Human Molecular Genetics, 2013, Vol. 22, No. 24 **5065–5074**
doi:10.1093/hmg/ddt356
Advance Access published on July 29, 2013

Methylomics of gene expression in human monocytes

Yongmei Liu[1,*], Jingzhong Ding[1], Lindsay M. Reynolds[1], Kurt Lohman[1], Thomas C. Register[1], Alberto De La Fuente[2], Timothy D. Howard[1], Greg A. Hawkins[1], Wei Cui[1], Jessica Morris[1], Shelly G. Smith[1], R. Graham Barr[3], Joel D. Kaufman[4], Gregory L. Burke[1], Wendy Post[5], Steven Shea[3], Charles E. Mccall[1], David Siscovick[4], David R. Jacobs, Jr[6], Russell P. Tracy[7], David M. Herrington[1,†] and Ina Hoeschele[8,†]

[1]Wake Forest School of Medicine, Winston-Salem, NC, USA, [2]Leibniz Institute for Farm Animal Biology, Mecklenburg-Vorpommern, Germany, [3]Columbia University Medical Center, New York, NY, USA, [4]University of Washington, Seattle, WA, USA, [5]John Hopkins University, Baltimore, MD, USA, [6]University of Minnesota, Minneapolis, MN, USA, [7]University of Vermont, Colchester, VT, USA and [8]Virginia Tech, Blacksburg, VA, USA

Received December 19, 2012; Revised and Accepted July 23, 2013

DNA methylation is one of several epigenetic mechanisms that contribute to the regulation of gene expression; however, the extent to which methylation of CpG dinucleotides correlates with gene expression at the genome-wide level is still largely unknown. Using purified primary monocytes from subjects in a large community-based cohort ($n = 1264$), we characterized methylation (>485 000 CpG sites) and mRNA expression (>48K transcripts) and carried out genome-wide association analyses of 8370 expression phenotypes. We identified 11 203 potential *cis*-acting CpG loci whose degree of methylation was associated with gene expression (eMS) at a false discovery rate threshold of 0.001. Most of the associations were consistent in effect size and direction of effect across sex and three ethnicities. Contrary to expectation, these eMS were not predominately enriched in promoter regions, or CpG islands, but rather in the 3′ UTR, gene bodies, CpG shores or 'offshore' sites, and both positive and negative correlations between methylation and expression were observed across all locations. eMS were enriched for regions predicted to be regulatory by ENCODE (Encyclopedia of DNA Elements) data in multiple cell types, particularly enhancers. One of the strongest association signals detected ($P < 2.2 \times 10^{-308}$) was a methylation probe (cg17005068) in the promoter/enhancer region of the glutathione *S*-transferase theta 1 gene (*GSTT1*, encoding the detoxification enzyme) with *GSTT1* mRNA expression. Our study provides a detailed description of the epigenetic architecture in human monocytes and its relationship to gene expression. These data may help prioritize interrogation of biologically relevant methylation loci and provide new insights into the epigenetic basis of human health and diseases.

INTRODUCTION

Methylation of CpG dinucleotides is an important contributor to epigenetic regulation of gene expression in numerous cellular processes, including genomic imprinting and X-chromosome inactivation (1–5). However, the relationship between methylation of CpG dinucleotides and gene expression at the genome-wide levels is still not well understood. Previous studies of the relationship between DNA methylation and gene expression have been limited to discrete genomic regions (6,7), modest samples sizes or mixed cell types (8–16). For example, a recent study of the association of DNA methylation with gene expression in peripheral blood mononuclear cells of 55 subjects using Illumina HumanMethylation27 and Human Ref-8 v3.0 Expression BeadChips reported that only 97 of the 16 419 CpG sites tested (0.6%) had significant correlations

*To whom correspondence should be addressed at: Division of Public Health Sciences, Department of Epidemiology and Prevention, Wake Forest School of Medicine, Winston-Salem NC 27157, USA. Email: yoliu@wakehealth.edu
†These authors contributed equally to this work.

[false discovery rate (FDR) < 5%] between DNA methylation and gene expression (14). However, much larger sample sizes are required for genome-wide studies of CpG methylation and gene expression to correct for genome-scale multiple testing, and homogeneous or nearly homogeneous cell samples may be necessary to avoid false positives or false negatives that could arise from admixture of cell types.

Here, we describe the association between methylation of *cis*-acting CpG loci and 8370 expression phenotypes in purified monocytes from a subset of participants ($n = 1264$) in the Multi-Ethnic Study of Atherosclerosis (MESA), a large cohort of community-dwelling subjects, aged 55–94 years from four communities in the USA. For this study, we chose monocytes, in part, because they are key cells of innate immunity and major contributors to the pathogenesis of atherosclerosis. Our goals were to describe the distribution of *cis*-acting methylation/expression correlations with respect to other structural features of the genome [e.g. transcription start sites (TSSs), regulatory regions, etc.] and to identify strong locus-specific correlations between methylation and gene expression as high-value targets for future functional evaluation.

RESULTS

General features of the monocyte methylome

We simultaneously characterized the methylomic (>485K CpG sites) and transcriptomic (>48K transcripts) profiles in purified monocytes from 1264 Caucasian (47%), African-American (21%) and Hispanic (32%) MESA participants (51% female). We excluded methylation loci and transcript expression probes that failed quality control (QC) criteria (see Materials and Methods) as well as probes that were rarely expressed or on the X or Y chromosome. When multiple transcripts were available for a given gene, we selected the single transcript with the strongest association with any of the candidate *cis*-methylation loci. A total of 416 507 CpG sites and 8370 expressed genes remained for analysis. Across the CpG sites, the distribution of median %methylation was strongly bimodal (Supplementary Material, Fig. S1A). In contrast, the variation of %methylation within sites was generally quite uniform [median inter-decile range (IDR): 4.2% methylation], with a relatively small portion of sites demonstrating more substantial inter-individual variation (Supplementary Material, Fig. S1B). The CpG sites tended to be hypo-methylated in promoter regions, 5′ UTRs, first exons and CpG islands (regions of high CpG density), and hyper-methylated in gene bodies, 3′ UTRs, intergenic regions and 'offshore' sites (sites outside of annotated CpG islands; Supplementary Material, Fig. S2A and B).

In a previous study of CpG sites in promoter regions (15), methylation levels at probes located in close proximity to one another (up to 2 kb) were highly correlated. We investigated the pair-wise correlations of methylation levels (across individuals) in relationship to the distance between CpG sites. The pair-wise correlations were generally low across the genome, including the promoter regions (median: 0.07) where the average of pair-wise correlations of methylation was weak even for loci ≤60 base pairs apart (median: 0.09; Supplementary Material, Fig. S3).

Identification of potential *cis*-acting regulatory methylation sites

We defined potential *cis*-acting regulatory methylation sites (eMS) as CpG sites whose %methylation was associated (at an FDR threshold of 0.001) with expression of any autosomal gene within 1 Mb of the CpG site in question. These associations were estimated using linear regression, with transcript expression as the outcome variable and methylation level at the corresponding CpG site as the explanatory variable after adjustment for age, sex, race/genetically inferred ethnicity, study site and residual sample contamination with non-monocytes. We identified 11 203 eMS (1.2% of CpG sites examined) associated with 3093 genes (37.0% of 8370 expressed genes examined) at a FDR threshold of 0.001 (P-values ranged from 5.0×10^{-6} to $<2.2 \times 10^{-308}$) in the overall analysis. In analyses stratified by sex, we identified 4599 and 4224 eMS (FDR < 0.001) for women and men, respectively, with 3058 (53.0%) significant in both sex groups (Supplementary Material, Fig. S4A). We also identified 4129, 1099 and 2168 eMS for Caucasians, African Americans and Hispanics, respectively (Supplementary Material, Fig. S4B). In total, 1921 eMS (40.4%) were significant in at least two ethnic groups, and 729 eMS (15.4%) were significant in all three ethnic groups. However, a combined analysis including methylation by sex or ethnicity interaction did not identify any sex-specific eMS, and only identified 52 African-American-specific eMS (FDR < 0.001). Most of the 11 203 eMS associations were consistent in the size and direction of effect across sex and ethnic subgroups (Supplementary Material, Fig. S5). The lower consensus of eMS across ethnicities is likely in part due to the lower statistical power in African Americans ($n = 272$) and Hispanics ($n = 402$).

Compared with all CpG sites, bimodality of median %methylation was attenuated in the 11 203 eMS (Supplementary Material, Fig. S1A), and the inter-individual variation as evaluated by the IDR was higher among the eMS (median IDR: 11.5% methylation, Supplementary Material, Fig. S1B). We note that among sites with low IDR, there remained instances of significant correlation between methylation and transcript expression. As expected, the eMS with lower IDR tend to have larger effects (Supplementary Material, Fig. S6). This finding argues against the frequently adopted approach of discarding methylation probes with low variance.

Among the 3093 genes with at least one identified eMS, 67% had two or more eMS. When methylation of more than one eMS was associated with expression of the same gene, we used a stepwise selection procedure to identify those eMS that were independently associated with the transcript of interest. The number of 'independent' eMS detected was 7029 (compared with 11 203 total eMS), and the median number per associated transcript was 2 (range: 1–28). The eMS jointly accounted for a median of 2.9% of the variation in expression of their associated gene transcripts (R^2; range: 0.3–84.4%); R^2 was >25% for 163 transcripts (Supplementary Material, Fig. S7).

Distribution of eMS with respect to annotated features of the genome

When considering the *cis*-acting eMS as a function of proximity to TSSs, we found that the magnitude of the correlations (in

either the positive or negative direction) was inversely related to distance from the TSS (Fig. 1). Nearly 45% of the *cis*-acting eMS were within 40 kb of the TSS of their cognate gene; however, evidence of association was also seen over distances up to 1 Mb.

When considering the location of eMS within genes, the number of eMS was highest in gene bodies, followed by promoter and intergenic regions. However, relative to the number of CpG sites assayed in each gene structural region, eMS were significantly enriched in the 3′ UTR (1.33-fold enrichment; $P = 1.5 \times 10^{-10}$) and gene bodies (1.14-fold enrichment; $P = 5.4 \times 10^{-25}$), whereas intergenic regions and the first exon contained relatively fewer eMS (Fig. 2A). There was a slight predominance of negative correlations between eMS methylation and transcript expression across all gene structural regions (Fig. 2B).

Similarly, the number of eMS was highest in 'offshore' sites, followed by CpG 'shores' (regions of low CpG density within ~2 kb of CpG islands) and CpG islands. However, relative to the number of sites assayed in each CpG annotation class, CpG 'shores' and 'offshore' sites were highly enriched with eMS, whereas CpG islands contained relatively fewer eMS (Fig. 3A). There were more negative correlations between eMS methylation and transcript expression across all regions related to CpG islands (Fig. 3B).

We also tested whether eMS were enriched for experimentally determined DNase I-hypersensitive sites (DHSs) or transcription factor-binding sites (TFBS), as well as bioinformatically determined functional sites defined as enhancers or insulators, using data from the Encyclopedia of DNA Elements (ENCODE) project (17) (Table 1). The enrichment test for DHSs in monocytes ($P = 2.5 \times 10^{-7}$) and DHSs in any cell type ($P < 2.2 \times 10^{-308}$) was significant with the same relative enrichment (1.18-fold enrichment). The enrichment tests for enhancers, insulators or TFBS determined in any cell type were also significant, with the strongest enrichment detected in enhancer regions (1.49-fold enrichment, $P < 2.2 \times 10^{-308}$). Supplementary Material, Figure S8, shows the distribution of methylation-expression correlations for eMS that reside in DHSs, enhancers, insulators and TFBS. We observed more negative correlations between eMS methylation and transcript expression across all the predicted functional sites. In addition, we tested the enrichment of eMS in ENCODE-annotated functional sites by specific genomic regions (Supplementary Material, Table S1). In 3′ UTR and gene bodies, we found significant enrichment of eMS for DHSs, enhancers, insulators and TFBS (1.20–1.73-fold enrichment) relative to the number of CpG sites assayed in each of the two regions. eMS were enriched for enhancers in all gene structural regions (1.43–2.13-fold enrichment) (Table S1).

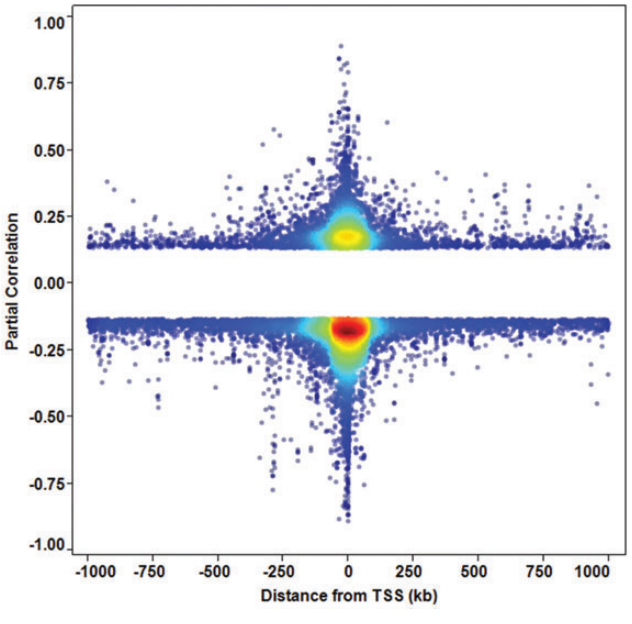

Figure 1. Scatter plot of partial expression–methylation correlations versus distance to the TSS for the 11 203 eMS. Most eMS were located close to TSSs and associations between transcript expression and eMS methylation were symmetric on both sides of TSSs (corrected for strand). The color scale is a measure of the density of points in the region.

Identification of individual genes and over-representation of biological processes among the most highly correlated *cis*-acting eMS/transcript pairs

The Manhattan plot (Supplementary Material, Fig. S9) of the *P*-values versus genomic locations for the 416 507 CpG sites revealed several dense regions of highly significant eMS on chromosomes 1, 6, 7, 12, 19 and 22. Gene set enrichment analysis suggested that the 11 203 eMS were most often associated with genes involved in immune response and regulation, protein transport and regulation of programmed cell death (Table 2). Supplementary Material, Table S2, shows the 84 most highly correlated *cis*-acting eMS/transcript pairs (FDR $\leq 1 \times 10^{-100}$). The entire list of *cis*-acting eMS/transcript pairs can be accessed through the eMS database at the MESA Epigenomics site (http://www.wakehealth.edu/mesaepigenomics).

One of the most significant findings ($P < 2.2 \times 10^{-308}$) was the association of GSTT1 mRNA expression with a methylation CpG site (cg17005068) located in the promoter/enhancer region of *GSTT1* on chromosome 22 (Fig. 4). GSTT1 is a detoxification enzyme that plays a significant role in the reduction of environmental pollutants, mutagens, carcinogens and anticancer drugs (2). There were 15 additional eMS associated with *GSTT1* expression (Fig. 4B). These 16 eMS are located in close proximity to CpG islands that overlap with DNaseI hypersensitivity clusters detected in monocytes, TFBS (ChIP-seq) in many cell types, and predicted promoter and enhancer regions in B cells (GM12878). Collectively, these eMS explained 77% of the variation in GSTT1 mRNA expression when jointly included in a multiple linear regression model; nine eMS remained significant ($P < 0.05$) in the joint model, with CpG site cg17005068 attaining the strongest independent signal ($P = 1.7 \times 10^{-27}$). When correlating GSTT1 mRNA expression and methylation of CpG site cg17005068 with all the single-nucleotide polymorphisms (SNPs) located within 1 Mb from the gene's TSS using single SNP regression, we identified a *cis*-acting SNP, rs407257, that was most strongly associated with both methylation and *GSTT1* expression in each of the three ethnicities (Fig. 4C). We performed causal inference using Mendelian randomization (18) by structural equation modeling (SEM) to compare the fit of six potential causal models in 590 Caucasians (Supplementary Material, Fig. S10).

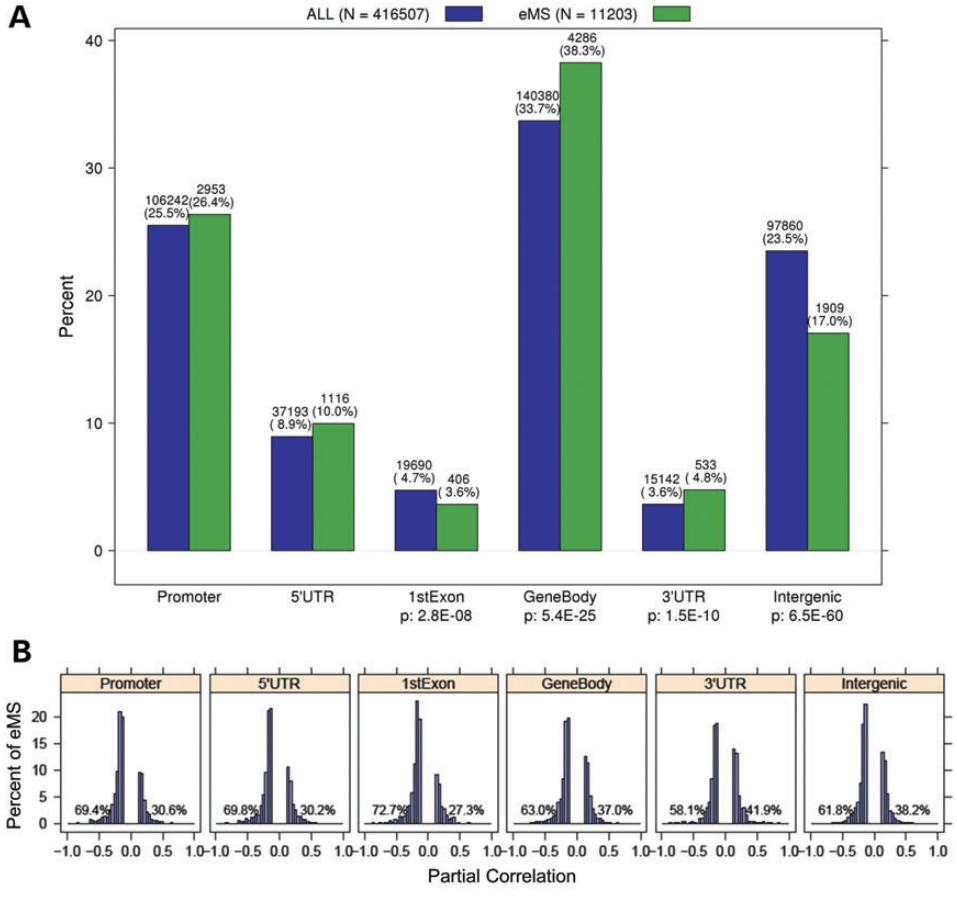

Figure 2. Distribution of 416K CpG sites on the Illumina array and 11 203 eMS across the different gene structural regions, and partial expression–methylation correlations of the 11 203 eMS for gene structural regions. (**A**) The absolute number of eMS was highest in gene bodies (end of first exon to end of last exon), followed by promoter regions (<1.5 kb upstream of the TSS), and intergenic regions (upstream of the promoter or downstream of the 3′ UTR, but still within 1 Mb of the TSS). Relative to the number of sites assayed in each gene structural region, the gene body and 3′ UTR regions were significantly enriched with eMS compared with other regions, whereas intergenic regions and 1st exons contained relatively fewer eMS ($P = 2.8 \times 10^{-8}$–6.5×10^{-60}, chi-square test). (**B**) There was a predominance of negative correlations between degree of eMS methylation and expression of their associated gene transcripts across all gene structural regions; however, this imbalance was most evident in the promoter regions, 5′ UTR and first exons.

Overall, the best-fitting model (Model 1) postulates a direct effect of the SNP on both methylation and gene expression, and a causal effect of methylation on gene expression (Supplementary Material, Table S3 and S4). Although the causal analysis favors a causal effect of methylation on expression with Model 1 providing an excellent fit to the data and Model 3 (with the reverse causality from expression to methylation) fitting the data poorly, we cannot rule out the possibility that the correlation between methylation and gene expression resulted from an uncontrolled (hidden) confounding effect (Model 5).

DISCUSSION

This large-scale epigenome-wide study identified and characterized a large number of CpG methylation sites associated with gene expression (eMS) in purified human monocytes—an important next step toward a comprehensive understanding of the methylome's functional topology. We show that eMS are located broadly across genome and enriched for regions predicted to be regulatory, particularly enhancers. The widespread sharing of eMS across sex and ethnic subgroups is a notable feature of these data which supports a hypothesis that basic features of the epigenetic regulation are common across sex and ethnicities. Clearly, more work is required to determine the precise mechanisms that underlie the correlations between methylation and expression described here. Nevertheless, these data provide a framework for development of more specific hypotheses concerning epigenetic regulation of gene expression, and identify high-value targets for further evaluation, such as GSTT1. In addition, we illustrate the use of Mendelian randomization and causal inference analyses to generate statistically derived (rather than experimentally derived) evidence from cross-sectional data to predict the direction of causation between methylation and expression.

Interestingly, our data suggest that eMS are similar in number and strength of correlation with expression to the much more frequently studied single-nucleotide expression quantitative trait loci (eQTLs) (19–22). However, the generally weak local pairwise correlation of CpG methylation is quite different from that of SNPs. This is anticipated given that linkage disequilibrium

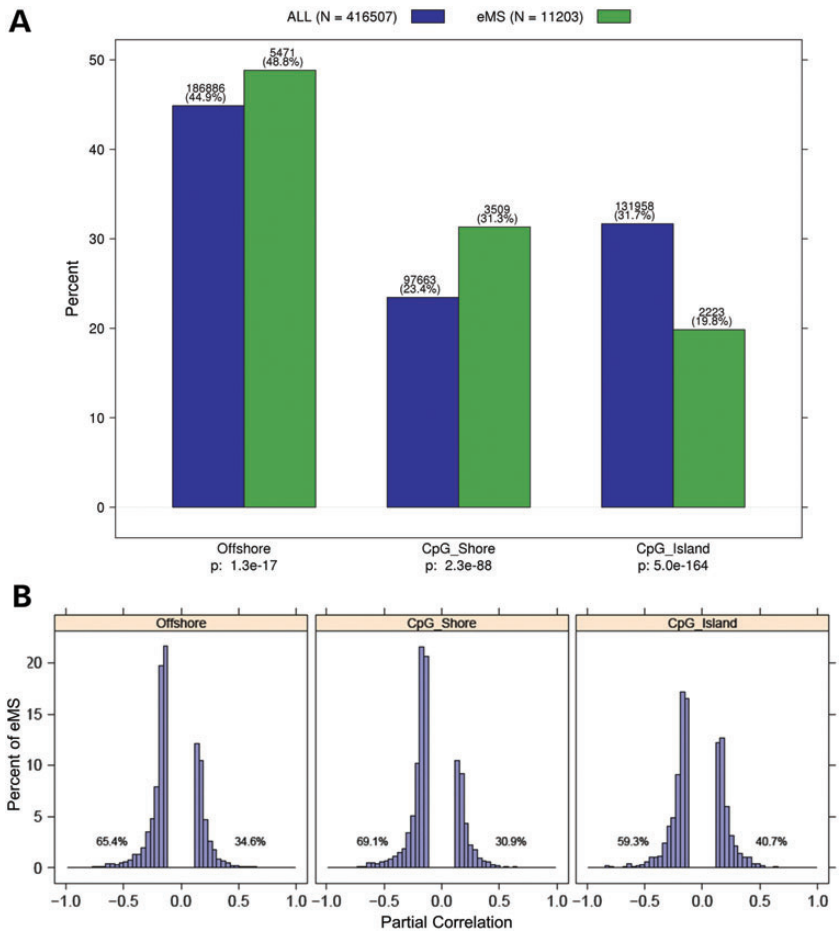

Figure 3. Distribution of 416K CpG sites on the Illumina array and 11 203 eMS across different regions related to CpG islands, and partial expression–methylation correlations of the 11 203 eMS for different regions related to CpG islands. (**A**) The absolute number of eMS was highest in 'offshore' sites (sites outside of annotated CpG islands—high CpG density regions), followed by CpG 'shores' [low CpG density regions in close proximity (~2 kb) to CpG islands], and CpG islands. However, relative to the number of sites assayed in each CpG annotation class, CpG shores and 'offshore' sites were highly enriched with eMS, whereas CpG islands contained relatively fewer eMS ($P = 1.3 \times 10^{-17}$–5.0×10^{-164}, chi-square test). (**B**) There were slightly more negative correlations between degree of eMS methylation and degree of expression of their associated gene transcripts across all locations in relation to CpG islands.

Table 1. Enrichment of eMS in ENCODE-annotated functional regions

Term	eMS count (%)	All CpG sites count (%)	Fold enrichment	*P*-value[a]
DHSs (m)[b]	853 (7.6)	26 782 (6.4)	1.18	2.5×10^{-7}
DHSs[c]	6011 (53.7)	189 873 (45.6)	1.18	$<2.2 \times 10^{-308}$
Enhancer[c]	7535 (67.3)	188 212 (45.2)	1.49	$<2.2 \times 10^{-308}$
Insulator[c]	686 (6.1)	21 690 (5.2)	1.18	1.1×10^{-5}
TFBS[c]	5905 (52.7)	195 174 (46.9)	1.12	$<2.2 \times 10^{-308}$

[a]Chi-square test.
[b]DHSs (m), DNase I-hypersensitive sites in monocytes (ENCODE).
[c]DHSs, transcription factor-binding sites (TFBS), enhancers and insulators reported in any available cell type (ENCODE).

Table 2. Gene set enrichment analysis of the 3093 genes associated with the 11 203 eMS using Gene Ontology

Term	Gene hit count	Fold enrichment[a]	*P*-value[b]	FDR
Immune response	162	1.50	2.6×10^{-8}	5.0×10^{-5}
Protein transport	169	1.42	7.1×10^{-7}	1.3×10^{-3}
Regulation of programmed cell death	165	1.31	1.6×10^{-4}	0.3

[a]Relative to the background set of 8370 genes examined.
[b]Fisher's exact test.

among SNPs is caused by historical patterns of mutation and recombination, whereas correlation between methylation sites is likely driven by biochemical processes that are limited to more narrowly defined genomic locations. Importantly, the less locally correlated structure of CpG methylation may allow better genomic localization of signals using eMS associations than eQTL analyses.

Hypermethylation of CpG sites is generally thought to be associated with transcriptional inactivation when occurring in gene promoter regions (4), but with transcriptional activation when occurring in gene bodies (4). Negative correlation

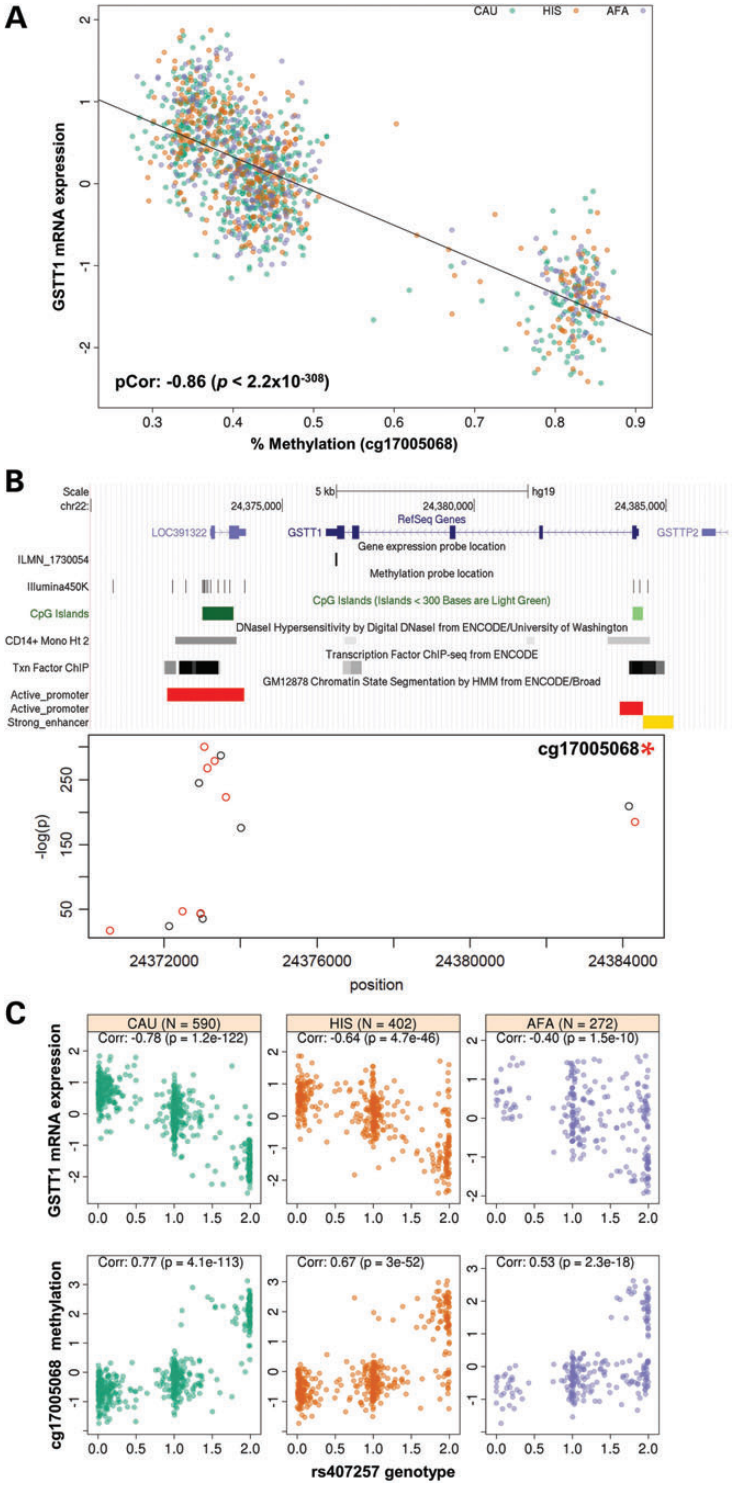

Figure 4. GSTT1 gene region. (**A**) Lower levels of *GSTT1* (glutathione *S*-transferase theta 1) mRNA expression are associated with hypermethylation of cg17005068. The inverse associations between the $\log_2$-transformed mRNA expression levels of *GSTT1* and %methylation of cg17005068 were consistent in Caucasians (CAU, green), Hispanics (HIS, orange) and African Americans (AFA, purple). (**B**) Genomic landscape (from UCSC genome browser, hg19) and regional association plot surrounding *GSTT1*; significance of *cis*-CpG methylation (eMS) associations with *GSTT1* mRNA gene expression ($-\log_{10}$ *P*-values) is plotted on the *y*-axis and eMS position on the *x*-axis for 16 eMS in this region that explained 77% of variation in mRNA expression in a linear regression model that included all 16 eMS as independent variables [shown as circles, with the asterisk denoting the most significant association (cg17005068)]. The red circles and the asterisk denote the independent eMS ($P < 0.05$) among the CpG sites shown in this figure (multiple linear regression). The eMS associating with *GSTT1* expression are located in close proximity to CpG islands (green) that overlap with DNase I-hypersensitivity clusters detected in monocytes, TFBS (ChIP-seq) in many cell types, and predicted promoter (red) and enhancer (orange) regions in B cells (GM12878). (**C**) A *cis*-acting SNP, rs407257, strongly associated with both cg17005068 methylation and *GSTT1* mRNA expression in Caucasians (CAU), Hispanics (HIS) and African Americans (AFA).

between gene expression and methylation in promoter regions was recently reiterated in a study of lymphoblastoid cell lines from 77 individuals (15). However, our data suggest that this paradigm may be only a weak approximation of the truth, as the majority of observed eMS associations were negative across all gene structural regions, with negative correlations being most frequent in promoter regions (69.4%), but only slightly less so in gene bodies (63.0%). On the other hand, positive correlations were not rare. Indeed, more than 30% of significant *cis*-methylation/expression correlations were positive—even in TFBS, where such positive correlations were previously thought to be rare or anomalous findings potentially due to specialized or cell-specific mechanisms (23). These findings suggest a great diversity and complexity of epigenetic regulatory mechanisms and highlight the need for further basic molecular biology investigations.

The strong association of cg17005068 methylation (and related loci) with mRNA expression of GSTT1 is potentially quite important given the protective role that GSTT1 plays against endogenous oxidative stress and exogenous toxins (24). More research is clearly warranted to understand the functional and ultimately the clinical implications of methylation of the cg17005068 loci. The full list of eMS harbors many other strong candidates for further functional evaluation as well.

It is likely that our results underestimate the total number of eMS. The Illumina methylation microarray does not assay every CpG site, although it covers most of the known or potentially important genomic regions with base-pair resolution. The Illumina transcript expression microarray design has limited sensitivity for low-abundance transcripts and limited coverage of alternative splice forms. Also, although we investigated CpG sites up to 1 Mb from TSSs, some regulation probably occurs beyond that distance. Since our study primarily focused on purified monocytes, it is likely that not all the identified eMS and their features will generalize to other cell types. Lastly, we cannot rule out effects of unmeasured confounding or reverse causality that gene expression can affect methylation (25).

Monocytes play a major role in immune function and are involved in the development of common diseases such as cardiovascular disease and type II diabetes. The identified eMS were enriched with immune response genes known to contribute to these and many other chronic diseases. This study establishes the feasibility of large-scale studies of methylation and expression, and lays the foundation for unraveling the mechanisms of epigenetic regulation of monocyte-related diseases. Future work in MESA will explore the associations of monocyte eMS and gene expression with the genetic, physiological, environmental and clinical characteristics of the MESA population.

MATERIALS AND METHODS

Study population

MESA was designed to investigate the prevalence, correlates and progression of subclinical cardiovascular disease in a population cohort of 6814 participants. Since its inception in 2000, five clinic visits collected extensive clinical, sociodemographic, lifestyle and behavior, laboratory, nutrition and medication data (25). The present analysis is primarily based on analyses of purified monocyte samples from the April 2010–February 2012 examination of 1264 randomly selected MESA participants [55–94 years old, Caucasian (47%), African American (21%), Hispanic (32%) and female (51%)] from four MESA field centers (Baltimore, MD; Forsyth County, NC; New York, NY; and St Paul, MN). The study protocol was approved by the Institutional Review Board at each site. All participants signed informed consent.

Purification of monocytes

Centralized training of technicians, standardized protocols and extensive QC measures were implemented for collection, on-site processing and shipment of MESA specimens and routine calibration of equipment. Blood was initially collected in sodium heparin-containing Vacutainer CPT^{TM} cell separation tubes (Becton Dickinson, Rutherford, NJ, USA) to separate peripheral blood mononuclear cells from other elements within 2 h from blood draw. Subsequently, monocytes were isolated with the anti-CD14-coated magnetic beads, using AutoMACs automated magnetic separation unit (Miltenyi Biotec, Bergisch Gladbach, Germany). Based on flow cytometry analysis of 18 specimens, monocyte samples were consistently $>90\%$ pure.

DNA/RNA extraction

DNA and RNA were isolated from samples simultaneously using the AllPrep DNA/RNA Mini Kit (Qiagen, Inc., Hilden, Germany). DNA and RNA QC metrics included optical density measurements, using a NanoDrop spectrophotometer and evaluation of the integrity of 18s and 28s ribosomal RNA. Additional RNA QC testing was performed using the Agilent 2100 Bioanalyzer with RNA 6000 Nano chips (Agilent Technology, Inc., Santa Clara, CA, USA) following manufacturer's instructions. RNA with RIN (RNA integrity) scores >9.0 was used for global expression microarrays. The median RIN for our 1264 samples was 9.9.

Global expression quantification

The Illumina HumanHT-12 v4 Expression BeadChip and Illumina Bead Array Reader were used to perform the genome-wide expression analysis, following the Illumina expression protocol. The Illumina TotalPrep-96 RNA Amplification Kit (Ambion/Applied Biosystems, Darmstadt, Germany) was used for reverse transcription, and amplification with 500 ng of input total RNA (at 11 μl). An amount of 700 ng of biotinylated cRNA was hybridized to a BeadChip at 58°C for 16–17 h. To avoid potential biases due to batch, chip and position effects, a stratified random sampling technique was used to assign individual samples (including 24 common control samples) to specific BeadChips (12 samples/chip) and chip position.

Epigenome-wide methylation quantification

The Illumina HumanMethylation450 BeadChip and HiScan reader were used to perform the epigenome-wide methylation analysis. The EZ-96 DNA Methylation^{TM} Kit (Zymo Research, Orange, CA, USA) was used for bisulfate conversation with 1 μg of input DNA (at 45 μl). An amount of 4 μl of bisulfite-converted DNA

were used for DNA methylation assays, following the Illumina Infinium HD Methylation Protocol. This consisted of a whole-genome amplification step followed by enzymatic end-point fragmentation, precipitation and resuspension. The resuspended samples were hybridized on HumanMethylation 450 BeadChips at 48°C for 16 h. The individual samples were assigned to the BeadChips and to chip position, using the same sampling scheme as that for the expression BeadChips.

Pre-processing of microarray data

Data pre-processing and QC analyses were performed in *R* (http://www.r-project.org/) using *Bioconductor* (http://www.bioconductor.org/) packages. For expression data, data corrected for local background were obtained from Illumina's proprietary software GenomeStudio. QC analyses and bead-type summarization (average bead signal for each type after outlier removal) were performed using the *beadarray* package (26). Detection *P*-values were computed using the negative controls on the array. The *neqc* function of the *limma* (27) package was used to perform a normal–exponential convolution model analysis to estimate non-negative signal, quantile normalization using all probes (gene and control, detected and not detected) and samples, addition of a recommended (small) offset, $\log_2$ transformation and elimination of control probe data from the normalized expression matrix. Multidimensional scaling plots showed that the five common control samples were highly clustered together and identified three outlier samples, which were excluded subsequently.

The Illumina HumanHT-12 v4 Expression BeadChip included 48K transcripts. Statistical analyses were limited to probes retained after applying the following QC elimination criteria: probes for the X or the Y chromosome, non-detectable expression in ≥90% of MESA samples using a detection *P*-value cut-off of 0.0001, existence of any known SNPs, overlap with a repetitive element or region, low variance across the samples (<10th percentile) or putative and/or not well-characterized genes, i.e. gene names starting with KIAA, FLJ, HS, Cxorf, MGC or LOC. We included 8370 autosomal gene transcripts for analysis.

Bead-level methylation data were summarized in *GenomeStudio*. Because a two-channel system and both Infinium I and II assays were used, normalization was performed in several steps using the *lumi* package (28). We first adjusted for color bias using 'smooth quantile normalization'. Next, the data were background-adjusted by subtracting the median intensity value of the negative control probes. Lastly, data were normalized across all samples by standard quantile normalization applied to the bead-type intensities and combined across Infinium I and II assays and both colors. QC measures included checks for sex and race/ethnicity mismatches, and outlier identification by multidimensional scaling plots. The final methylation value for each methylation probe was computed as the *M*-value, essentially the log ratio of the methylated to the unmethylated intensity (29). The *M*-value is well suited for high-level analyses and can be transformed into the beta-value, an estimate of the percent methylation of an individual CpG site that ranges from 0 to 1 [*M* is logit(beta-value)].

The Illumina HumanMethylation450 BeadChip included probes for 485K CpG sites. Of these 485K CpG sites, 416 507 passed the QC elimination criteria including: probes for the X or the Y chromosome, 'detected' methylation levels in <90% of MESA samples using a detection *P*-value cut-off of 0.05, existence of any SNPs within 10 base pairs of the targeted CpG site, location outside of the 1 MB intervals on both sides of each gene or overlap with a repetitive element or region.

Pre-processing with global normalization removed large position and chip effects across all probes; however, probe-specific chip effects were found for some CpG sites and gene transcripts, whereas probe-specific position effects existed for some CpG sites but were ignorable for all gene transcripts. These probe-specific effects were included as covariates in all subsequent analyses.

Association analyses

Association analyses were performed using the linear model (*lm*) function and the *stepAIC* function of the *MASS* package in *R*. To identify *cis*-methylation sites associated with gene expression, we fit separate linear regression models with the *M*-value for each CpG site (adjusted for methylation chip and position effects) as a predictor of transcript expression for any autosomal gene within 1 Mb of the CpG site in question. Covariates were age, sex, race/ethnicity, study site, expression chip and residual sample contamination with non-targeted cells (non-monocytes, see what follows). Sex- and ethnicity-stratified analyses were performed as an internal validation and check of generalizability. To look for potential population stratification, we used EIGENSTRAT (30) to compute principal components (PCs) to infer the genetic ancestry for each race and overall, based on MESA Affymetrix 6.0 array genotype data (31), and examined the association between the first five PCs and gene expression. Less than 0.03% expression transcripts were associated with the first two PCs in the Caucasian and African-American populations; however, 14.7% of gene expression transcripts in the Hispanic population were associated with the first two PCs (FDR < 0.05). Therefore, analyses in the Hispanic population were adjusted for the first two PCs. Adjustment for PCs ($n = 1202$) yielded similar results as self-reported ethnicity ($n = 1264$). Therefore, analyses for overall were adjusted for the self-reported ethnicity to maximize the sample size and control for the potential cultural and socioeconomic factors. Normality-based *P*-values were obtained for all tests (they were highly similar to permutation-based *P*-values and produced essentially the same ranking in a subset of 360 samples). *P*-values were adjusted for multiple testing using the *q*-value false recovery rate (FDR) method (32). To minimize false-positive results, we used the stringent FDR threshold of 0.001.

To estimate residual sample contamination for monocyte data analyses, we generated separate enrichment scores for neutrophils, B cells, T cells and natural killer cells. We implemented a gene set enrichment analysis (33) to calculate the enrichment scores, using the gene signature of each blood cell type in the ranked list of expression values for each MESA sample. The cell type-specific signature genes were selected from previously defined lists (34) and passed the following additional QC filters: at least 4-fold more highly expressed in the targeted cell type than in any other cell populations and low expression levels in monocytes.

To minimize spurious associations due to a few highly influential data points, we calculated Cook's distance (35) for each data point and repeated the analysis after removing the four samples with the highest Cook's distance. We then removed associations whose *P*-values no longer fell below the FDR-based significance threshold. Because the correlation structure of the expression and methylation profiles may increase bias and variance of the FDR estimates (36), we verified control of the FDR by a permutation approach, where the columns of the methylation matrix were permuted. The permutation-based estimate of the FDR level based on five replicates of the permuted data was 0.0018, quite close to the FDR threshold of 0.001.

Because the methylation profiles of eMS associated with the same gene showed various degrees of correlation, for each expressed gene with >1 significantly associated eMS, we performed step-wise selection (backward/forward starting with the full model, using the Akaike Information Criterion) to estimate the 'independent' number of eMS jointly affecting the gene's expression.

Causal inference

Inference of causal relationships, utilizing Mendelian randomization, between gene expression and biomedical phenotypes is well established in Genetical Systems Biology (18,37). The SNP data were derived from MESA Affymetrix 6.0 array genotype data (31). To infer the relationships between eMS and gene expression, we used structural equation modeling as implemented in the *R* package *lavaan* (38) to compare the fit to the data of six alternative causal models depicted in Supplementary Material, Figure S10. We then compare these six models using well-established SEM fit indices (39) shown in Supplementary Material, Table S3.

Functional annotation analysis

Functional annotation analysis was performed using data from the ENCODE project (17) accessed through the UCSC Genome Browser at http://genome.ucsc.edu/ (40). DNaseI hypersensitive areas were assayed in a large collection of cell types including monocytes (Digital DNaseI Hypersensitivity Clusters from ENCODE). We examined the eMS enrichment of DHSs in monocytes, although the monocyte samples were collected from only one subject, as well as the enrichment of DHSs in all available cell types. TFBS were assayed in a large collection of cell types; however, monocytes were investigated with only one transcription factor (CTCF); therefore, functional studies included all TFBS, from all available cell types (Transcription Factor ChIP-seq from ENCODE). Enhancer and insulator regions were bioinformatically predicted in nine different cell types, not including monocytes (chromatin state segmentation by HMM from ENCODE/Broad). The CpG sites that overlapped a predicted strong enhancer (State 4/5) or weak/poised enhancer (State 6/7) in any cell type were reported as overlapping an enhancer. The CpG sites that overlapped a predicted insulator (State 8) in any cell type were reported as overlapping an insulator. Additionally, annotation analysis of the methylation-associated transcripts was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (41) to test for over-representation of functional categories or pathways.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

The authors thank the investigators, the staff and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org. DHSs data were collected by the University of Washington ENCODE group. TFBS data were collected by the Myers Lab at the HudsonAlpha Institute for Biotechnology and by the laboratories of Michael Snyder, Mark Gerstein and Sherman Weissman at Yale University, Peggy Farnham at UC Davis, Kevin Struhl at Harvard, Kevin White at the University of Chicago and Vishy Iyer at the University of Texas Austin. Enhancer and insulator data were generated at the Broad Institute and in the Bradley E. Bernstein laboratory at the Massachusetts General Hospital/Harvard Medical School, and the chromatin state segmentation was produced by Manolis Kellis's Computational Biology group at the Massachusetts Institute of Technology.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by contracts N01-HC from the National Heart, Lung and Blood Institute and by grants (UL1-RR-024156, UL1-RR-025005) from the NIH. The MESA Epigenomics and Transcriptomics Study was funded by a National Heart, Lung and Blood Institute grant (R01HL101250) to Wake Forest University Health Sciences. The Encyclopedia of DNA Elements (ENCODE) project data generation and analysis was supported by funds from the National Human Genome Research Institute (ENCODE), the Burroughs Wellcome Fund, Howard Hughes Medical Institute, National Science Foundation, Sloan Foundation, Massachusetts General Hospital and the Broad Institute. ENCODE data comes from grants led by Bradley Bernstein (Broad Institute), Richard Myers (HudsonAlpha Institute), Michael Snyder (Stanford), Gregory Crawford (Duke) and John Stamatoyannopoulos (University of Washington).

REFERENCES

1. Bell, A.C. and Felsenfeld, G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482–485.
2. Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33** (suppl.), 245–254.
3. Esteller, M. (2007) Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum. Mol. Genet.*, **16** (Spec no. 1), R50–R59.
4. Portela, A. and Esteller, M. (2010) Epigenetic modifications and human disease. *Nat. Biotechnol.*, **28**, 1057–1068.
5. Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R., Deaton, A., Andrews, R., James, K.D. *et al.* (2010) Cpg islands

influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, **464**, 1082–1086.
6. Hui, R., Macmillan, R.D., Kenny, F.S., Musgrove, E.A., Blamey, R.W., Nicholson, R.I., Robertson, J.F. and Sutherland, R.L. (2000) INK4a gene expression and methylation in primary breast cancer: overexpression of p16INK4a messenger RNA is a marker of poor prognosis. *Clin. Cancer Res.*, **6**, 2777–2787.
7. Pike, B.L., Greiner, T.C., Wang, X., Weisenburger, D.D., Hsu, Y.H., Renaud, G., Wolfsberg, T.G., Kim, M., Weisenberger, D.J., Siegmund, K.D. *et al.* (2008) DNA methylation profiles in diffuse large B-cell lymphoma and their relationship to gene expression status. *Leukemia*, **22**, 1035–1043.
8. Ball, M.P., Li, J.B., Gao, Y., Lee, J.H., LeProust, E.M., Park, I.H., Xie, B., Daley, G.Q. and Church, G.M. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.*, **27**, 361–368.
9. Li, M., Balch, C., Montgomery, J.S., Jeong, M., Chung, J.H., Yan, P., Huang, T.H., Kim, S. and Nephew, K.P. (2009) Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC Med. Genomics*, **2**, 34.
10. Sun, Z., Asmann, Y.W., Kalari, K.R., Bot, B., Eckel-Passow, J.E., Baker, T.R., Carr, J.M., Khrebtukova, I., Luo, S., Zhang, L. *et al.* (2011) Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One*, **6**, e17490.
11. Lee, S.T., Xiao, Y., Muench, M.O., Xiao, J., Fomin, M.E., Wiencke, J.K., Zheng, S., Dou, X., de Smith, A., Chokkalingam, A. *et al.* (2012) A global DNA methylation and gene expression analysis of early human B-cell development reveals a demethylation signature and transcription factor network. *Nucleic Acids Res.*, **40**, 11339–11351.
12. van Eijk, K.R., de Jong, S., Boks, M.P., Langeveld, T., Colas, F., Veldink, J.H., de Kovel, C.G., Janson, E., Strengman, E., Langfelder, P. *et al.* (2012) Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, **13**, 636.
13. Pai, A.A., Bell, J.T., Marioni, J.C., Pritchard, J.K. and Gilad, Y. (2011) A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.*, **7**, e1001316.
14. Lam, L.L., Emberly, E., Fraser, H.B., Neumann, S.M., Chen, E., Miller, G.E. and Kobor, M.S. (2012) Factors underlying variable DNA methylation in a human community cohort. *Proc. Natl Acad. Sci. USA*, **109** (Suppl. 2), 17253–17260.
15. Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y. and Pritchard, J.K. (2011) DNA Methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.*, **12**, R10.
16. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J. *et al.* (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.*, **6**, e1000952.
17. ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
18. Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C. *et al.* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, **37**, 710–717.
19. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.
20. Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M. and Burdick, J.T. (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, 1365–1369.
21. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.
22. Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H. *et al.* (2010) Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS One*, **5**, e10693.
23. Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegan, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D. *et al.* (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490–495.
24. Kordi-Tamandani, D.M., Hashemi, M., Birjandian, E., Bahari, A., Valizadeh, J. and Torkamanzehi, A. (2011) Lack of association of GSTT1 and GSTP1 genes methylation and their expression profiles with risk of NAFLD in a sample of Iranian patients. *Clin. Res. Hepatol. Gastroenterol.*, **35**, 387–392.
25. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R. Jr., Kronmal, R., Liu, K. *et al.* (2002) Multi-ethnic study of atherosclerosis: objectives and design. *Am. J. Epidemiol.*, **156**, 871–881.
26. Dunning, M.J., Smith, M.L., Ritchie, M.E. and Tavare, S. (2007) beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, **23**, 2183–2184.
27. Smyth, G.K., Michaud, J. and Scott, H.S. (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067–2075.
28. Du, P., Kibbe, W.A. and Lin, S.M. (2008) Lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, **24**, 1547–1548.
29. Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L. and Lin, S.M. (2010) Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.
30. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
31. Fox, C.S., White, C.C., Lohman, K., Heard-Costa, N., Cohen, P., Zhang, Y., Johnson, A.D., Emilsson, V., Liu, C.T., Chen, Y.D. *et al.* (2012) Genome-wide association of pericardial fat identifies a unique locus for ectopic fat. *PLoS Genet.*, **8**, e1002705.
32. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
33. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
34. Abbas, A.R., Baldwin, D., Ma, Y., Ouyang, W., Gurney, A., Martin, F., Fong, S., van Lookeren, C.M., Godowski, P., Williams, P.M. *et al.* (2005) Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun.*, **6**, 319–331.
35. Henderson, A.R. (2006) Information for authors: is the advice regarding the reporting of residuals in regression analysis incomplete? Should Cook's distance be included? *Clin. Chem.*, **52**, 1848–1850.
36. Schwartzman, A. and Lin, Y. (2011) The effect of correlation in FDR estimation. *Biometrika*, **98**, 199–214.
37. Aten, J.E., Fuller, T.F., Lusis, A.J. and Horvath, S. (2008) Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Syst. Biol.*, **2**, 34.
38. Rosseel, Y. (2012) Lavaan: an R package for structural equation modeling. *J. Stat. Software*, **48**, 1–36.
39. Hooper, D., Coughlan, J. and Mullen, M. (2008) Structural equation modelling: guidelines for determining model fit. *Electron. J. Bus. Res. Methods*, **6**, 53–60.
40. Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G. *et al.* (2013) ENCODE data in the UCSC genome browser: year 5 update. *Nucleic Acids Res.*, **41**, D56–D63.
41. Dennis, G. Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, 3.