EPIGENETICS
2022, VOL. 17, NO. 12, 1753–1773
https://doi.org/10.1080/15592294.2022.2079293

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

Large-scale integration of DNA methylation and gene expression array platforms identifies both *cis* and *trans* relationships

Eva E. Lancaster [a], Vladimir I. Vladimirov [b], Brien P. Riley [a,c], Joseph W. Landry [c], Roxann Roberson-Nay [a,d], and Timothy P. York [c,e]

[a]Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA; [b]Department of Psychiatry, Texas A&M University, College Station, TX, USA; [c]Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, USA; [d]Department of Psychology, Virginia Commonwealth University, Richmond, VA, USA; [e]Department of Obstetrics and Gynecology, Virginia Commonwealth University, Richmond, VA, USA

ABSTRACT

Although epigenome-wide association studies (EWAS) have been successful in identifying DNA methylation (DNAm) patterns associated with disease states, any further characterization of etiologic mechanisms underlying disease remains elusive. This knowledge gap does not originate from a lack of DNAm–trait associations, but rather stems from study design issues that affect the interpretability of EWAS results. Despite known limitations in predicting the function of a particular CpG site, most EWAS maintain the broad assumption that altered DNAm results in a concomitant change of transcription at the most proximal gene. This study integrated DNAm and gene expression (GE) measurements in two cohorts, the Adolescent and Young Adult Twin Study (AYATS) and the Pregnancy, Race, Environment, Genes (PREG) study, to improve the understanding of epigenomic regulatory mechanisms. CpG sites associated with GE in *cis* were enriched in areas of transcription factor binding and areas of intermediate-to-low CpG density. CpG sites associated with *trans* GE were also enriched in areas of known regulatory significance, including enhancer regions. These results highlight issues with restricting DNAm-transcript annotations to small genomic intervals and question the validity of assuming a *cis* DNAm–GE pathway. Based on these findings, the interpretation of EWAS results is limited in studies without multi-omic support and further research should identify genomic regions in which GE-associated DNAm is overrepresented. An in-depth characterization of GE-associated CpG sites could improve predictions of the downstream functional impact of altered DNAm and inform best practices for interpreting DNAm–trait associations generated by EWAS.

ARTICLE HISTORY
Received 16 August 2021
Revised 23 March 2022
Accepted 6 May 2022

KEYWORDS
Epigenetics; DNA methylation; gene expression; transcriptional regulation; multi-omics; data integration

Introduction

Epigenome-wide association studies (EWAS), aiming to test the theory that marks of DNA methylation (DNAm) are involved in the pathophysiology of disease, have successfully identified associations between complex traits and DNAm. Specific DNAm patterning has been associated with environmental exposures, as well as short- and long-term health outcomes [1–6]. Several attributes of DNAm potentially link this epigenetic mark to the development or progression of complex disease. Appropriate DNAm patterning is essential for normal development and ageing, and DNAm regulatory mechanisms are implicated in a multitude of molecular processes, such as cellular differentiation, X–inactivation, and genomic imprinting [7–10]. As an epigenetic mark, DNAm is both dynamic and persistent; modifiable by environmental exposures yet heritable during cell division, so that any alterations to DNAm patterns may be carried through future populations of cells [11–14]. Importantly, altered DNAm has been linked to downstream functional changes, particularly in the regulation of gene expression (GE) [15]. These properties suggest that DNAm may be contributing to mechanisms in which previous exposures and genetic predispositions can have lasting effects on disease risk. While EWAS methods are promising, their current utility beyond biomarker discovery is questionable due to study design

CONTACT Eva E. Lancaster Eva.Lancaster@vcuhealth.org Department of Psychiatry, Virginia Commonwealth University, Richmond, VA 23220, USA
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15592294.2022.2079293

limitations that impact the interpretability of results, particularly those stemming from the omission of GE measurements [14,16].

The canonical mechanism describes DNAm as a repressor of proximal transcription, in which DNAm within promoter regions is able to silence GE by either blocking the binding of essential transcriptional machinery or by recruiting chromatin modifying proteins that transition the local DNA conformation to a more heterochromatic state [9,17]. Despite accumulating evidence that suggests this model is overly simplistic when applied on a global scale, [17,18] many researchers rely on this paradigm to interpret an association between DNAm and a disease of interest [14,16,19]. In a typical EWAS, any significantly associated CpG sites (also known as differentially methylated positions or DMPs) are each mapped to their most proximal gene and the biological function of those genes is reported in the context of the tested phenotype. This method is problematic since the canonical DNAm mechanism applies specifically to CpGs within promoter regions whereas many DMPs are intergenic, within gene bodies, or downstream of the proposed gene. Given that this interpretation emphasizes the functional relevance of specific genes to disease biology, an argument can be made that current EWAS are primarily interested in examining a theory of DNAm-driven transcriptional regulation [14,16]. By inferring transcriptional activity from DNAm–trait associations, this approach relies on assumptions without directly testing for functional evidence. Given that accurately inferring the functional consequences of modified DNAm at any particular site is still very limited, this practice may lead to inaccurate conclusions about disease biology [18,20].

Accurately predicting the functional impacts of altered DNAm remains challenging, in part, due to the limited characterization of genome-wide DNAm–GE relationships [20]. DNAm often does not block transcription independently but rather works in concert with other regulatory elements to coordinate GE [21–23]. These regulatory mechanisms involve a complex crosstalk between DNAm, higher-order chromatin modifiers, and other epigenetic marks, further contributing to difficulties in determining the functional impact from DNAm measurements alone [21–23]. Moreover, linking genes to their putative regulatory regions is not always straightforward [24]. DMPs are often located outside of proximal regulatory elements, within intergenic or intronic regions with no known regulatory function. Since a frequently utilized approach for interpreting these results involves linking all DMPs to their nearest gene, any features of the genomic landscape beyond distance are disregarded. Even if CpG-GE pairs are identified, predicting the regulatory consequence of altered DNAm remains difficult as exceptions to the theory that describes DNAm as a proximal gene repressor have accumulated. For example, increased DNAm, particularly within the gene body, is frequently positively correlated with local transcription [25–30]. Although mechanisms linking hypermethylation to increased GE are still unclear, a recent study identified more transcription factors that preferred binding methylated sequences than those inhibited by DNAm [23]. These functional complexities suggest that assumptions regarding transcriptional activity should not be inferred by DNAm patterns alone. Instead, if the fundamental theory being explored is a mechanism of transcriptional regulation modulated by DNAm, measurements of GE should be included in the analysis [14,16].

Multi-omic studies integrating global DNAm and GE measurements can provide evidence for DNAm-driven transcriptional regulatory mechanisms. Measuring GE in parallel with DNAm could allow for not only a direct test of the association between DNAm and the outcome of interest, but also whether this relationship can be explained by a mediator, in this case, GE [31]. This approach tests the hypothesized mechanism while avoiding assumptions regarding the regulatory function of DNAm that typically cloud the interpretation of EWAS results. Although cross-sectional studies are still unable to eliminate the possibility of reverse causation (i.e., GE changes proceeding DNAm changes), they can provide a more comprehensive understanding of biological processes involved in disease. However, current studies often integrate DNAm and GE measures by discovering differentially methylated and differentially expressed genes separately, and regulatory mechanisms are inferred by the observation of overlaps between DNAm–

trait and GE–trait associations at the gene level [32–39]. Since the relationship between DNAm and GE is not tested *a priori*, this method still relies on assumptions that DMPs strictly influence expression of the nearest transcript. A limitation impacting current studies is the lack of information on the extent that DNAm loci are associated with GE in *cis* or *trans*, which could vary by, for instance, cell type or developmental stage.

An extended EWAS approach integrating both DNAm and GE measurements holds promise in uncovering biological processes important to the development or progression of disease, however, a mechanistic interpretation requires prior knowledge regarding specific DNAm–GE relationships across the genome, which has yet to be resolved. Several studies have attempted to clarify this functional relationship by integrating DNAm and GE measurements [26,27,29,40,41]. Although a variety of complex relationships between DNAm and GE have been identified, including long-range associations, focus has been primarily placed on outlining proximal relationships and relatively few studies have examined distal associations on a genome-wide scale [26,41]. The objective of this study was to expand upon this work by cataloguing and characterizing the relationships between DNAm and both proximal and distal GE (i.e., *cis* and *trans* relationships, respectively) in peripheral blood, a tissue commonly assayed in EWAS. To identify attributes that replicate across disparate samples, analyses were conducted in two previously described cohorts, the Adolescent and Young Adult Twin Study (AYATS) [6,42], and the Pregnancy, Race, Environment, Genes Study (PREG) [43]. To our knowledge, this is the first study to test for genome-wide associations between DNAm and GE in two cohorts of the same tissue.

Methods

Study cohorts

Adolescent and Young Adult Twin Study (AYATS)

The AYATS study was designed to examine genetic and environmental contributions to internalizing pathways (e.g., depression and anxiety) during development. A sample of monozygotic twins were chosen for their adherence to the study's inclusion criteria (e.g., 15–20 years of age, no current use of psychotropic medications) [6,42]. Peripheral blood collected from 141 participants at a single time point was assayed for both DNAm and GE. An overview of study characteristics and further demographic information can be accessed in the supplement (Supplementary Table S1).

Pregnancy, Race, Environment, Genes (PREG) study

The PREG Study is a prospective longitudinal study with the purpose of identifying how environmental determinants of health and DNAm remodelling relate to racial health disparities in perinatal health outcomes [43]. Of the 240 women who enrolled in the study, 177 met all birth and pregnancy inclusion criteria (e.g., mother and father self-identify as either both Caucasian or both African American) and no exclusion criteria (e.g., preeclampsia, fetal congenital anomaly, placental anomaly, fewer than 3 study time points completed). Peripheral blood samples were collected up to four times throughout pregnancy. Sample collection was scheduled during gestational weeks 0–15, 10–25, 20–40, and 37–42. DNAm was assessed at all time points, whereas GE was measured once at the final collection during weeks 37–42. Only those DNAm measurements from specimen simultaneously collected with GE were analysed in this study. A total of 151 women had concomitant DNAm and GE measured. An overview of study characteristics and further demographic information can be found in the supplement (Supplementary Table S2).

DNAm measurement and data processing

In both samples, DNAm and GE was measured from peripheral blood. The Infinium 450k HumanMethylation BeadChip assayed genome-wide DNAm and the Affymetrix HG-U133A 2.0 array measured GE. A description of platform characteristics as well as the methods used for measurement and preprocessing can be found in the supplement.

Table 1. Study characteristics.

	AYATS	PREG
N	137	131
Study phenotype	Internalizing disorders (e.g., early-onset major depression)	Perinatal health outcomes (e.g., preterm birth)
Age	16.96 (1.28)	29.06 (4.99)
Sex (% female)	97 (71%)	131 (100%)
Ethnicity (% Caucasian)	132 (97%)	67 (51%)
Methylation probes tested	445,120	421,729
Expression probes tested	10,913	12,249

Study characteristics were assessed after preprocessing and removal of poor quality samples.
Mean (standard deviation) or *N* (%).
Abbreviations. AYATS = Adolescent and Young Adult Twin Study; PREG = Pregnancy, Race, Environment, Genes cohort.

Association analysis

The relationship between all pairwise combinations of measured DNAm and GE (Table 1) was tested by linear regression in the R statistical environment (version 3.5) [44]. Log-transformed expression values (dependent variable) were regressed on DNAm M-values (independent variable), while covariates controlled for differences in cell type heterogeneity. Cell type proportions were derived from the Houseman algorithm, which estimates proportions for granulocytes, monocytes, CD8-positive T cells, CD4-positive T cells, B lymphocytes, and natural killer cells based on cell type-specific DNAm profiles [45]. Granulocytes were selected to account for overall differences in cell type proportions based on high correlations with other cell type estimates (absolute correlations ranged from 0.47 to 0.71), and included as a covariate in all models. Natural killer proportions were included in AYATS models exclusively to adjust for the atypical variation in this cellular fraction characteristic of depressed patients [6,46].

Additional covariates were selected to adjust for potential confounding influences specific to the characteristics of each cohort, while also maintaining a similar analytical approach across the two studies. Since PREG is a racially diverse sample (Table 1), ancestrally informative principal components were estimated from the DNAm data using the method described in Barfield et al. The third principal component was highly correlated with self-reported race and included as a covariate in the PREG cohort models [47]. A linear mixed-model framework was used to account for twin structure in the AYATS cohort [48]. The limma Bioconductor package was used to estimate within-family correlations from 1,000 randomly sampled CpGs in order to appropriately adjust model standard errors and account for the non-independence of twin pair DNAm observations [48].

Both DNAm and GE measurements were adjusted for technical artefacts prior to analysis (see supplement), so that variables related to slide or row effects were not included as covariates in subsequent analyses.

Although measurements were generated using the same technology in both cohorts, differing numbers of probes remained after quality control procedures (Table 1). A within-study Bonferroni correction was used to adjust for multiple testing at an alpha threshold of 0.05. While estimates of genomic inflation are typically used to identify spurious associations driven by artefacts in genome-wide association studies (GWAS), it has been recently suggested that inflated test statistics should be similarly reviewed in epigenetic studies [49]. To mitigate the presence of false positives, genomic inflation was assessed using the method described in Kennedy et al. [26]. Briefly, genomic inflation factors were calculated for each transcript, across all CpG associations, as the median $(\text{T-statistic})^2/0.4549$. Appropriate thresholds for test statistic inflation are not as well established in the epigenetics field. To facilitate cross-study comparisons, any transcript with an inflation factor > 2 was flagged for removal [26].

Every pairwise relationship between measured DNAm and GE was modelled and classified as either *cis* or *trans*, since molecular mechanisms linking proximal DNAm may differ from more long-range interactions. DNAm–GE pairs were in *cis* if the CpG site was located within a gene or 2,500 base pairs (bp) upstream. This extension is expected to capture important transcript-specific regulatory regions, given that many promoters are located up to 1 kilobase upstream of the transcriptional start site (TSS) [50]. CpG–transcript pairs located outside this range were categorized as *trans* relationships, with the rationale that more distal regulatory features (e.g., enhancers) may be responsible for the relationship between CpG methylation and transcript expression.

Characterization of results

CpG sites were mapped to biologically relevant annotations to test for feature enrichment among sites significantly associated with GE. Annotation selection was based on evidence that local CpG density, gene feature location, and proximity to regulatory elements are important characteristics that may impact the functional consequences of DNAm [17]. Transcription factor activity is regulated by DNAm, and a history of transcription factor binding also appears to influence the susceptibility of CpG methylation in specific locations [23,51–53]. Moreover, processes involving transcription factors were previously enriched among CpGs associated with GE [26,54]. Similarly, non-coding RNAs are regulated by methylation patterning, while also contributing to the regulatory activity of DNAm [55,56]. Chromatin states are accurately able to distinguish variable transcriptional activity by describing the specific patterns of histone modifications that impact the regulation of GE [57,58]. Histone modifications are intricately linked to both DNAm and GE, potentially serving to mediate the influence of methylation on transcription [17,57].

Selected features described local CpG densities (UCSC CpG island classifiers and HIL annotations) [59,60], genomic regions (UCSC knownGenes track, hg19 build) [61], chromatin states (ENCODE 15-state ChromHMM) [58], transcription factor binding (ENCODE TF ChIP-seq) [62], non-coding RNAs (GENCODE version 37) [63], and other annotations related to regulatory activity (i.e., FANTOM5-defined enhancer and ENCODE-defined insulator regions) [64]. All cell type-specific annotations (e.g., chromatin states, enhancer regions, etc.) were defined in the lymphoblastoid cell line GM12878. The specific gene feature annotations that were sourced using Bioconductor packages can be found on the Open Science Framework project page (https://osf.io/dk3cg/).

Enrichment analyses were performed separately for *cis* and *trans* groups. The proportion of significant findings annotated to each category was compared to the proportion of total number of tested CpG sites using Fisher's exact test. Duplicate mappings were conserved for gene region annotations, so that individual CpG sites could be annotated to more than one gene region. Other annotations assigned each CpG to mutually exclusive categories (e.g., chromatin states and CpG classifiers). A Bonferroni correction for 20 enrichment tests was used to adjust for unique annotation categories examined (e.g., CpG density classifiers, chromatin states, transcription factor binding, etc.)

Results

Participant demographics and initial findings

After performing the preprocessing procedures described in the supplementary methods, all 137 of the remaining GE measurements also had corresponding DNAm of sufficient quality in AYATS. In PREG, 131 samples had both DNAm and GE that passed quality control. While the tissue and platforms were consistent across studies, these cohorts differed in other characteristics (Table 1). PREG was an older (aged 18–40 years) and more racially diverse sample, with 49% of participants identifying as African American (Supplementary Table S2). Notably, all participants in the PREG sample were pregnant women, while the AYATS sample consisted of both male (29%) and female (71%) adolescents (aged 15–20 years; Supplementary Table S1).

Genome-wide methylomic and transcriptomic data was generated using the Illumina HumanMethylation 450k BeadChip and Affymetrix HG-U133A 2.0 array, respectively. After performing quality control procedures separately in both cohorts, a differing number of probes were identified as poor quality. In

Table 2. Overview of pairwise DNAm–GE Association results.

Study	Significant associations[a]	Direction of Relationship	Unique CpGs	Unique transcripts[d]	Mean adjusted R-squared[e]
AYATS[b]					
cis	169	65% negative	107	42	0.58 (0.11)
trans	734	78% negative	266	96	0.48 (0.10)
PREG[c]					
cis	121	79% negative	87	31	0.46 (0.10)
trans	258	49% negative	156	43	0.40 (0.08)

[a]After Bonferroni adjustment for total number of tests performed within cohort.
[b]P-value < 1.03× 10^{-11}.
[c]P-value < 9.68× 10^{-12}.
[d]Defined as unique by Entrez identifier.
[e]Mean (standard deviation).
Abbreviations. AYATS = Adolescent and Young Adult Twin Study; PREG = Pregnancy, Race, Environment, Genes cohort.

AYATS, 40,392 DNAm probes and 10,809 GE probe sets were removed during preprocessing, while 63,783 DNAm and 9,473 GE measurements were removed in PREG (Table 1).

Associations between DNA methylation and gene expression

Overall findings

An overview of significant DNAm–GE relationships is presented in Table 2 (see the Open Science Framework project page at https://osf.io/dk3cg/ for summary statistics). Due to the differing number of measurements surviving quality control, associations with P-values < 9.68 $\times 10^{-12}$ (0.05 alpha corrected for 5,165,758,521 total tests) in PREG and P-values < 1.03 $\times$ 10^{-11} (0.05 alpha corrected for 4,857,594,560 tests) in AYATS were considered significant (Table 2). A total of 903 associations were identified in the AYATS cohort, 169 of which were in *cis* (4.72 $\times$ 10^{-61} < P < 1.01 $\times$ 10^{-11}) and 734 in *trans* (2.64 $\times$ 10^{-61} < P < 1.03 $\times$ 10^{-11}).

Within the PREG sample, 379 DNAm–GE associations were statistically significant, of which 121 were *cis* (5.15 $\times$ 10^{-58} < P < 8.50 $\times$ 10^{-12}) and 258 *trans* (2.86 $\times$ 10^{-53} < P < 9.51 $\times$ 10^{-12}). Since GE probe sets measuring expression of the same gene were retained, some transcripts and CpG sites are represented more than once in the results. A total of 340 unique CpG sites and 105 unique genes comprised the 903 significant associations identified in AYATS, while 228 CpGs and 69 genes were unique in PREG across both *cis* and *trans* relationships (total n = 379). Across all categories (i.e., AYATS/PREG *cis*/*trans*), many significant relationships occurred between one transcript and one CpG site (Supplementary Figures S1 and S2), although instances in which a single CpG site was associated with multiple transcripts, and vice versa, were also common. Both positive and negative relationships were identified, although the

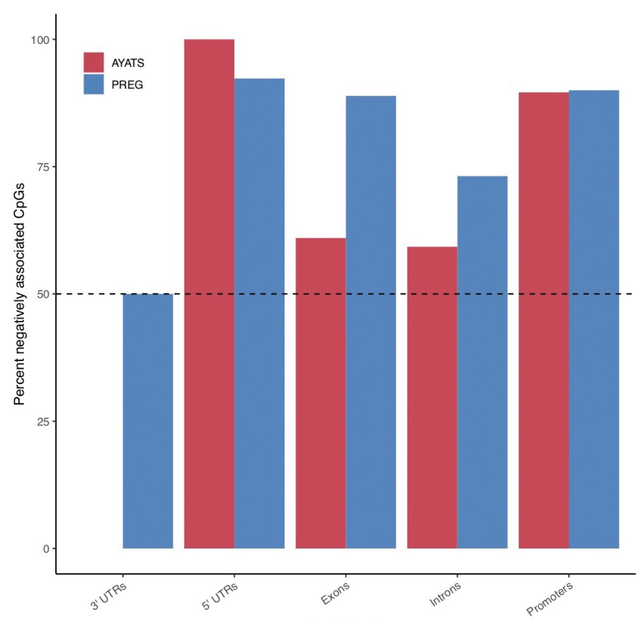

Figure 1. Percent negative CpG-GE associations by gene region. With the exception of 3′ untranslated regions (UTRs), the majority of *cis* CpG-GE relationships were negative across gene regions in both the AYATS (red) and PREG (blue) cohorts. Promoters and 5′ UTRs had the highest fraction of negative associations, aligning with canonical descriptions of promoter DNAm as a repressor of local gene transcription.

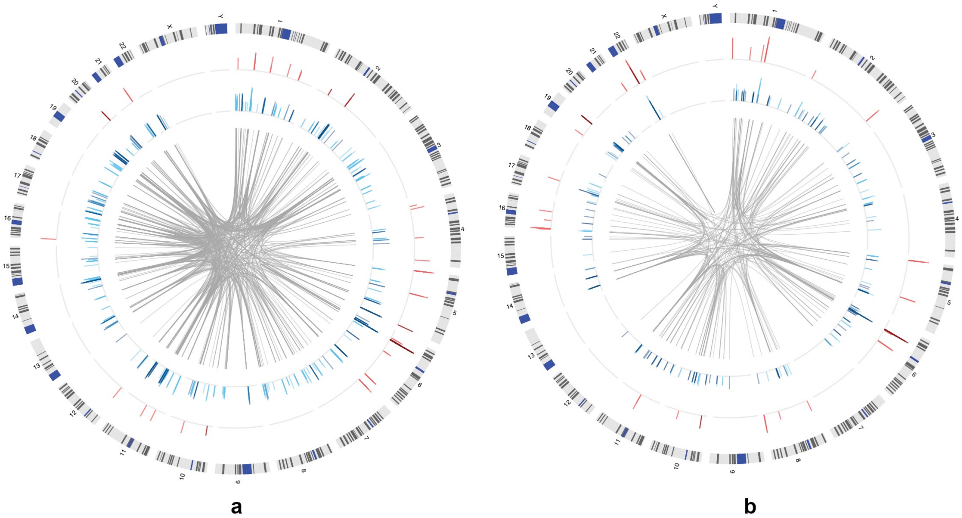

Figure 2. Distribution of significant connections between DNA methylation and transcript expression across the genome in the AYATS (2a) and PREG (2b) cohorts. The location of significant *cis* (red track) and *trans* relationships (blue track) across the genome (ideogram of human chromosomes, outer track) is shown. Bar graphs show the direction of the relationship (positive relationships are shown in the darker color) and the relative magnitude of the effect (height of bars; defined by adjusted R-squared values). *Trans* CpG-GE relationships often spanned chromosomes (location of associated CpG-GE pairs shown by center grey links).

majority of significant associations had negative coefficients (49% to 78% negative across tested categories; Table 2 and Figure 1). Effect sizes were relatively large throughout (adjusted R-squared range = 0.23–0.90), with *cis* DNAm explaining more GE variability on average (mean adjusted R-squared = 0.58 and 0.46 for AYATS and PREG, respectively) when compared to *trans* (mean adjusted R-squared = 0.48 and 0.40 for AYATS and PREG, respectively).

AYATS Associations. The distribution of significant *cis* and *trans* connections is shown in Figure 2(a). On average, each significant *cis* CpG site was associated with 1.58 transcripts (median = 1, range = 1–4; Supplementary Figure S1). *Trans* CpGs were more likely to associate with multiple transcripts than *cis* (mean = 2.75, median = 1, range = 1–22). Effect sizes, defined by adjusted R-squared values, ranged from 0.23 to 0.90. *Cis* DNAm explained more variation in GE on average (Welch's *t*-test $P = 2.2 \times 10^{-16}$). DNAm–GE relationships were predominantly negative (65% of *cis* and 78% of *trans* relationships; Table 2 and Figure 1).

PREG Associations. The distribution of significant *cis* and *trans* connections is shown in Figure 2b. On average, each significant *cis* CpG site is associated with 1.39 transcripts (median = 1, range = 1–4; Supplementary Figure S1). Significant *trans* CpGs were more likely to associate with multiple transcripts (mean = 1.65, median = 1, range = 1–11). Effect sizes ranged from 0.31 to 0.86, with *cis* DNAm explaining more variation in GE on average (Welch's *t*-test $P = 5.10 \times 10^{-8}$). *Cis* DNAm–GE relationships were predominantly negative (79%) while *trans* relationships were split almost equally between positive and negative associations (Table 2).

Between study comparison

A total of 86 individual DNAm-transcript pairs replicated across cohorts (57 *cis* and 29 *trans*). In *cis*, 34% of significant CpG-GE pairs identified in AYATS were replicated, and 47% of those found in PREG overlapped with AYATS. In *trans*, only 4% of AYATS and 11% of PREG connections were replicated. Notably, all overlapping *trans* relationships consisted of same-chromosome CpG–transcript pairs. Replicated *trans* CpGs were located anywhere from 2,654 to 91,895 bp away from their associated gene (mean = 22,826 bp). On average, relationships that replicated across cohorts had stronger effects than those that were cohort-specific (Welch's *t*-test $P = 1.9 \times 10^{-7}$).

Enrichment analyses characterizing those GE-associated CpG sites that replicated across cohorts identified similar attributes to the whole sets of

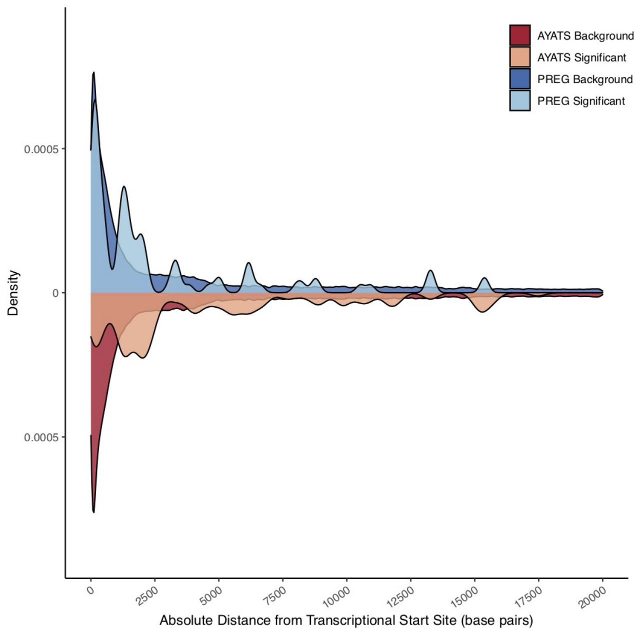

Figure 3. Absolute distance between CpG probes and transcriptional start sites (TSS) of proximal genes. A density plot depicting the absolute distance of DNA methylation microarray probes (darker colour) relative to the GE-associated CpG sites (lighter colour) from the transcriptional start sites of *cis* genes. Both the AYATS (red) and PREG (blue) cohorts showed enrichment in areas flanking active transcription. The proportion of GE-associated CpGs compared to the CpGs represented on the microarray was highest in areas directly downstream of the TSS.

significant CpGs. Replicated GE-associated CpG sites were significantly ($P < 0.0025$) enriched in South shores ($P = 7.0 \times 10^{-5}$), regions of transcription factor binding ($P = 2.7 \times 10^{-6}$), in chromatin states corresponding to areas flanking active transcription (chromatin state 3 $P = 0.0008$) and zinc-finger binding sites (chromatin state 4 $P = 0.0024$), and within introns ($P = 3.5 \times 10^{-7}$). These results were not significant when the replicated relationships were compared to all significant relationships rather than all tested relationships, indicating that similar proportions of replicated CpG sites and cohort-specific CpG sites were located within these regions.

Location relative to transcriptional start sites

Previous research has observed that methylation status at CpGs adjacent to gene TSS is more likely to correlate with proximal GE than that at CpG sites within other genomic regions [26,54,65]. To explore this topic further, CpG sites associated with GE in *cis* were mapped to their associated TSS, as defined by gene annotations from the UCSC hg19 build knownGene track [66]. This annotation defines the most 5' TSS as the primary gene TSS. The location of

Table 3. Results of AYATS Enrichment Analyses.[a]

Annotation	*cis*		*trans*	
	Enriched	Depleted	Enriched	Depleted
Chromatin States[b]	**TxFlnk (3),** **ZNF/Rpts (8)**	**Tx (4),** **TxWk (5),** Quies (15)	**TssAFlnk(2),** **TxFlnk (3),** **TxWk (5),** **EnhG (6),** **Enh (7)**	**TssA (1)** **Het (9),** TssBiv (10), BivFlnk (11), **ReprPC (13),** **ReprPCWk (14)**
CpG Classifiers	South Shore	North shore, **Island**	**South shore,** **Open sea**	**North shelf,** **North shore,** **Island**
Gene Regions		**5'-UTRs,** **Promoters**	**Introns**	5'-UTRs, **Promoters**
Other[c]	Enhancers, **TF binding**	lncRNAs	**Enhancers**	**lncRNAs,** **Insulators**

Abbreviations. UTR = untranslated region; TF = transcription factor; lncRNAs = long non-coding RNAs.
Bolded items. *P*-value <0.0025 (Bonferroni corrected for 20 tests)
Underlined items. Concordance across both the AYATS and PREG study.
[a]*P*-value < 0.05.
[b]ENCODE ChromHMM 15-state model; 1 = Active transcriptional start site (TSS), 2 = Flanking active TSS, 3 = Flanking strong transcription, 4 = Strong transcription, 5 = Weak transcription, 6 = Genic enhancer, 7 = Active enhancer, 8 = Zinc-finger genes and repeats, 9 = Heterochromatin, 10 = Bivalent/poised TSS, 11 = Flanking bivalent TSS, 12 = Bivalent Enhancers, 13 = Polycomb-repressed, 14 = Weak Repressed Polycomb, 15 = Quiescent.
[c]FANTOM5-defined enhancers, transcription factor binding sites derived from ENCODE TF ChIP-seq, GENCODE long non-coding RNAs.

Table 4. Results of PREG Enrichment Analyses.[a]

Annotation	*cis*		*trans*	
	Enriched	Depleted	Enriched	Depleted
Chromatin States[b]	TssA (1), <u>**ZNF/Rpts (8)**</u>	**TssAFlnk (2),** <u>**Tx (4)**</u>	<u>**TssAFlnk(2),**</u> <u>**TxFlnk (3),**</u> <u>**Enh (7),**</u> **EnhBiv (12)**	**TxWk (5),** <u>**Het (9),**</u> <u>**ReprPCWk (14),**</u> **Quies (15)**
CpG Classifiers	<u>South Shore</u>		**North shore**, <u>South shore</u>	<u>**Island**</u>
Gene Regions			Exons	<u>Promoters,</u> 3′-UTRs
Other[c]	<u>**TF binding**</u>		<u>**Enhancers,**</u> <u>**TF binding**</u>	

Abbreviations. UTR = untranslated region; TF = transcription factor; lncRNAs = long non- coding RNAs.
Bolded items. *P*-value <0.0025 (Bonferroni corrected for 20 tests)
Underlined items. Concordance across both the AYATS and PREG study.
[a]*P*-value < 0.05.
[b]ENCODE ChromHMM 15-state model; 1 = Active transcriptional start site (TSS), 2 = Flank- ing active TSS, 3 = Flanking strong transcription, 4 = Strong transcription, 5 = Weak tran- scription, 6 = Genic enhancer, 7 = Active enhancer, 8 = Zinc-finger genes & repeats, 9 = Heterochromatin, 10 = Bivalent/poised TSS, 11 = Flanking bivalent TSS, 12 = Bivalent En- hancers, 13 = Polycomb-repressed, 14 = Weak Repressed Polycomb, 15 = Quiescent.
[c]FANTOM5-defined enhancers, transcription factor binding sites derived from ENCODE TF ChIP-seq, GENCODE long non-coding RNAs.

cis CpG sites relative to the TSS of their associated gene is shown in Figure 3. Although many significant sites were located near the 5' end of gene boundaries, these areas are also overrepresented in the 450k microarray. Overall, the relative proportion ([number GE-associated CpGs within 2500 bp/total number GE-associated CpGs]/[number microarray CpGs within 2500 bp/total number microarray CpGs]) of TSS-proximal CpGs was higher in the GE-associated CpG sites compared to the microarray background. Interestingly, this observation was driven by CpG sites located downstream of the TSS. The relative proportion of CpG sites 2500 bp upstream of the TSS was lower in GE-associated CpGs than was present on the microarray, while the opposite relationship was observed immediately downstream of the TSS (relative proportion upstream = 0.58 and 0.65; relative proportion downstream = 1.59 and 2.21 in AYATS and PREG, respectively).

Enrichment analyses

CpG sites were annotated by genomic regions, local CpG densities, chromatin states, bound transcription factors, and other related regulatory regions (e.g., insulator regions, regulatory RNAs). Enrichment tests were then performed within annotation type. An overview of results from enrichment analyses is outlined in Table 3 (AYATS) and Table 4 (PREG). Annotation categories with *P*-values <0.0025 exhibited significant depletion or enrichment, while *P*-values <0.05 were considered suggestive. A number of depleted and enriched categories overlapped between the two cohorts (underlined in Tables 3 and 4). Overall, regions of high CpG density were depleted across all groups (Supplementary Table S3 and Supplementary Figures S3 and S4) while annotations indicative of regulatory activity (e.g., transcription factor binding, enhancers) were enriched among GE-associated CpGs (Tables 3 and 4).

Characterization of *cis* connections

South shore regions (i.e., shore regions located downstream from a CpG island) were either significantly ($P < 0.0025$) or suggestively ($P < 0.05$) enriched across all groups (AYATS/PREG and *cis*/*trans*; Figure 4). In addition to CpG classifiers, HIL annotations were used to describe local CpG density [60]. Regions of low CpG density were enriched in the AYATS cohort, while high-density regions were consistently depleted across AYATS and PREG (Table S3 and Figure S3). CpG islands are often associated with promoters, and both of these annotations were depleted in AYATS but were neither significantly enriched

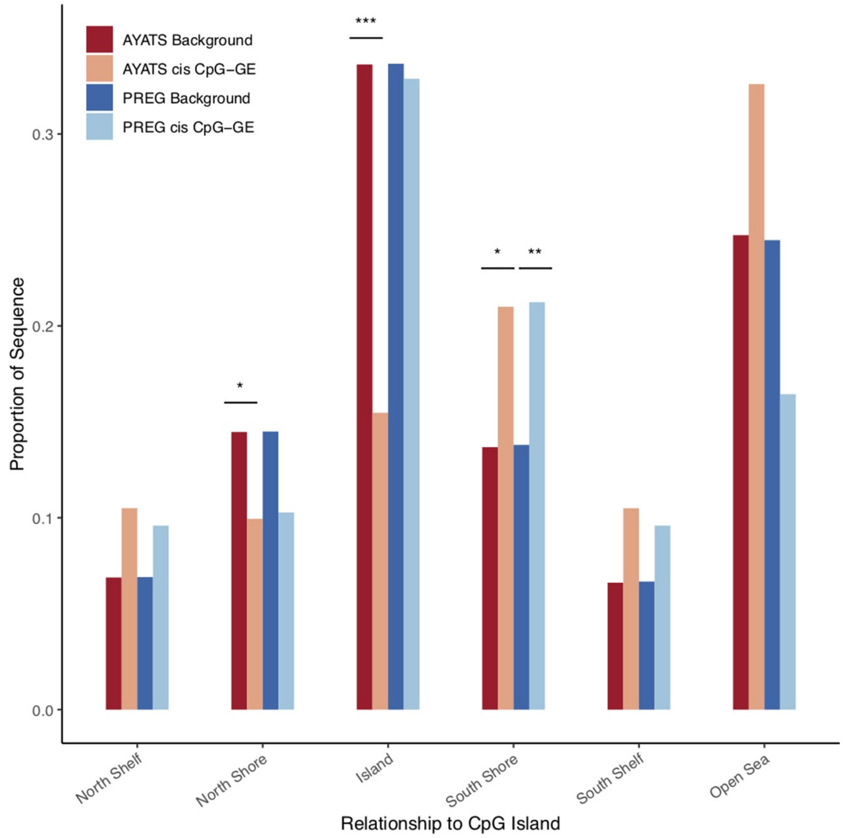

Figure 4. Enrichment for CpG classifiers in *cis* CpG–transcript relationships. CpG classifiers based on the distribution around CpG island regions were defined by the UCSC hg19 knownGene track. Islands and regions directly upstream from islands were depleted in AYATS. However, downstream regions bordering islands (South shores), were significantly enriched in both cohorts (***$P < 0.0005$; ** $P < 0.005$; * $P < 0.05$).

or depleted in PREG (Figures 4 and 5). Transcription factor binding sites, defined by significant peaks identified in ChIP-seq analyses of 134 transcription factors in lymphoblastoid cells [62], were enriched in both cohorts (Figure 6). Chromatin state characteristics, which assign a function to genomic regions based on the presence of specific histone methylation marks, also showed some concordance between the two studies (Figure 7). Specifically, zinc-finger genes and repeats were consistently found to be enriched, whereas areas of strong transcription were consistently depleted. Regions flanking active transcription were more variably assigned, with one category found to be enriched in AYATS (areas flanking strong transcription) and another depleted in PREG (areas flanking active TSS).

Characterization of *trans* connections

Like *cis* CpG-GE pairings, CpGs associated with GE in *trans* were overall depleted in areas of high CpG density (Table S3 and Figure S4) and within the promoter regions of genes, while South shore regions and regions of intermediate CpG density were enriched (Tables 3–4 and Figures 8–9). Again, GE-associated CpG sites were overrepresented in areas of known regulatory importance, such as sites of transcription factor binding and enhancer regions (Figure 10). With the exception of lncRNAs, which were depleted in AYATS, noncoding RNAs were neither over- or underrepresented. The chromatin state analysis highlighted distinct differences between *cis* and *trans* results. Chromatin states reflecting enhancer regions were enriched in both AYATS and PREG, as

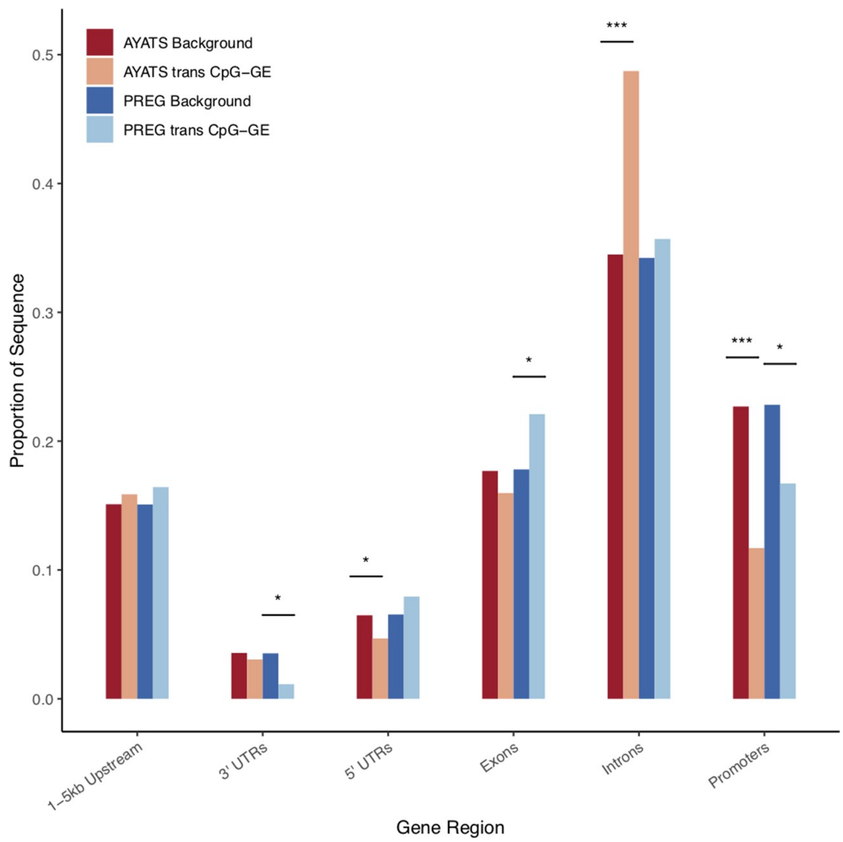

Figure 5. Enrichment for gene regions in *cis* CpG–transcript relationships. Gene regions were annotated based on the UCSC hg19 knownGene track. GE-associated CpG sites were depleted in 5′ untranslated regions (UTRs) and in promoters in the AYATS cohort only. (*** $P < 0.0005$; **$P < 0.005$; * $P < 0.05$).

were areas flanking sites of active transcription. Repressed states, including heterochromatic regions and polycomb-repressed regions, were consistently depleted (Figure 11).

Functional enrichment analysis

The Gene Ontology (GO) Consortium and the Kyoto Encyclopaedia of Genes and Genomes (KEGG) were used to assess overrepresented gene functions and pathways within significant *cis* results (see supplement for more information). Terms with a false discovery rate (FDR) <0.05 were deemed significant [67]. Common themes were uncovered in both cohorts (Supplementary Tables S4–S11), and include functions related to the activation and regulation of immune response and cellular detoxification. A total of 33 significantly enriched GO terms overlapped between the two cohorts, and all significant KEGG pathways identified in AYATS (n = 31) were also found in PREG (n = 39). However, these consistencies were supported by relatively few genes.

Discussion

Although genome-wide epigenetic studies aim to uncover the role of DNAm in disease development and progression, they often do not utilize an experimental framework that provides evidence for a mechanistic relationship. Most EWAS operate under the assumption that DNAm influences proximal gene transcription. However, the absence of measured GE makes relying on this interpretation difficult, especially as mounting evidence suggests that DNAm does not always follow a canonical *cis* relationship [17]. Given the complicated network of interactions between DNAm, GE, higher-order chromatin modifiers, and other regulatory elements, it is challenging to draw accurate conclusions about the downstream functional effects of altered DNAm without, at minimum, integrating concomitant measurements of GE

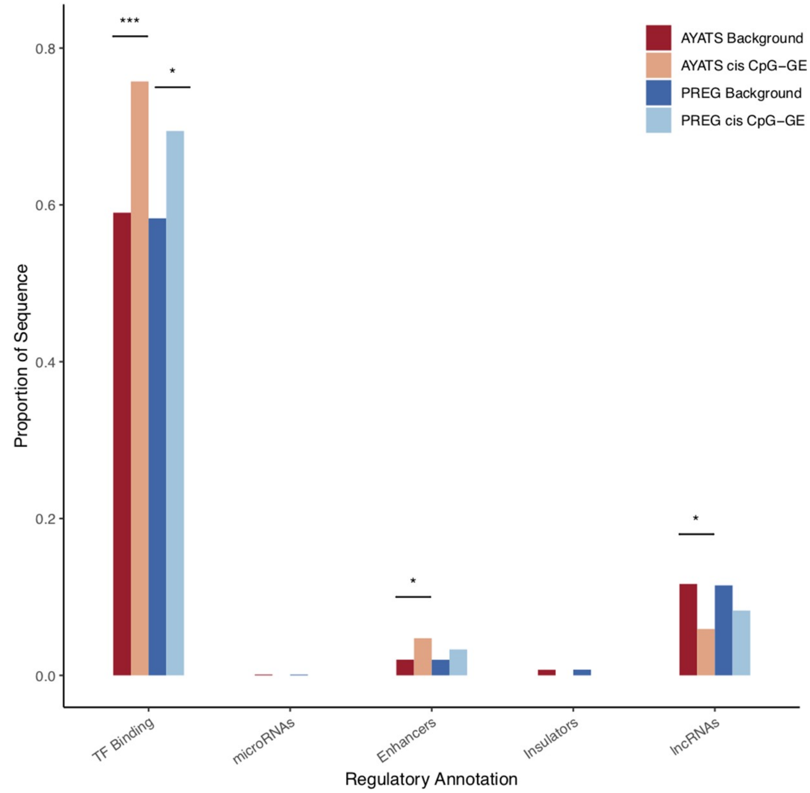

Figure 6. Enrichment for additional regulatory annotations in *cis* CpG-transcript relationships. Sites of transcription factor binding, as defined by ENCODE TF ChIP-seq annotations, were significantly enriched across cohorts. FANTOM5 enhancers were enriched in AYATS (*** $P < 0.0005$; ** $P < 0.005$; * $P < 0.05$).

[14,17,22,62]. To investigate the relationship between DNAm and GE further, this study tested genome-wide associations between DNAm and GE in peripheral blood collected from two cohorts. Both proximal and distal relationships were identified, highlighting potential inaccuracies in the current functional interpretation of trait-associated CpG sites within the frequently adopted EWAS framework.

Across cohorts, DNAm was significantly associated with both proximal (*cis*) and distal (*trans*) GE. The primary findings of this study align with other reports of long-range DNAm–GE relationships, adding to the growing body of literature questioning the accuracy of current EWAS interpretations [18,26,68]. The effect sizes detected among DNAm–GE pairs were relatively large, with DNAm predicting 42–50% percent of GE variability on average. Although this result is likely influenced by a lack of statistical power to detect more attenuated relationships, it reiterates that while DNAm may not be an appropriate proxy for changes in GE, strong links between the two measurements exist. Approximately 23% of connections identified in PREG were also significant in the AYATS cohort, suggesting a consistent programme of gene regulation even among the disparate cohorts tested. While this proportion is similar to DNAm–GE connections identified in peripheral blood and isolated monocytes [26], discrepancies between the two cohorts could be related to differences in statistical power or differences in demographic and clinical features (i.e., genetic ancestry, developmental stage, etc.). On average, *cis* connections were more likely to replicate between studies and account for larger proportion of GE variability when compared to within-cohort *trans* associations. Larger samples are likely necessary to detect more subtle *cis* and *trans* CpG-GE pairings and provide a balanced assessment of the expected replication across samples.

Interpreting DNAm–disease relationships is hindered not only by limitations in identifying DNAm–GE pairs but also by challenges in predicting the precise functional impact of altered DNAm

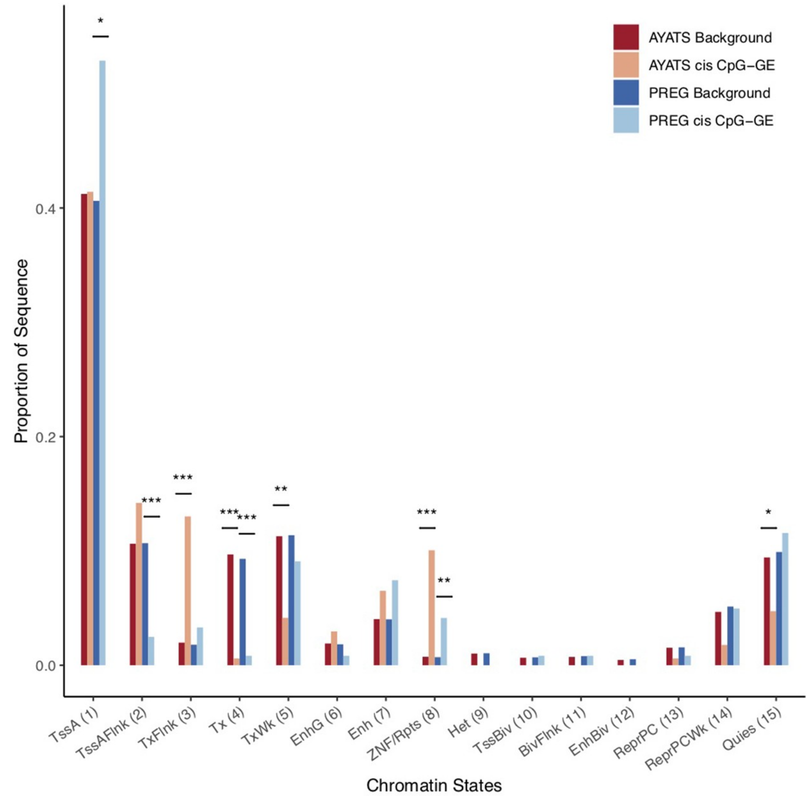

Figure 7. Enrichment for ENCODE chromatin states in cis CpG–transcript relationships. The 15-state ChromHMM model was used to determine regional chromatin states. Overall, GE-associated CpGs were depleted in transcriptionally active regions but enriched at zinc-finger binding sites (*** $P < 0.0005$; ** $P < 0.005$; * $P < 0.05$). Abbreviations: 1 = Active transcriptional start site (TSS), 2 = Flanking active TSS, 3 = Flanking strong transcription, 4 = Strong transcription, 5 = Weak transcription, 6 = Genic enhancer, 7 = Active enhancer, 8 = Zinc-finger genes & repeats, 9 = Heterochromatin, 10 = Bivalent/poised TSS, 11 = Flanking bivalent TSS, 12 = Bivalent Enhancers, 13 = Polycomb-repressed, 14 = Weak Repressed Polycomb, 15 = Quiescent.

on an associated gene's expression. Both negative and positive relationships between *cis* and *trans* DNAm–GE pairs were identified (Figure 1). Although DNAm is usually considered a repressive mark, inhibiting GE by either blocking transcription factor binding or by promoting a more condensed DNA conformation [17], positive DNAm–GE relationships could be explained by several mechanisms. Within genomic regulatory elements, transcription factors with repressive, rather than activating properties, may bind unmethylated sequences [25]. Furthermore, many transcription factors actually exhibit an increased affinity for heavily methylated sites [23]. Besides influencing the binding affinity of regulatory proteins, DNAm patterns may also reflect a history of transcription factor binding, a phenomenon that cannot be separately identified by a classic EWAS design [69]. In recent years, speculation has emerged regarding potential alternative roles of DNAm in the cell, including theories that DNAm may serve to direct splicing regulation or in maintaining genomic stability within specific regions [4,70–73]. Although this study found that the associations were predominantly negative across the majority of gene regions (Figure 1), these findings agree with other reports that strong positive DNAm–GE relationships exist [26,27,54]

Given the large effect size distribution of detected associations and the modest number of participants in each cohort, it was expected that only a small proportion of DNAm–GE connections would reach statistical significance. Instead of only focusing on individual connections, this study sought to outline genome-wide trends by identifying attributes of GE-associated CpG sites. Despite the modest number of DNAm–GE pairs overlapping across cohorts, GE-associated CpG sites displayed similar annotation characteristics (Tables 3 and 4). Annotations uniquely

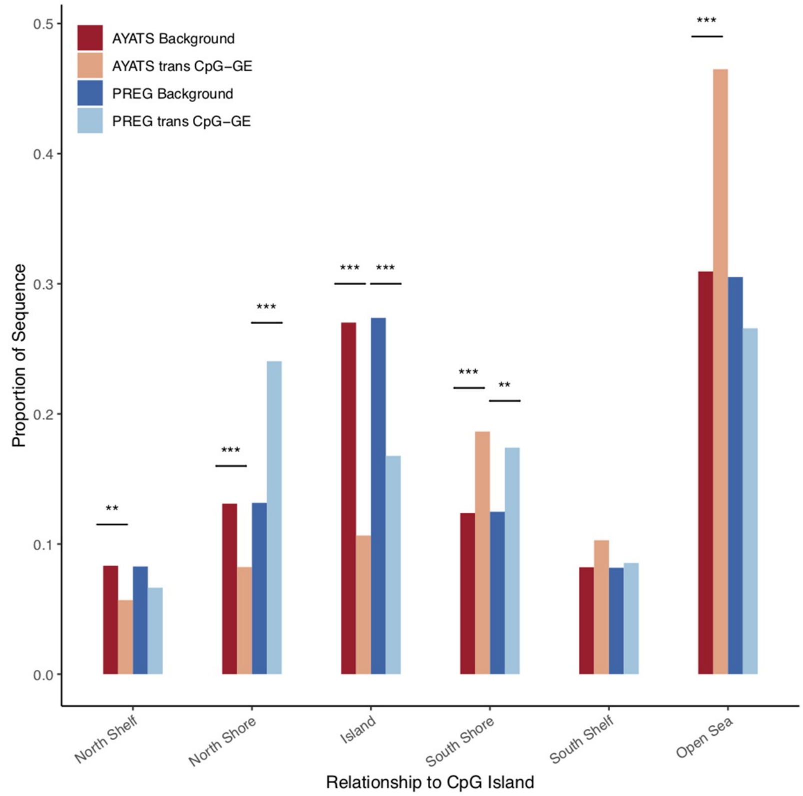

Figure 8. Enrichment for CpG classifiers in *trans* CpG–transcript relationships. CpG classifiers based on the distribution around CpG island regions were defined by the UCSC hg19 knownGene track. Islands were depleted while downstream regions bordering islands were significantly enriched in both cohorts. The North shore region directly upstream of CpG islands was more variable, with significant CpGs showing depletion in AYATS and enrichment in PREG (*** $P < 0.0005$; **$P < 0.005$; * $P < 0.05$).

characterized attributes of *cis* and *trans* GE-associated CpGs, indicating that separate paradigms may exist for proximal and distal connections. In general, DNAm within intermediate CpG density regions were more likely to be associated with GE. Regions of intermediate CG density are more variable compared with low- or high-density regions, and appear more dynamic across tissues and developmental stages [74,75]. Conversely, CpG sites in high-density regions, which are most often associated with CpG islands and promoter regions, were consistently depleted. Interestingly, both transcription factor binding and chromatin states in regions near active TSS were enriched, with those regions directly downstream of the TSS particularly characterized by a high proportion of GE-associated CpGs (Figure 3). Other studies have noted a similar relationship with DNAm located in the first intron, while also observing high transcription factor activity typical of intronic enhancers within these areas [65,76,77]. Although mechanisms of DNAm transcriptional inactivation usually focus on the hypermethylation of CpG sites within promoter and island regions, these results agree with other studies showing enrichment for “off-island” DNAm among GE-associated CpG sites [26,74]

The DNAm regulatory paradigm is predicated on the importance of DNAm in promoter regions, yet in this study CpG associations in promoters were depleted. Results from this study instead reiterated the significance of DNAm within enhancer regions and suggest that additional chromatin areas should be considered in addition to promoter-associated DNAm [26,29]. Multiple enhancer definitions (i.e., enhancer-like chromatin states and enhancer annotations generated from cap analysis of gene expression [CAGE]) were enriched within *trans* results. Enhancers have an established role in long-range gene regulation [8,78], often looping over more

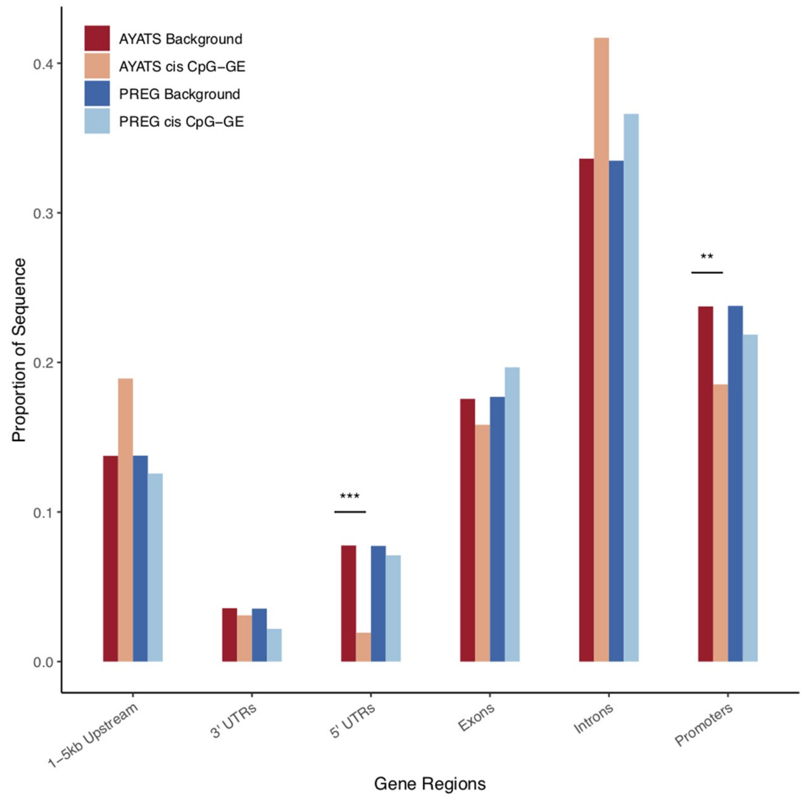

Figure 9. Enrichment for gene regions in *trans* CpG–transcript relationships. Gene regions were annotated based on the UCSC hg19 knownGene track. GE-associated CpG sites were depleted in 3′ untranslated regions (PREG), 5′ untranslated regions (AYATS), and in promoters (AYATS and PREG). Exons and introns were enriched in PREG and AYATS, respectively (***$P < 0.0005$; **$P < 0.005$; *$P < 0.05$).

proximal genes to interact with those farther away [79]. Enhancer regions are often characterized by intermediate DNAm and chromatin accessibility, demonstrate greater DNAm variability than promoters [57,74], and exhibit ongoing *de novo* methylation and demethylation activity [53]. The high rate of DNAm remodelling within enhancers, coupled with the strong DNAm–GE relationships found within these regions, align with hypotheses that suggest environmental exposures can influence complex disease risk through epigenetic mechanisms of transcriptional dysregulation. While mapping enhancers to their putative genes is a fundamental aim in identifying transcriptional regulatory networks, current methods are still under development [79], adding to the uncertainty in predicting the downstream functional effects of DNAm within these distal regulatory regions. Further challenges arise from evidence that many genes actually interact with multiple enhancers, and that these compounded interactions can result in additive effects on target GE [79].

This study serves to improve understanding of the relationships between DNAm and GE across the genome and cautions that, in the absence of concomitant GE measurements, EWAS should interpret DNAm–trait associations with care. Overall, these results identified that strong distal relationships between DNAm and GE are prevalent across the genome, highlighting issues with restricting DNAm-transcript annotations to small genomic intervals only [26,29]. The results from this study underscore concerns in predicting the biological mechanisms underlying disease from DNAm measurements alone and question the validity of assuming a *cis* DNAm–GE pathway without considering relevant features of the surrounding genomic landscape. EWAS relying on a DNAm-mediated transcriptional regulatory mechanism to interpret DNAm–trait associations may reach inaccurate conclusions about disease pathoetiology, as this approach fails to consider that the downstream effect of DNAm is likely

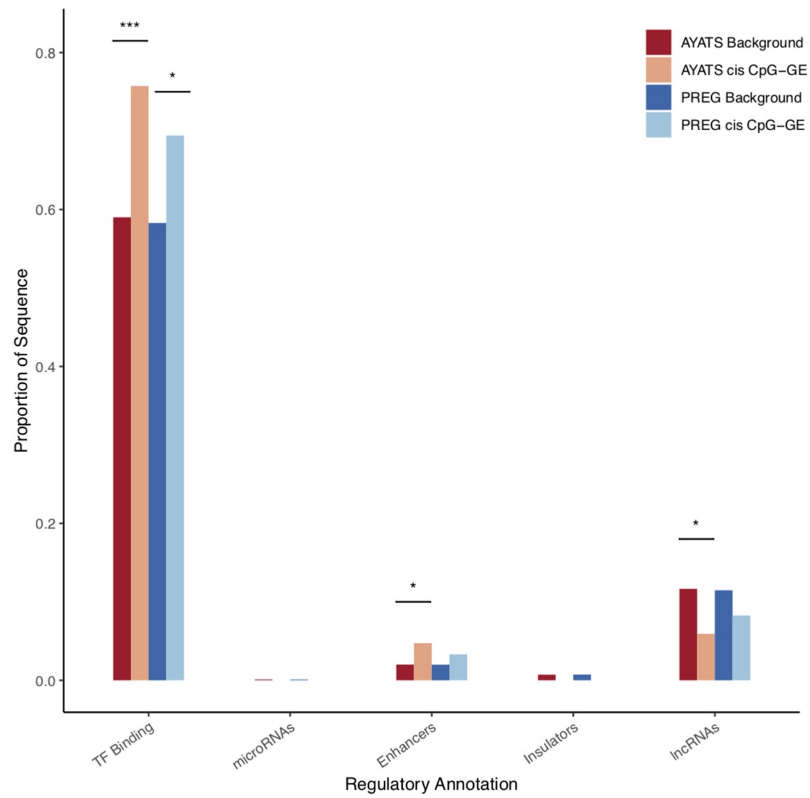

Figure 10. Enrichment for additional regulatory annotations in *trans* CpG–transcript relationships. Sites of transcription factor binding, as determined by ENCODE TF ChIP-seq, were significantly enriched in the PREG cohort. Enhancers were enriched across cohorts, and insulator regions were depleted in AYATS (*** $P < 0.0005$; **$P < 0.005$; *$P < 0.05$).

context-dependent and may be impacted by the attributes of specific CpG sites. While modified EWAS that incorporate GE information by performing a differential expression analysis (i.e., testing GE-disease associations) alongside testing for DNAm–disease associations avoid relying on assumptions of altered GE, biologically relevant information may be lost with this study design since *a priori* assumptions link CpG sites to putative genes and distal DNAm–GE relationships remain uninvestigated.

Based on these results, epigenetic research should continue moving towards multi-omic approaches that integrate DNAm with other levels of data (e.g., GE, genotypes, transcription factor binding) to study complex traits. Although DNAm-GE relationships are highly complex, the integration of DNAm with data outlining regional chromatin architecture and transcription factor activity may assist in predicting the functional impact of altered DNAm [26,80]. However, as an emerging and heterogeneous field, several obstacles can interfere with the implementation and interpretation of multi-omic studies [81–84]. Standardized analytical pipelines have yet to be developed, leading to difficulties in cross-study comparisons and in assessing rigour, and it may be some time before multi-omic studies generate replicable findings [83]. Even as some of these issues are mitigated, determining how to meaningfully but accurately decipher results from DNAm-only studies remains paramount. EWAS have generated a wealth of information outlining relationships between DNAm, diseases, and environmental exposures [19]. Researchers should reconsider how best to interpret these findings moving forward, given the emerging understanding of the complexity of epigenetic regulatory mechanisms. Currently, only a handful of studies have tested genome-wide associations between GE and DNAm [26,29,41,68,74,85], but variability in the methodologies used has led to difficulties in determining the replicability and generalizability of identified relationships. Although cross-study comparisons are challenging, several consistent themes have emerged

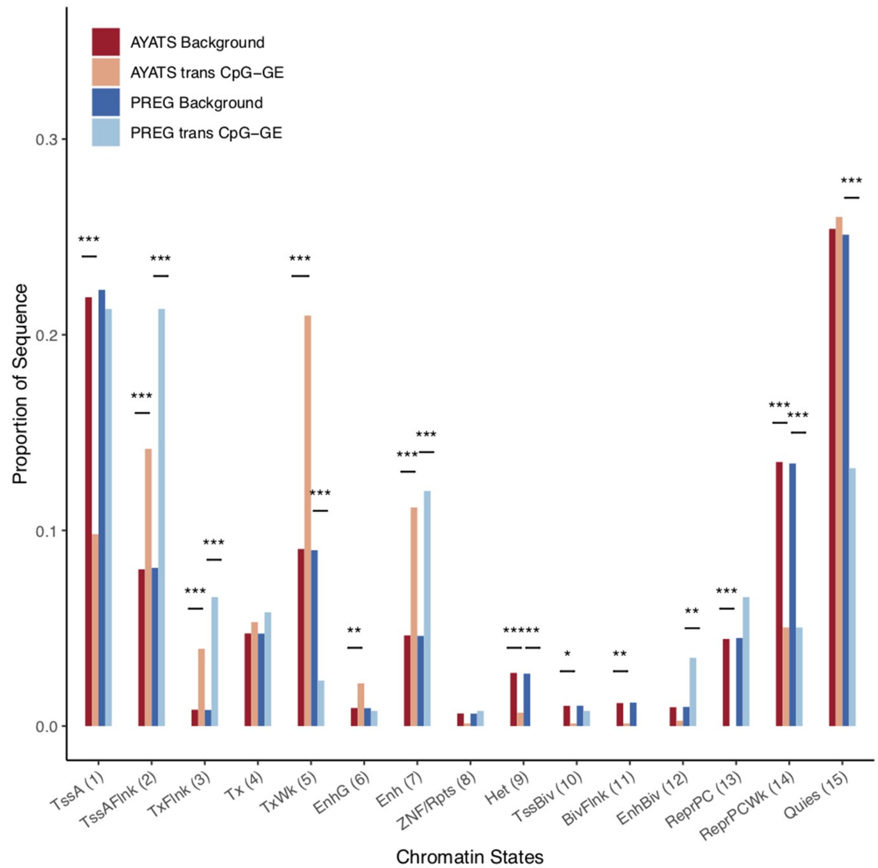

Figure 11. Enrichment for ENCODE chromatin states in *trans* CpG–transcript relationships. The 15-state ChromHMM model was used to determine regional chromatin states. Overall, GE-associated CpGs were depleted in repressive states but enriched at enhancers and areas flanking actively transcribed genes (***$P < 0.0005$; **$P < 0.005$; *$P < 0.05$). Abbreviations: 1 = Active transcriptional start site (TSS), 2 = Flanking active TSS, 3 = Flanking strong transcription, 4 = Strong transcription, 5 = Weak transcription, 6 = Genic enhancer, 7 = Active enhancer, 8 = Zinc-finger genes & repeats, 9 = Heterochromatin, 10 = Bivalent/poised TSS, 11 = Flanking bivalent TSS, 12 = Bivalent Enhancers, 13 = Polycomb-repressed, 14 = Weak Repressed Polycomb, 15 = Quiescent.

from this modest body of literature. This study replicates the overrepresentation of GE-associated CpGs within enhancers and at transcription factor binding sites, as well as the depletion within islands and promoter regions[26]. By further cataloguing specific genome-wide DNAm–GE associations and identifying attributes of GE-associated CpG sites, results from this study can be used to inform EWAS design and the interpretation of DNAm–trait associations. Moving forward, continued examination of DNAm–GE relationships in large, diverse cohorts should be prioritized to advance our understanding of the role of DNAm within the cell and disease biology.

Strengths and Limitations

To our knowledge, this was the first study to assess the global relationship between peripheral blood DNAm and GE in both a primary and replication sample. However, results of this study should be considered in the context of the following limitations. First, both DNAm and GE were measured by microarray technologies that provided coverage within well-characterized locations, but were unable to assay the full extent of RNA and CpG sites in the genome [86]. GE was measured on a gene-level, rather than transcript-level, array platform, so that specific transcript variants were not analysed individually. Future analyses with more comprehensive measurements, particularly those that utilize technology with transcript-level resolution (e.g., RNA-seq), are necessary for confirming genome-wide trends. Second, only relationships of relatively large effect size were detected in this study (adjusted R-squared range = 0.23–0.90). Especially given that a conservative multiple testing correction was applied, it is assumed that many more DNAm–GE connections exist but were undetected in this study, which could influence the results of the feature enrichment tests. Third, both cohorts were analysed cross-sectionally, a study design that is unable to provide evidence for causation or directionality [14,16]. Mechanisms of reverse causation, in which changes to DNAm occur in response to modified GE, have been observed [87]. Therefore, it is unknown whether changes to DNAm are actually proceeding changes in GE as described in the canonical mechanism. Fourth, some annotations were derived from experiments conducted on a well-described lymphoblastoid cell line (GM12787), which was selected based on data that supports the genetic and functional similarity to mature blood cells (i.e., T cells and B cells) [88]. One benefit of using this approach is that annotations were kept consistent across the different functional enrichment categories (e.g., chromatin landscapes, enhancer definitions, etc.). It remains important to consider that this study focused on the association between DNAm at individual CpG sites and GE. In actuality, regional changes in DNAm may be co-regulating GE in some instances, and future studies can examine how regions of CpGs work in concert to regulate GE [68,89]. Finally, this study only investigated DNAm and GE in the peripheral blood and may not generalize to other tissues [65]. Future analyses with more comprehensive measurements in alternative tissues will be crucial for characterizing genome-wide trends across cell types.

Abbreviations

AYATS	Adolescent and Young Adult Twin Study
DMPs	Differentially methylated positions
DNAm	DNA methylation
EWAS	Epigenome-wide association study
FDR	False discovery rate
GE	Gene expression
GO	Gene ontology
KEGG	Kyoto Encyclopaedia of Genes and Genomes
PREG	Pregnancy, Race, Environment, Genes study
TSS	Transcriptional start site
UTRs	Untranslated regions

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Center for Advancing Translational Sciences [UL1TR000058]; National Institute of Mental Health [MH106924]; National Institute of Mental Health [MH101518]; National Institute on Minority Health and Health Disparities [P60MD002256]; Brain and Behavior Research Foundation [21976].

Author contributions

All authors assisted with the design of the study. EL cleaned the data, performed the analyses, and wrote the initial draft of the manuscript. RRN, VV, BR, JL and TY provided substantive feedback and revisions, and TY and RRN planned and secured funding for the PREG and AYATS studies.

Data availabilty

Sharing PREG and AYATS study data is limited by Institutional Review Board agreements and participant consent forms, which restrict openly sharing individual-level measures. Anyone interested in data access or collaboration is encouraged to contact Dr. Timothy P. York (timothy.york@vcuhealth.org) or Dr. Roxann Roberson-Nay (roxann.robersonnay@vcuhealth.org) for more information.

Informed consent and ethical approvals

Both the AYATS and PREG study received Virginia Commonwealth University Institutional Review Board approval and obtained written participant consent for each participant.

ORCID

Eva E. Lancaster http://orcid.org/0000-0002-7371-8094
Brien P. Riley http://orcid.org/0000-0002-2408-8268
Roxann Roberson-Nay http://orcid.org/0000-0002-1037-3121
Timothy P. York http://orcid.org/0000-0003-4068-4286

References

[1] Kaur G, Begum R, Thota S, et al. A systematic review of smoking-related epigenetic alterations. Arch Toxicol. 2019 Oct;93(10):2715–2740.

[2] Anderson OS, Sant KE, Dolinoy DC. Nutrition and epigenetics: an interplay of dietary methyl donors,

one-carbon metabolism and DNA methylation. J Nutr Biochem. 2012 Aug;23(23):853–859.
[3] Vinkers CH, Kalafateli AL, Rutten BPF, et al. Traumatic stress and human DNA methylation: a critical review. Epigenomics. 2015;7(4):593–608.
[4] Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. 2019 Oct;20(10):590–607.
[5] York TP, Eaves LJ, Neale MC, et al. The contribution of genetic and environmental factors to the duration of pregnancy. 3rd. Am J Obstet Gynecol. 2014 May;210(5):398–405.
[6] Roberson-Nay R, Lapato DM, Wolen AR, et al. An epigenome-wide association study of early-onset major depression in monozygotic twins. Transl Psychiatry. 2020 Aug;10(10):301.
[7] De Bustos C, Ramos E, Young JM, et al. Tissue-specific variation in DNA methylation levels along human chromosome 1. Epigenetics Chromatin. 2009 June;2(1):7.
[8] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13(7):484–492.
[9] van der Maarel SM. Epigenetic mechanisms in health and disease. Ann Rheum Dis. 2008 Dec;67(3):iii97–100.
[10] Zhang X, Gierman HJ, Levy D, et al. Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs. BMC Genomics. 2014 June;15(15):532.
[11] Hannon E, Knox O, Sugden K, et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. PLoS Genet. 2018 Aug;14(8):e1007544.
[12] Teh AL, Pan H, Chen L, et al. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. Genome Res. 2014 July;24(7):1064–1074.
[13] van Dongen J, Ehli EA, Slieker RC, et al. Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells. Genes (Basel). 2014 May;5(2):347–365.
[14] Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nat Rev Genet. 2017 July;18(7):441–451.
[15] Gibney ER, Nolan CM. Epigenetics and gene expression. Heredity (Edinb). 2010 July;105(1):4–13.
[16] Birney E, Smith GD, Greally JM. Epigenome-wide association studies and the interpretation of disease -omics. PLoS Genet. 2016 June;12(6):e1006105.
[17] Schübeler D. Function and information content of DNA methylation. Nature. 2015 Jan;517(7534):321–326.
[18] Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. Nat Rev Genet. 2018 Mar;19(3):129–147.
[19] Michels KB, Binder AM, Dedeurwaerder S, et al. Recommendations for the design and analysis of epigenome-wide association studies. Nat Methods. 2013 Oct;10(10):949–955.
[20] Zhong H, Kim S, Zhi D, et al. Predicting gene expression using DNA methylation in three human populations. PeerJ. 2019 May;7:e6757.
[21] Du J, Johnson LM, Jacobsen SE, et al. DNA methylation pathways and their crosstalk with histone methylation. Nat Rev Mol Cell Biol. 2015 September;16:519–532.
[22] Liu H, Chen Y, Lv J, et al. Quantitative epigenetic co-variation in CpG islands and co-regulation of developmental genes. Sci Rep. 2013;3(1):2576.
[23] Yin Y, Morgunova E, Jolma A, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356(6337). 10.1126/science.aaj2239
[24] Doane AS, Elemento O. Regulatory elements in molecular networks. Wiley Interdiscip Rev Syst Biol Med. 2017 May;9(3). 10.1002/wsbm.1374.
[25] Smith J, Sen S, Weeks RJ, et al. Promoter DNA hypermethylation and paradoxical gene activation. *Trends Cancer Res*. 2020 May;6(5): 392–406.
[26] Kennedy EM, Goehring GN, Nichols MH, et al. An integrated -omics analysis of the epigenetic landscape of gene expression in human blood cells. BMC Genomics. 2018 June;19(19):476.
[27] Wagner JR, Busche S, Ge B, et al. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014 Feb;15(2):R37.
[28] Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009 Nov;462(7271):315–322.
[29] Aran D, Hellman A. DNA methylation of transcriptional enhancers and cancer predisposition. Cell. 2013 July;154(1):11–13.
[30] Varley KE, Gertz J, Bowling KM, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013 Mar;23(3):555–567.
[31] Huang Y-T. Integrative modeling of multi-platform genomic data under the framework of mediation analysis. Stat Med. 2015 Jan;34(1):162–178.
[32] Mamrut S, Avidan N, Truffault F, et al. Methylome and transcriptome profiling in myasthenia gravis monozygotic twins. J Autoimmun. 2017 Aug;82:62–73.
[33] Gillberg L, Perfilyev A, Brøns C, et al. Adipose tissue transcriptomics and epigenomics in low birthweight men and controls: role of high-fat overfeeding. Diabetologia. 2016 Apr;59(4):799–812.
[34] Tian W, Li Y, Zhang J, et al. Combined analysis of DNA methylation and gene expression profiles of osteosarcoma identified several prognosis signatures. Gene. 2018 Apr;650:7–14.
[35] Song D, Qi W, Lv M, et al. Combined bioinformatics analysis reveals gene expression and DNA methylation patterns in osteoarthritis. Mol Med Rep. 2018 June;17(6):8069–8078.

[36] Miao L, Yin R-X, Zhang Q-H, et al. Integrated DNA methylation and gene expression analysis in the pathogenesis of coronary artery disease. Aging (Albany NY). 2019 Mar;11(5):1486–1500.
[37] Wang Z, Wu X, Wang Y. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. BMC Bioinformatics. 2018 Apr;19(S5):115.
[38] Zhang Y, Fang L, Zang Y, et al. Identification of core genes and key pathways via integrated analysis of gene expression and DNA methylation profiles in bladder cancer. Med Sci Monit. 2018 May;24:3024–3033.
[39] Li H, Wang F-L, Shan L-P, et al. Comprehensive analysis of gene expression and DNA methylation for human nasopharyngeal carcinoma. *Eur Arch Otorhinolaryngol*. 2019 Sept;276(9): 2565–2576.
[40] Bell JT, Pai AA, Pickrell JK, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011 Jan;12(1):R10.
[41] van Eijk KR, de Jong S, Boks MPM, et al. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics. 2012 Nov;13(13):636.
[42] Cecilione JL, Rappaport LM, Hahn SE, et al. Genetic and environmental contributions of negative valence systems to internalizing pathways. Twin Res Hum Genet. 2018 Feb;21(21):12–23.
[43] Lapato DM, Moyer S, Olivares E, et al. Prospective longitudinal study of the pregnancy DNA methylome: the US Pregnancy, Race, Environment, Genes (PREG) study. BMJ Open. 2018 May;8(5):e019721.
[44] R Core Team. “R: a language and environment for statistical computing. Vienna Austria”: R Foundation for Statistical Computing; 2016.
[45] Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012 May;13(1):86.
[46] Maes M, Meltzer HY, Stevens W, et al. Natural killer cell activity in major depression: relation to circulating natural killer cells, cellular indices of the immune response, and depressive phenomenology. Prog Neuropsychopharmacol Biol Psychiatry. 1994 July;18 (18):717–730.
[47] Barfield RT, Almli LM, Kilaru V, et al. Accounting for population stratification in DNA methylation studies. *Genet Epidemiol*. 2014 Apr;38(3): 231–241.
[48] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43 (7):e47.
[49] Guintivano J, Shabalin AA, Chan RF, et al. Test-statistic inflation in methylome-wide association studies. Epigenetics. 2020 May;15(11):1163–1166.
[50] Goñi JR, Pérez A, Torrents D, et al. Determining promoter location based on DNA structure first-principles calculations. Genome Biol. 2007;8(12):R263.
[51] Brandeis M, Frank D, Keshet I, et al. Sp1 elements protect a CpG island from de novo methylation. Nature. 1994 Sept;371(6496):435–438.
[52] Mummaneni P, Yates P, Simpson J, et al. The primary function of a redundant sp1 binding site in the mouse aprt gene promoter is to block epigenetic gene inactivation. Nucleic Acids Res. 1998 Nov;26 (22):5163–5169.
[53] Feldmann A, Ivanek R, Murr R, et al. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. 2013 Dec;9(12):e1003994.
[54] Xie -F-F, Deng F-Y, Wu L-F, et al. Multiple correlation analyses revealed complex relationship between DNA methylation and mRNA expression in human peripheral blood mononuclear cells. Funct Integr Genomics. 2018 Jan;18(18):1–10.
[55] Zhao Y, Sun H, Wang H. Long noncoding RNAs in DNA methylation: new players stepping into the old game. Cell Biosci. 2016 July;6(1):45.
[56] Yu L, Xia K, Cen X, et al. DNA methylation of noncoding RNAs: new insights into osteogenesis and common bone diseases. Stem Cell Res Ther. 2020 Mar;11 (11):109.
[57] Roadmap Epigenomics Consortium AK, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb;(518):317–330. https://doi.org/10.1038/nature14248
[58] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012 Feb;9(9):215–216.
[59] Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol. 1987 July;196(2):261–282.
[60] Price ME, Cotton AM, Lam LL, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethy- lation450 BeadChip array. Epigenetics Chromatin. 2013 Mar;6(6):4.
[61] Cavalcante RG, Sartor MA. annotatr: genomic regions in context. Bioinformatics. 2017 Aug;33 (15):2381–2383.
[62] Project Consortium ENCODE. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sept;489:57–74.
[63] Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 2012 Sept;22:1760–1774.
[64] Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014 Mar;507(7493):455–461.
[65] Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the

first intron and gene expression across tissues and species. Epigenetics Chromatin. 2018 June;11(11):37.
[66] Carlson M, Maintainer BP. *TxDb.Hsapiens.UCSC.hg19.knownGene: annotation package for TxDb object(s)*, 2015. R package version 3.2.2.
[67] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289–300.
[68] Cheng J, Wei D, Ji Y, et al. Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. Genome Med. 2018 May;10(1):42.
[69] Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. Nat Rev Genet. 2016 Aug;17(9):551–565.
[70] Maunakea AK, Nagarajan RP, Bilenky M, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010 July;466(7303):253–257.
[71] Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. Trends Genet. 2015 May;31(5):274–280.
[72] Irizarry RA, Ladd-Acosta C, Wen B, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009 Feb;41(2):178–186.
[73] Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007 Apr;8(4):272–285.
[74] Liu Y, Ding J, Reynolds LM, et al. Methylomics of gene expression in human monocytes. Hum Mol Genet. 2013 Dec;22(24):5065–5074.
[75] Doi A, Park I-H, Wen B, et al. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat Genet. 2009 Dec;41(12):1350–1353.
[76] Vanderkraats ND, Hiken JF, Decker KF, et al. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. Nucleic Acids Res. 2013 Aug;41(14):6816–6827.
[77] Schlosberg CE, VanderKraats ND, Edwards JR. Modeling complex patterns of differential DNA methylation that associate with gene expression changes. Nucleic Acids Res. 2017 May;45(9):5100–5111.
[78] Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011 Feb;144(3):327–339.
[79] Schoenfelder S, Fraser P. Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet. 2019 Aug;20(8):437–455.
[80] Chen BH, Marioni RE, Colicino E, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany NY). 2016 Sept;8(9):1844–1865.
[81] Pinu FR, Beale DJ, Paten AM, et al. Systems biology and multi-omics integration: viewpoints from the metabolomics research community. Metabolites. 2019 Apr;9(4):76.
[82] Haas R, Zelezniak A, Iacovacci J, et al. Designing and interpreting 'multi- omic' experiments that may change our understanding of biology. Curr Opin Syst Biol. 2017 Dec;6:37–45.
[83] Subramanian I, Verma S, Kumar S, et al. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020 Jan;14(14):1177932219899051.
[84] Bersanelli M, Mosca E, Remondini D, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016 Jan;17(2):15.
[85] Taylor DL, Jackson AU, Narisu N, et al. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. Proc Natl Acad Sci U S A. 2019 May;116(22):10883–10888.
[86] Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011 Oct;98(4):288–295.
[87] Pacis A, Mailhot-Léonard F, Tailleux L, et al. Gene activation precedes DNA demethylation in response to infection in human dendritic cells. Proc Natl Acad Sci U S A. 2019 Apr;116(14):6938–6943.
[88] Sie L, Loong S, Tan EK. Utility of lymphoblastoid cell lines. J Neurosci Res. 2009 July;87(9):1953–1959.
[89] Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics Chromatin. 2016 June;9(1):26.