with DMPs and bivalent promoters were depleted for DMPs. Enrichment of enhancers was also seen, but the magnitude of enrichment was smaller. When designing future arrays, these results suggest that continuing to target promoters and enhancers, whilst avoiding gene regions that are less likely to be actively transcribed may yield more associations in EWAS.

Despite the tissue specific nature of DNA methylation, the regions for which DMPs identified in The EWAS Catalog were found to be enriched were fairly consistent across tissues. However, enrichment of DMPs tended to be greater for blood-based genomic annotations, perhaps reflecting the fact the majority of EWAS in the EWAS Catalog were conducted using DNA methylation measured in whole blood.

4.5.5 Limitations

Individual participant data were not available and thus to calculate standardised betas, the variance of the trait had to be estimated from external measures of DNA methylation. If the GoDMC sample is not representative of the sample used for the study EWAS then these estimates may be substantially biased. Further, many studies do not report the effect estimates from their statistical analyses. If there is a marked difference in the studies that do not report effect sizes and those that do, then any associations between standardised effect estimates and DNA methylation site characteristics are likely to be biased.

Like other observable phenotypes, DNA methylation varies under many contexts. Time, sex, tissue type, population, socioeconmic position and many other factors may influence the results of EWAS. The majority of EWAS conducted have used DNA methylation measured in whole blood from European adults making the results not necessarily apply broadly outside those bounds. The need for tissue-specific data has been discussed previously in **Section 1.3.4**. Differences in DNA methylation between ethnic groups has been shown previously (161) and the predictive value of a smoking-related methylation score was shown to differ between Europeans and South Asians (140). This suggests any biological insight and population health benefits that may be the result of EWAS is likely to to be maximised by diversifying populations. It is unclear from this study whether the CpG characteristics and genomic annotations that show evidence that they influence EWAS results, will also influence EWAS results in the same way within a more ethnically diverse selection of samples.

4.6 Conclusion

This chapter demonstrates the potential for using large-scale EWAS databases to understand DNA methylation-trait associations. It was found that study design flaws can help explain some associations. However, it is noteworthy that the vast majority of studies have accounted for some potential biasing factors, for example 79.1262136% of studies adjusted for batch effects and cell composition. Further, there was an invese association between DNA methylation variability and effect size, suggesting that studies that remove variable sites prior to analysis could be excluding important regions from the analysis. Finally, cg06500161 *ABCG1* was identified as being associated with 71 traits that share known biological relationships. This highlights the potential to use The EWAS Catalog to identify molecular markers that might underlie the relationship between traits.