Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits

Elis Hannon,1 Tyler J. Gorrie-Stone,2 Melissa C. Smart,3 Joe Burrage,1 Amanda Hughes,9 Yanchun Bao,3 Meena Kumari,3 Leonard C. Schalkwyk,2 and Jonathan Mill1,*

Characterizing the complex relationship between genetic, epigenetic, and transcriptomic variation has the potential to increase understanding about the mechanisms underpinning health and disease phenotypes. We undertook a comprehensive analysis of common genetic variation on DNA methylation (DNAm) by using the Illumina EPIC array to profile samples from the UK Household Longitudinal study. We identified 12,689,548 significant DNA methylation quantitative trait loci (mQTL) associations (p < 6.52 × 10^{-14}) occurring between 2,907,234 genetic variants and 93,268 DNAm sites, including a large number not identified by previous DNAm-profiling methods. We demonstrate the utility of these data for interpreting the functional consequences of common genetic variation associated with > 60 human traits by using summary-data-based Mendelian randomization (SMR) to identify 1,662 pleiotropic associations between 36 complex traits and 1,246 DNAm sites. We also use SMR to characterize the relationship between DNAm and gene expression and thereby identify 6,798 pleiotropic associations between 5,420 DNAm sites and the transcription of 1,702 genes. Our mQTL database and SMR results are available via a searchable online database as a resource to the research community.

Introduction

DNA methylation (DNAm), an epigenetic modification to cytosine, is involved in mediating the developmental regulation of gene expression and function, as well as transcriptional processes such as genomic imprinting and X chromosome inactivation.1,2 Although often regarded as a mechanism of transcriptional repression, the relationship between DNAm and gene expression is highly complex and not fully understood.3 Gene-body DNAm, for example, is often associated with active expression4 and also influences other transcriptional processes, including alternative splicing and promoter usage.5 This dynamic property of DNAm means it can vary across samples and might underlie phenotypic differences. There is growing interest in characterizing the variation of DNAm across populations6,7 and in the role of DNAm in disease, and recent epigenome-wide association studies (EWASs) have identified robust associations between variable DNAm and cancer,8 as well as a diverse range of other complexphenotypes, including rheumatoid arthritis [MIM: 180300],9 body-mass index,10 schizophrenia [MIM: 181500],11 and Alzheimer disease [MIM: 104300].12 Characterizing the complex relationship between genetic, epigenetic, and transcriptomic variation will increase understanding about the mechanisms underpinning health and disease phenotypes. Twin and family studies have demonstrated that population-level variation in DNAm is under considerable genetic control, although these effects vary across genomic loci, developmental stages, and different cell and tissue types.11–17 Studies in a variety of tissues, including brain, whole blood, pancreatic islet cells, and adipose tissue, have identified widespread associations between common DNA sequence variants and DNAm.17–22 These DNAm quantitative trait loci (mQTLs) are primarily cis-acting, are enriched in regulatory chromatin domains and transcription-factor binding sites, and have been shown to colocalize with gene expression quantitative trait loci (eQTLs).3,17,23

There is considerable interest in using mQTLs, along with other types of molecular QTLs, to interpret the functional consequences of common genetic variation associated with human traits, especially because the actual gene(s) involved in mediating phenotypic variation are not necessarily the most proximal to the lead SNPs identified in genome-wide association studies (GWASs). Of note, GWAS variants are enriched in enhancers and regions of open chromatin,24,25 reinforcing the hypothesis that most common genetic risk factors influence gene regulation rather than directly affecting the coding sequences of transcribed proteins.26 Importantly, evidence for the co-localization of genetic variants associated with both phenotypic and regulatory variation is not sufficient to show that the overlapping association signals are causally related; additional analytical steps are needed to distinguish pleiotropic effects—i.e., where the same variant is influencing both outcomes, although not necessarily dependently—from those that are an artifact of linkage disequilibrium (LD). We recently extended the use of one approach—summary-data-based Mendelian randomization (SMR), which was initially used in conjunction with expression quantitative trait loci (eQTL) data27—to prioritize genes for GWAS-nominated loci using mQTL data.28

1University of Exeter Medical School, University of Exeter, Exeter EX2 5DW, United Kingdom; 2School of Biological Sciences, University of Essex, Colchester, CO4 3SQ, United Kingdom; 3Institute for Social and Economic Research, University of Essex, Colchester CO3 3LG, United Kingdom
*Correspondence: j.mill@exeter.ac.uk
https://doi.org/10.1016/j.ajhg.2018.09.007.
© 2018 The Authors. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Building on our previous work, we used the Illumina EPIC array and imputed SNP data to identify mQTLs associated with variable DNAm at ~850,000 sites across the genome in samples from the Understanding Society UK Household Longitudinal Study (UKHLS) (n = 1,111). We then used these mQTLs within the SMR framework to refine genetic association data from publicly available GWAS datasets in order to prioritize genes involved in 63 complex traits and diseases. We subsequently used the SMR approach to identify pleiotropic relationships between DNAm and variable gene expression by using publicly available whole-blood gene eQTL data. Our mQTL database and SMR results are available via a searchable online database as a resource to the research community (see Web Resources).

Subjects and Methods

Sample Description

The British Household Panel Survey (BHPS) began in 1991, and in 2010 it was incorporated into the larger UKHLS32 (also known as Understanding Society), which is a longitudinal panel survey of 40,000 UK households from England, Scotland, Wales, and Northern Ireland. Since 1991, annual interviews have collected sociodemographic information, and in 2011–2012, biomedical measures and blood samples for BHPS participants were collected during a nurse visit in the participant’s home. Respondents were eligible to give a blood sample if they had taken part in the previous main interview in English; were 16 or older; lived in England, Wales, or Scotland; were not pregnant; and met other conditions detailed in the user guide.33 For each participant, non-fasting blood samples were collected through venipuncture; these were subsequently centrifuged so that plasma and serum were separated, and samples were aliquoted and frozen at ~80°C. DNA has been extracted and stored for genetic and epigenetic analyses.

Genome-wide Quantification of DNAm

DNAm was profiled in DNA extracted from whole blood for 1,193 individuals who were aged from 28 to 98; who were eligible for and consented to both blood sampling and genetic analysis; who had been present at all annual interviews between 1999 and 2011; and whose time between blood sample collection and processing did not exceed 3 days. Eligibility requirements for genetic analyses meant that the epigenetic sample was restricted to participants of white ethnicity. The EZ-96 DNA Methylation-Gold kit (Zymo Research) was used for treating 500 ng of DNA from each sample with sodium bisulfite. DNAm was quantified with the Illumina Infinitum HumanMethylationEPIC BeadChip run on an Illumina iScan System according to the manufacturer’s standard protocol. Samples were randomly assigned to chips and plates so that batch effects would be minimized. In addition, the inclusion of a fully methylated control (CpG Methylated HeLa Genomic DNA; New England BioLabs) in a random position on each plate facilitated sample tracking and helped to resolve experimental inconsistencies and confirm data quality.

DNAm Data Preprocessing

Raw signal intensities were imported from .idat files into the R statistical environment31 and converted into beta values (the proportion of DNA methylation at individual sites was measured) with the bigmelon package.32 These data were processed via a standard pipeline including the following steps: (1) detection of outlier samples via principal-component analysis and Mahalanobis distance equivalents, (2) confirmation of complete bisulphite conversion via control probes, (3) comparison of estimated age from the data via the Horvath Epigenetic Clock algorithm43 and reported age at sampling, and (4) visualization of principal components. Data were normalized with the desen function within the watemelon package,34 which performs background adjustment and between-sample quantile normalization of methylated (M) and unmethylated (U) intensities separately for type I and type II probes. Samples that were dramatically altered as a result of normalization were excluded on the basis of the difference between the normalized and raw data; those with a root mean square and standard deviation > 0.05 were removed. Samples were then filtered so that those with >1% of sites with a detection p value > 0.05 were excluded. Finally, DNA-methylation sites with a bead count <3 were excluded along with those in which >1% of the sample had a detection p value > 0.05. The raw DNA methylation data from the final sample set was then re-normalized with the desen function. The final dataset included 857,071 DNA-methylation sites and 1,175 individuals for subsequent analysis. These DNAm data are available upon request through the European Genome-Phenome Archive under accession code EGAS00001001232.

Annotation of DNAm Sites

The genomic location of DNAm sites along with genic, DNase hypersensitive sites and open chromatin annotation were taken from the manifest files provided by Illumina and downloaded from the product support pages (see Web Resources).

Genotyping and Imputation

UKHLS samples were genotyped with the Illumina Infinium HumanCoreExome BeadChip Kit as previously described (12v1-0).35 This array contains a set of >250,000 highly informative genome-wide tagging single-nucleotide polymorphisms (SNPs) as well as a panel of functional (protein-altering) exonic markers, including a large proportion of low-frequency (MAF 1%–5%) and rare (MAF < 1%) variants. Genotype calling was performed with the gencall algorithm within GenomeStudio (Illumina). After only the samples with matched DNAm data were selected, variants were refiltered prior to imputation. PLINK36 was used for removing samples with >5% missing data. We also excluded SNPs characterized by >5% missing values, a Hardy-Weinberg equilibrium p value < 0.001, and a minor-allele frequency of <5%. For identification of related samples, SNPs underwent LD pruning, and the --genome command in PLINK was used for calculating the proportion of identity-by-descent for all pairs of samples; 58 pairs of related samples (PI_HAT > 0.2) were identified, and randomly excluding one individual from each pair ensured that the samples were independent. These data were then imputed with the 1000 Genomes phase 3 version5 reference panels SHAPEIT and minimac3.37 Best-guess genotypes were called, and variants were filtered to those with a minor-allele frequency >0.01 and an INFO score >0.8. Because variants were named using their locations (“chr:pos”) and variant type (SNP/INDEL), duplicate variants were also excluded. Principal components were calculated from the imputed genotype data via GCTA (a tool for genome-wide complex-trait analysis).38 16 samples
were identified as being outliers (defined as more than 2 standard deviations from the mean) in a scatterplot of the first two principal components and were excluded from subsequent genetic analyses. Principal components were then recalculated for inclusion as covariates in QTL analyses. The imputed genetic variants were then filtered so that variants characterized by >5% missing values, a Hardy-Weinberg equilibrium p value < 0.001, a minor-allele frequency of <5%, and a minimum of five observations in each genotype group were excluded. These genotype data are available on application through the European Genome-phenome Archive under accession code EGAS00001001232.

**DNAm Quantitative-Trait Loci**

Cross-hybridizing probes, probes with a common SNP (-European population minor-allele frequency > 0.01) within 10 bp of the CpG site or a single base extension9,40 and probes on the sex chromosomes were excluded from the QTL analysis. In addition, 977 subprobes identified by Illumina were also excluded. We performed a genome-wide mQTL analysis; in total, we tested 766,714 DNAm sites against 5,210,475 genetic variants by using the R package MatrixEQTL.41 This package enables fast computation of QTLs by only saving those more significant than a pre-defined threshold (set to p = 1 × 10^{-8} for this analysis). We fitted an additive linear model to test whether the number of alleles (coded 0,1,2) predicted DNAm at each site; we included covariates for age, sex, six estimated cellular composition variables (B cells, CD8 T cells, CD4 T cells, monocytes, granulocytes, natural killer T cells),42,43 two binary batch variables, and the first ten principal components from the genotype data to control for ethnic differences. We used a Bonferroni-corrected multiple-testing threshold, set to genome-wide significance for GWAS and divided by the number of DNAm sites tested (i.e., 5 × 10^{-8}/766714 = 6.52 × 10^{-14}). We used the clump command in PLINK36 to identify the number of independent associations for each DNAm site with more than 1 significant mQTL by using the following parameters: --clump-p1 1e-8--clump-p2 1e-8--clump-r2 0.1--clump-kb 250.

**Bayesian Co-localization**

Out of all DNAm sites with at least 1 significant mQTL (p < 1 × 10^{-10}), all pairs of DNAm sites located on the same chromosome and within 250 kb of each other were tested for co-localization. Because data for all SNPs (regardless of significance) are required for this analysis, first, the mQTL analysis was rerun for these DNAm sites so that all association statistics (p value, regression coefficient, and r^2, so that the standard error could be inferred) could be recorded for all SNPs within 500 kb of the DNAm site. Co-localization analysis was performed as previously described44 with the R coloc package (see Web Resources). From our mQTL results we input the regression coefficients, their variances, and SNP minor-allele frequencies, and we left the prior probabilities as their default values. This methodology allowed us to quantify the support across the results of each GWAS for five hypotheses by calculating the posterior probabilities, denoted as Ppi for hypothesis Hi.

\[\begin{align*}
H_0 &: \text{there exist no causal variants for either CpG site;} \\
H_1 &: \text{there exists a causal variant for CpG}_1\text{ only;} \\
H_2 &: \text{there exists a causal variant for CpG}_2\text{ only;} \\
H_3 &: \text{there exist two distinct causal variants, one for each CpG;} \text{or} \\
H_4 &: \text{there exists a single causal variant common to both CpGs.}
\end{align*}\]

**Summarized Mendelian Randomization Analysis 1:** Identifying Putative Pleiotropic Relationships between DNAm and Complex Traits

SMR analysis between DNAm and complex traits was performed with publicly available software (see Web Resources) as previously described.27,50 Publicly available genome-wide association study (GWAS) results were downloaded from a range of sources and converted to the appropriate format for the SMR analysis. We renamed SNPs in the 1000 Genomes format (chr:bp) to align them with the mQTL output by using dbSNP version 141 (where SNP locations for hg19 were not provided in the results file). Where allele frequency was not provided, it was taken from the European subset of 1000 Genomes (phase 3, version 5). Details for how each set of results was processed can be found in Table S4. We used significant mQTLs (p < 1 × 10^{-10}) calculated in the UKHLS sample to identify genetic instruments for 126,457 DNAm sites that were included in the SMR analysis. The SMR test comprises of two steps. First, we performed a two-sample Mendelian randomization with the two-step least-squares (2LSL) approach by using the effect size of the top cis-QTL SNP and its corresponding effect in the GWAS. The significance threshold for this part of test was set at 3.95 × 10^{-7}, calculated by the Bonferroni correction method and adjusted for the number of DNAm sites tested (0.05/126,457). Second, we tested for heterogeneity of effects by using alternative SNPs as the instrumental variable, on the basis of the theory that if both DNAm and the GWAS trait were associated with the same causal variant, the choice of SNP would be irrelevant, whereas if they were associated with different causal variants, the differing linkage disequilibrium relationships between the instruments and each causal variant would lead to variation in the estimated effect between the trait and DNAm. Non-significant heterogeneity (heterogeneity in dependent instruments [HEIDI] p > 0.05) indicates that there is a pleiotropic effect on a GWAS trait and DNAm. This approach was repeated with publicly available eQTL data from Westra et al.;45 in this analysis, significant pleiotropic associations between gene expression and complex traits were selected as those with SMR p < 8.38 × 10^{-5} (corrected for 5,966 gene expression probes tested) and HEIDI p > 0.05.

**Summarized Mendelian Randomization Analysis 2:** Identifying Putative Pleiotropic Relationships between DNAm and Gene Expression

We used a second application of the SMR analysis to identify pleiotropic relationships between DNAm and gene expression. Gene eQTL results from the Westra eQTL study45 were downloaded along with the SMR software. SNP IDs were converted to the 1000 Genomes format (so they would match the mQTL output), and SNP frequencies were taken from the European subset of 1000 Genomes (phase 3, version 5). These data included eQTLs at 5,966 probes. All pairs of CpG and genes where tested as long as (1) the CpG had a significant mQTL (p < 1 × 10^{-10}), (2) the gene had a significant eQTL (p < 5 × 10^{-8}), and (3) there was a common genetic variant tested within 500 kilobases of the gene expression probe and DNAm site. In total, 488,342 pairs of DNAm sites and gene expression transcripts were tested; therefore, the significance threshold for the first stage of the SMR test was set to p < 1.02 × 10^{-7} after a Bonferroni correction for the number of tests was applied. Consistent with all other SMR analyses in this manuscript, a non-significant heterogeneity test (HEIDI p > 0.05) in step 2 of the SMR analysis was used for classifying pleiotropic relationships from artifacts of linkage disequilibrium.

Please cite this article in press as: Hannon et al., Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylytic Variation,..., The American Journal of Human Genetics (2018), https://doi.org/10.1016/j.ajhg.2018.09.007
Enrichment Analyses

DNAm sites were annotated to genes and CpG islands with the information provided in the Illumina manifest file, which is based on the UCSC RefGene and CpG island databases. Sites are annotated to genes if they are located within the gene body or up to 1,500 base pairs form the transcription start site. Sites are annotated to CpG islands if they are located within the boundaries of a CpG island, to a shore if they are located up to 2,000 base pairs from an island, or to a shelf if they are between 2,000 and 4,000 base pairs from an island. Frequency tables were used for recording the number of sites annotated to each feature category, and Chi-square tests were used for identifying different distributions across these annotation categories between all tested DNAm sites and the subset of sites considered for enrichment analysis (e.g., all DNAm sites with at least one significant mQTL).

Data Availability

Summary statistics for all Bonferroni-significant DNA-methylation quantitative-trait loci are available for download from the Complex Disease Epigenomics Group website, where readers can also explore many of the results included in this manuscript through our interactive web application. Analysis scripts used in this manuscript are available on GitHub, and data on phenotypes linked to DNA methylation are available on METADAC. See the Web Resources and the Accession Numbers sections.

Results

Additional mQTL Associations Identified with the Illumina EPIC Array

An overview of our study design is presented in Figure S1. We tested 5,210,475 imputed genetic variants against the 766,714 DNAm sites that were on the Illumina EPIC array and that passed our stringent QC criteria (see Subjects and Methods). We identified 12,689,548 significant mQTL associations (we used a conservative Bonferroni-corrected threshold of \( p < 6.52 \times 10^{-14} \)) between 2,907,234 genetic variants and 93,268 DNAm sites (Table S1; Figure 1A); there was a mean percentage point change in DNAm per additional reference allele of 3.46% (SD = 3.01%) across all mQTL-associated sites. Existing mQTL databases have been almost exclusively generated with the Illumina 450K array; more than half of the DNAm sites (\( n = 48,099, 51.6\% \); Table S2) that we identify as being associated with genetic variation with the Illumina EPIC array involve additional content not previously interrogated (Figure S2).

Importantly, these additional mQTL associations are annotated to 5,172 genes not included in mQTL databases generated with the Illumina 450K array (Figure S3). DNAm sites associated with genetic variation are associated with a median of 65 (interquartile range = 22–162) mQTLs, probably reflecting linkage disequilibrium (LD) relationships between proximal variants. In contrast, each mQTL variant is associated with a median of two (interquartile range = 1–5) DNAm sites, and the majority of mQTL SNPs (\( n = 1,003,238, 34.5\% \)) are associated with DNAm at only a single site (Figure S4). We performed LD clumping of the results for each DNAm site to identify the number of independent associations for each DNAm site (see Subjects and Methods); this process reduced the number of mQTL associations (\( p < 6.52 \times 10^{-14} \)) to 161,761 (1.27% of the total number of unclumped significant mQTL associations); a median of 1 (interquartile range = 1–2) mQTL variant associated with each DNAm site (Figure S5). At a more relaxed “discovery” threshold (\( p < 1 \times 10^{-10} \)), we identified a total of 17,051,673 mQTL associations between 3,281,391 genetic variants and 114,595 DNAm sites; these results are available in a searchable database (see Web Resources).

mQTL Associations Predominantly Occur in cis and Influence DNAm at Sites Known to Be Influenced by Heritable Factors

Consistent with the results of previous studies, we found that the majority of mQTL associations (\( n = 11,679,376; \)
(Figure 1A). Cis mQTL variants are typically associated with larger effects on DNAm than those acting in trans (average cis effect = 3.48% change in DNAm per allele, average trans effect = 3.26% change in DNAm per allele; Mann-Whitney p < 2.23 x 10^-308) (Figure S6). Furthermore, among cis mQTL associations, both significance and effect size increase as the distance between the genetic variant and DNAm site decreases (Figure S7). Compared to other tested DNAm sites, those associated with at least one mQTL variant (after correction for the number of tests performed [see Subjects and Methods], p < 6.52 x 10^-14) are significantly enriched in intergenic regions and less likely to be located within both gene bodies (Chi square test: p < 2.23 x 10^-308; Figure S8; Table S3A) and CpG islands (Chi square test p < 2.23 x 10^-308; Figure S9; Table S3B). We used quantitative genetic data from a study of DNAm in monozygotic and dizygotic twins1 to show that DNAm at sites associated with at least one mQTL variant is more strongly influenced by heritable (additive genetic) factors than are other tested DNAm sites (mQTL sites: median heritability, h^2 = 55% [interquartile range = 38%-71%]; all DNAm sites: median h^2 = 12% [interquartile range = 5%-31%]; Mann-Whitney p < 2.23 x 10^-308; Figure S10). Overall, the proportion of sites at which DNAm is associated with a mQTL variant increases as a function of the estimated additive genetic influence derived from twin analyses (Figure 1B). Interestingly, there is no significant difference in the contribution of additive genetic effects to variance in DNAm at sites associated with cis (median h^2 = 56%; interquartile range = 39%-72%) and trans (median h^2 = 57%; interquartile range = 32%-76%) mQTL variants (Mann-Whitney p = 0.910).

Proximal DNA-Methylation Sites Share Genetic Associations

Similar to the LD relationships that exist between proximal genetic variants, DNAm levels are often correlated between proximally located DNAm sites.14,46 To further characterize the genetic architecture of DNA methylation, we investigated whether shared genetic effects on multiple DNAm sites underlies this regional correlation structure. Although genetic variants are often associated with variation at multiple DNAm sites (Figure S4), this does not establish a shared genetic effect; shared genetic signals influencing a pair of DNAm sites might result from two distinct causal genetic variants that are in strong LD. To formally test whether neighboring DNAm sites are influenced by the same causal variant, we used a Bayesian co-localization approach44 to interrogate all pairs of DNAm sites characterized as being located within 250 kb of each other and associated with at least one significant mQTL variant at our “discovery” significance threshold (p < 1 x 10^-16). Our analyses assessed 3,535,812 pairs of DNAm sites with a median distance between DNAm sites of 110,493 bp (interquartile range = 47,914–178,085) and compared the pattern of mQTL associations for both DNAm sites to test whether they index an association with either the same causal variant or two distinct causal variants. We found that the posterior probabilities for virtually all of these (n = 3,520,781 [99.6%], median distance of 110,319 bp [interquartile range = 47,803–177,948]) supported a co-localized association within the same genomic region (PP3 + PP4 > 0.99). Of these, 281,898 pairs (8%) had sufficient support for the association of both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 and PP4/PP3 > 1; Table S4); 234,460 pairs (6.6%) had “convincing” evidence (PP3 + PP4 > 0.99 and PP4/PP3 > 5) for co-localization of the same mQTL association according to the criteria of Guo and colleagues.47 DNAm sites that shared genetic effects with at least one other DNAm site co-localize with a median of three other DNAm sites, indicating a complex relationship between genetic variation and DNAm in cis. Figure 2, for example, demonstrates that chromosome 9 contains a broad genomic region (>400 kb) where 38 DNAm sites—spanning seven genes—have a common underlying genetic signal. Of note, these DNAm sites are not contiguous; a small number of genetically mediated DNAm sites located within this region do not share the same mQTL signal. Pairs of DNAm sites with a shared causal mQTL variant are enriched for concordant directions of effect (71.2% pairs with positive correlations versus 28.8% pairs with negative correlations, binomial test p = 1.48 x 10^-323; Figure S11). Furthermore, these pairs are located relatively close together (median distance between convincing co-localized pairs = 12,394 bp [interquartile range = 1,004–49,110]), with evidence that the shared genetic architecture is structured around annotated genomic features. Co-localized pairs of DNAm sites are significantly more likely to be annotated to the same gene (OR = 6.08, Fisher’s test p < 2.23 x 10^-308) or CpG island (OR = 1.54, Fisher’s test p < 2.23 x 10^-308) than non-co-localized pairs. Where pairs of DNAm sites with a shared genetic signal are annotated to the same gene, they are nominally less likely to be annotated to the same feature than are pairs of DNAm sites annotated to different genes (OR = 0.956, Fisher’s test p = 2.52 x 10^-7), suggesting that where genetic variation influences DNAm at multiple sites across a gene these sites do not necessarily cluster by genic feature and can be located anywhere from the transcription start site to the end of the last exon. DNAm is more likely to be positively correlated between pairs of co-localized sites annotated to the same gene than between pairs of sites annotated to different genes (OR = 1.85, Fisher’s p < 2.23 x 10^-308), a result driven predominantly by pairs of DNAm sites annotated to the same feature within that gene (OR = 1.57, Fisher’s test p = 3.41 x 10^-15) rather than those annotated to different features within a gene. Finally, pairs of DNAm sites with shared genetic effects annotated to the same genic feature, although not necessarily the same gene, are more likely to be positively correlated than pairs annotated to different
genic features (OR = 1.73, Fisher’s p < 2.23 × 10⁻³⁰⁸, Figure S12).

**DNAm QTL Have Utility for Refining GWAS Signals for Complex Traits**

Genetic variants identified in GWAS analyses rarely index protein-coding changes. Instead, they are hypothesized to influence gene regulation because they are enriched in regulatory motifs, including enhancers and regions of open chromatin.⁴⁻⁶ There is considerable interest in using regulatory QTLs to refine genetic association signals and prioritize potentially causal genes within the extended genomic regions identified in GWASs.⁴⁻⁷,⁴⁸⁻⁴⁹

We next extended our previous application of the SMR approach⁴⁸ to test 126,457 DNAm sites identified at our “discovery” mQTL threshold (p < 1 × 10⁻¹⁰) against 63 complex phenotypes with GWAS data (Table S5). The first stage of the SMR approach uses the most significantly associated mQTL SNP—that has also been tested in the GWAS dataset—as an instrumental variable and implements a two-step least-squares (2SLS) approach to compare the estimated associations. Using this approach, we identified 5,848 associations (p < 3.95 × 10⁻⁷ corrected for 126,457 DNAm sites) between 40 complex traits and 5,849 unique DNAm sites (Figure S13). Because the associations identified in this way potentially reflect two highly correlated but different causal variants for the GWAS trait and DNAm, the second stage of the SMR method repeats the analysis with alternative mQTL SNPs as the instrument. If there is a single causal variant associated with both the phenotype and DNAm, the association statistics will be identical regardless of the selected instrument. In contrast, if there are two separate causal variants, each correlated with the instrument, there will be variation in the results. To distinguish between these scenarios, we applied the heterogeneity in dependent instruments (HEIDI) test to select associations with non-significant heterogeneity (HEIDI p > 0.05) and identified a refined set of 1,662 associations between 36 complex traits and 1,246 DNAm sites (Table S6).

Because the power of the SMR approach to detect pleiotropic associations reflects, in part, the power of the initial complex-trait GWAS, it is unsurprising that the highest number of SMR associations was found for traits characterized by the largest number of GWAS signals, such as height (423 significant GWAS loci, 506 SMR pleiotropic associations)⁵⁰ and inflammatory bowel disease [MIM: 266600] (168 significant GWAS variants, 127 SMR pleiotropic associations).⁵¹ In contrast, no SMR associations were found...
of the identified pleiotropic associations characterized the genic location of associated DNAm sites.

Figure S14), suggesting that there are many additional true signals in those that did not meet the stringent criteria.

Figure 3. Summary-Data-Based Mendelian Randomization (SMR) Analysis Using Quantitative Trait Loci Associated with DNA Methylation (mQTL) and Gene Expression (eQTL) Implicates a Role for LIME1 in Crohn Disease

Shown is a genomic region on chromosome 20 (chr20: 62335000–62371000) identified in a recent Crohn disease GWAS performed by Liu et al.51 Genes located in this region are shown at the top, exons are indicated by thicker bars, and the red arrows indicate the direction of transcription. The scatterplot depicts the −log10 p value (y axis) against genomic location (x axis) from the SMR analysis (where circles represent Illumina EPIC array DNA-methylation sites, squares represent gene expression probes, and solid green and red highlight those with a non-significant HEIDI test for DNA methylation and gene expression, respectively). The green and red horizontal lines represent the multiple-testing corrected threshold for the SMR test using mQTL and eQTL, respectively.

for traits with few or no genome-wide-significant SNPs; such traits included parental age at death (0–1 significant GWAS variants),52 insulin secretion rate (no significant GWAS variants),53 and whether a person has ever smoked (no significant GWAS variants).54 We compared our SMR results to those obtained with our previous mQTL dataset—generated from a smaller number of samples—and observed high rates of replication for loci that were tested in both analyses. Because our previous SMR analysis was based on a subset of 43 traits and the reduced content of the Illumina 450K array, 842 pleiotropic associations reported in the current analysis were taken forward for replication; DNAm at 519 (33.0%) of these was associated with an mQTL variant, and therefore these associations had been tested in our previous SMR study; 268 (51.6%) were characterized by significant pleiotropic association in both studies. Furthermore, the vast majority of associations tested in both datasets (516; 99.4%) were in the same direction; this was significantly more than would be expected by chance (sign test \( p = 2.72 \times 10^{-19} \); Figure S14), suggesting that there are many additional true signals in those that did not meet the stringent criteria for significance used in both studies.

In order to prioritize genes for each complex trait, we characterized the genic location of associated DNAm sites. 1,269 (76.3%) of the identified pleiotropic associations involve DNAm sites located either within a gene or less than 1500 bp from the transcription start site; this rate is significantly higher rate than that for all DNAm sites tested in our SMR analysis (OR = 1.64, Fisher’s test \( p = 1.12 \times 10^{-18} \)). To further explore these 786 pleiotropic associations—occurring between 577 genes and 32 complex traits—we extended our SMR analyses to incorporate a publicly available whole-blood gene eQTL (\( n = 5,311 \) individuals) dataset.45 Expression of 232 (40.2%) of our identified genes was significantly associated with an eQTL variant, and we used these to test for pleiotropic associations between gene expression level and the GWAS trait. These analyses provided additional support for 138 of the pleiotropic associations identified with mQTL data, supporting a relationship between 33 genes and 17 complex traits (Table S7). Figure 3, for example, highlights an association between the regulation and expression of LIME1 and Crohn disease [MIM: 266600]; this association is supported by SMR analyses incorporating both mQTL and eQTL data.
The majority of significant gene expression probes (n = 1,192; 62.3%) are associated with a median of two DNAm sites (interquartile range = 1–4) spanning a median distance of 66,846 bp (interquartile range = 19,062–155,737) at a median density of 19,959 bp (interquartile range = 6,387–54,445) between sites. Interestingly, DNAm sites pleiotropically associated with gene expression are enriched in the gene body and transcription start sites of genes and depleted intergenically (Chi square test p = \(7.08 \times 10^{-133}\); Figure S16; Table S9). We identified a small but significant enrichment of scenarios where DNAm is negatively associated with gene expression at sites located in the 5’ UTR (mean effect = \(-0.0211\); p = 0.00108), TSS200 (mean effect = \(-0.0479\); p = 6.38 \times 10^{-7} ), TSS1500 (mean effect = \(-0.0350\); p = 5.82 \times 10^{-13} ) and 1st exon (mean effect = \(-0.0506\); p = 6.19 \times 10^{-5} ), consistent with the hypothesis that promoter DNAm often represses gene expression (Figure S17).

**Using QTL Data to Refine the Genic Annotation Associated with DNAm Sites**

A key challenge in epigenetic epidemiology relates to the genic annotation of DNAm sites; such annotation is critical for the biological interpretation of significant EWAS associations. DNAm sites are usually annotated to specific genes on the basis of proximity, although the extent to which this approach is valid for inferring downstream transcriptional effects is not known. Among the identified pleiotropic associations between DNAm and gene expression, we selected instances where the DNAm site is not intergenic—i.e., <1500 bp from the transcription start site of a gene (n = 5,593 [82.3%])—and found that these were annotated to the same gene whose expression level they were associated with at a much higher rate than were DNAm sites significantly associated with expression levels at another gene (OR = 9.67; Fisher's test p < 2.23 \times 10^{-308} ). Of the 5,460 DNAm sites significantly associated with expression of at least one gene, 1,790 (32.8%) were associated with the gene they were annotated to, although 276 (5.05%) of these were also associated to an additional gene and 2,686 (50.0%) were associated with a different gene. Of note, not all CpGs were tested against the gene they were annotated to because the gene lacked a significant eQTL; this was the case for the majority of DNAm sites (n = 2,701; 80.4%) identified as being associated with a gene other than the one they were annotated to. Of particular interest are the 944 (18.3%) intergenic sites that are associated with gene expression; these potentially enable additional gene annotations for interpreting the results of EWAS analyses. Overall, although the proximity-based annotation of DNAm sites appears to be appropriate in many instances, we identified notable exceptions. For example, Figure 4A shows that the DNAm site cg00331210, located within the body of NARFL on chromosome 16, is not associated with expression of that gene but with the FAM173A gene, which is located 7.9 kilobases away. Likewise, Figure 4B shows that the DNAm site cg00072720, located within the gene body of CLDN7, is not associated with expression of that gene but with that of two other genes (ACADVL and ELPS/C17ORF81) on chromosome 17.
Discussions

In this study, we present a comprehensive assessment of the genetic architecture of DNA methylation and identify associations between common genetic variants and specific DNA methylation sites (mQTLs) by using the Illumina EPIC array. We explored our database of mQTLs to characterize genetic influences on individual and proximally located DNA methylation sites. We found that there are many instances of shared genetic signals on neighboring DNA methylation sites and that these associations are structured around both genes and CpG islands. Our results are in line with the Genome-wide Study of DNA Methylation and Expression (GeneMes) groups reported by Liu et al., who observed that multiple DNA methylation sites were influenced by overlapping genetic variants; their observations included examples where these DNA methylation sites were not contiguous. Moreover, we report that these shared genetic effects on DNA methylation are generally associated with positive correlations between the DNA methylation sites. This has implications for studies of trait-associated differentially methylated regions (DMRs) because it suggests that associations with phenotypic variation could be genetically mediated.

In an extension of our work prioritizing genes in Genome-wide Study of DNA Methylation and Expression (GeneMes)–nominated regions, we found robust agreement with our previous SNP findings (obtained from mQTLs identified with the Illumina 450K array) for shared content by using independent datasets. The additional content present on the EPIC array, however, enabled us to identify gene-trait associations not detected with the older array technology, increasing the potential yield of biological information. This augments the existing literature integrating results from Genome-wide Association Studies (GWASs) of complex traits and quantitative trait loci (QTLs) studies of gene expression and DNA methylation and substantiates the hypothesis that Genome-wide Association Studies (GWAS) variants act via gene regulation. Finally, we use these data to explore the relationship between DNA methylation and gene expression by using genetic instruments rather than correlations to infer associations between specific DNA methylation sites and genes. Although most DNA methylation sites associated with gene expression were found to be located within the gene body or close to the transcription start site, there are many relationships that challenge the commonly used gene expression in the sole basis of physical proximity. Furthermore, although the expression of most genes is associated with one or more DNA methylation sites, not all DNA methylation sites are associated with gene expression, implying that variable DNA methylation does not always have an effect on gene expression. These findings are consistent with those reported previously by Bonder et al. in their expression quantitative trait methylation (eQTM) analysis; they also report the association of multiple DNA methylation sites with each gene, the presence of both negative and positive correlations between DNA methylation and gene expression, and an enrichment of DNA methylation sites associated with gene expression in the expression quantitative trait methylation (eQTM) analysis; they also report the association of multiple DNA methylation sites with each gene, the presence of both negative and positive correlations between DNA methylation and gene expression, and an enrichment of DNA methylation sites associated with gene expression in the TSS and enhancers. Although we could only test for associations between DNA methylation sites with significant mQTLs and the expression of genes with a significant eQTL, our results provide a potentially effective method for annotating results from expression quantitative trait methylation (eQTM) analysis, particularly where the influence of DNA methylation on gene expression is hypothesized and candidates are taken forward for transcriptional analysis.

Our study has a number of important limitations. The analyses presented here are based on an unrelated subset of participants from the UK Biobank; although these represent a large sample (>1,000) of European ancestry with a broad age range, the extent to which our results are applicable to other ethnic groups characterized by a different genetic architecture is not known. Despite using the most comprehensive, high-throughput technology for profiling DNA methylation across the genome (the Illumina EPIC array), our study only assayed a small proportion of the total number of DNA methylation sites and included sparse coverage of regulatory features that are often represented by a single DNA methylation site. Moreover, DNA methylation was profiled in whole blood, which potentially limits the interpretation of candidate disease genes where the presumed tissue of interest is not blood. Given the tissue-specific nature of some mQTLs and eQTL effects, these associations should be confirmed in additional disease-relevant tissues and cell types. Although Mendelian randomization is proposed as a methodology for quantifying causal relationships between variables, it relies on a number of key assumptions, all of which also apply to mendelian randomization. Therefore, our approach did not seek to establish the direction of association between DNA methylation and outcome; we are consequently careful in our use of terminology and refrain from describing our associations as causal, especially because the SMR approach is unable to distinguish two causal variants in approximately perfect LD from one causal variant; instead, we refer to these as “pleiotropic” associations. Furthermore, given that our application of Mendelian randomization is proposed as a methodology for quantifying causal relationships between variables, we cannot rule out the possibility of horizontal pleiotropy. Finally, a limitation of the HEIDI approach to distinguishing pleiotropic associations from LD artifacts is that it looks to accept the null hypothesis of homogeneity of effects rather than reject it. However, we are confident in the set of pleiotropic associations we report given the strong replication of our previous results based on mQTLs estimated in an independent dataset.

Taken together, our results add to an increasing body of evidence showing that genetic influences on DNA methylation are widespread across the genome. We show that integrating these relationships with the results from Genome-wide Association Studies (GWAS) of complex traits and genetic studies of gene expression can improve our understanding about the interplay between gene regulation and expression and facilitate the prioritization of candidate genes implicated in disease etiology.

Accession Numbers

Individual-level DNA methylation and genetic data are available upon application through the European Genome-phenome Archive under accession code EGAS00001001232.
Supplemental Data

Supplemental Data include seventeen figures and nine tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.09.007.

Acknowledgments

We acknowledge the Wellcome Trust Sanger Institute and Ele Zeggini for generating the genotype data. Both genotyping and DNA methylation in UKHLS were funded through enhancements to the Economic and Social Research Council (ESRC) grants ES/K005146/1 and ES/N00812X/1. A.H., M.C.S., and Y.B. are supported by the ESRC (ES/M008592/1). M.K. is supported by the University of Essex and ESRC (RES-596-28-0001). E.H., J.M., and E.C.S. were supported by Medical Research Grant K013807 to J.M.. Analysis was facilitated by access to the Genome DNA methylation in UKHLS were funded through enhancements to the Economic and Social Research Council (ESRC) grants ES/K005146/1 and ES/N00812X/1. A.H., M.C.S., and Y.B. are supported by the ESRC (ES/M008592/1). M.K. is supported by the University of Essex and ESRC (RES-596-28-0001). E.H., J.M., and L.C.S. were supported by Medical Research Grant K013807 to J.M.. Analysis was facilitated by access to the Genome High-Performance Computing Cluster at the University of Essex School of Biological Sciences.

Declaration of Interests

The authors declare no conflict of interest.

Received: June 15, 2018
Accepted: September 14, 2018
Published: October 25, 2018

Web Resources

Complex Disease Epigenomics Group, http://www.epigenomicslab.com/onlinedata-resources/
European Genome-Phenome Archive (EGA), https://www.ebi.ac.uk/ega/home
GitHub, https://github.com/ehli243/UKHLS_mQTL.git
METADAC, www.metadac.ac.uk
Online Mendelian Inheritance in Man: http://www.omim.org
R coloc package, http://cran.r-project.org/web/packages/coloc
Understanding Society, https://www.understandingsociety.ac.uk

References


