Genome-Wide Variation of Cytosine Modifications Between European and African Populations and the Implications for Complex Traits

Erika L. Moen,^{*,1} Xu Zhang,^{†,1} Wenbo Mu,[‡] Shannon M. Delaney,[§] Claudia Wing,[§] Jennifer McQuade,[§]

Jamie Myers,[§] Lucy A. Godley,^{*,§,***} M. Eileen Dolan,^{*,§,***,††} and Wei Zhang^{*,††,§§}

^{*}Committee on Cancer Biology, [§]Section of Hematology/Oncology, Department of Medicine, and ^{††}Committee on Clinical Pharmacology and Pharmacogenomics, University of Chicago, Chicago, Illinois 60637, [†]Section of Hematology/Oncology, Department of Medicine, [‡]Department of Pediatrics, and ^{**}Institute of Human Genetics, University of Illinois, Chicago, Illinois 60612, ^{***}University of Chicago Comprehensive Cancer Center, Chicago, Illinois 60637, and ^{§§}University of Illinois Cancer Center, Chicago, Illinois 60612

ABSTRACT Elucidating cytosine modification differences between human populations can enhance our understanding of ethnic specificity in complex traits. In this study, cytosine modification levels in 133 HapMap lymphoblastoid cell lines derived from individuals of European or African ancestry were profiled using the Illumina HumanMethylation450 BeadChip. Approximately 13% of the analyzed CpG sites showed differential modification between the two populations at a false discovery rate of 1%. The CpG sites with greater modification levels in European descent were enriched in the proximal regulatory regions, while those greater in African descent were biased toward gene bodies. More than half of the detected population-specific cytosine modifications could be explained primarily by local genetic variation. In addition, a substantial proportion of local modification quantitative trait loci exhibited population-specific effects, suggesting that genetic epistasis and/or genotype × environment interactions could be common. Distinct correlations were observed between gene expression levels and cytosine modifications in proximal regions and gene bodies, suggesting epigenetic regulation of interindividual expression variation. Furthermore, quantitative trait loci associated with population-specific modifications can be colocalized with expression quantitative trait loci and single nucleotide polymorphisms previously identified for complex traits with known racial disparities. Our findings revealed abundant population-specific cytosine modifications and the underlying genetic basis, as well as the relatively independent contribution of genetic and epigenetic variations to population differences in gene expression.

DNA methylation is a covalent cytosine modification that occurs at the C-5 position of cytosines at CpG dinucleotides and is dispersed unevenly over the genome (Bird 2002). Interindividual variation in cytosine modifications can be affected by both the stable underlying genetic sequence and dynamic environmental influences (Flanagan *et al.* 2006; Bock *et al.* 2008). Cytosine modifications are

known to play an important role in the regulation of gene expression, with promoter methylation acting to silence gene expression (Grewal and Moazed 2003). Previous studies of human variation in gene expression have shown that differential gene expression can influence a variety of complex traits, including susceptibilities to common diseases and variation in drug response (Schadt *et al.* 2005; Emilsson *et al.* 2008; Cookson *et al.* 2009).

The lymphoblastoid cell lines (LCLs) from the International HapMap Project (HapMap 2003; HapMap 2005) have been used recently for investigating within- and between-population differences in promoter methylation (Bell *et al.* 2011; Fraser *et al.* 2012). Furthermore, previous work from our group and others has demonstrated that common genetic variants and microRNAs contributed to the variation in gene expression between LCLs derived from individuals of

Copyright © 2013 by the Genetics Society of America

doi: 10.1534/genetics.113.151381

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.151381/-/DC1>.

Manuscript received March 17, 2013; accepted for publication June 3, 2013
Raw and processed data have been deposited into the Gene Expression Omnibus (GEO series number GSE39672).

¹These authors contributed equally to this work.

²Corresponding authors: 840 S. Wood St., CSB 1200 (M/C 856), Chicago, IL 60612.
E-mail: weizhan1@uic.edu; and Room 7100, 900 E. 57th St., Chicago, IL 60637.

E-mail: edolan@medicine.bsd.uchicago.edu

Yoruba people from Ibadan, Nigeria (YRI) and Caucasian residents of European ancestry from Utah (CEU) (Stranger *et al.* 2007; Zhang *et al.* 2008a; Huang *et al.* 2011). Due to tissue-specific DNA methylation patterns (Rakyan *et al.* 2008), studies in LCLs have the distinct advantage of being a pure population of B cells, as primary samples from humans can include more than one cell type, which may confound downstream analyses.

We reasoned that an evaluation of the natural variation in cytosine modifications at single-base resolution beyond the promoter regions could help elucidate the epigenetic contribution to interpopulation gene expression differences as well as other phenotypic differences, such as risks for common diseases and drug response variation. Furthermore, an improved understanding of the genetic basis of cytosine modification variation and the relationships between modification quantitative trait loci (mQTL) and expression quantitative trait loci (eQTL) would provide novel insights into the genetic architectures of gene expression and other complex traits.

Specifically, we collected genome-wide cytosine modification data on more than 480,000 CpG sites using the Illumina Infinium HumanMethylation450 BeadChip (450K array) (Bibikova *et al.* 2011) in a collection of 60 unrelated CEU and 73 unrelated YRI samples, for which comprehensive genotypes of single nucleotide polymorphisms (SNPs) (Frazer *et al.* 2007) and gene expression (Zhang *et al.* 2008a) data are available. Technically, the 450K array requires bisulfite-converted DNA to distinguish modified from unmodified cytosines. Although 5-methylcytosine (5-mC) is the most common covalent cytosine modification in the human genome, 5-mC can be oxidized by the ten-eleven translocation (TET) enzymes to 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine, and 5-carboxylcytosine (Tahiliani *et al.* 2009; Ito *et al.* 2011). Because bisulfite conversion cannot distinguish between 5-mC and 5-hmC (Huang *et al.* 2010; Jin *et al.* 2010), we have chosen to use the term “cytosine modification” to avoid any implied bias. In this study, we identified the differentially modified cytosines between the CEU and YRI samples, assessed the genetic contribution to these differences, and investigated the implications of mQTL on gene expression and other complex traits with known racial disparities.

Materials and Methods

Sample preparation and cytosine modification profiling

We purchased 50 µg of genomic DNA for 60 unrelated CEU and 74 unrelated YRI (59 phase 1/2 and 15 phase 3 samples) from the Coriell Institute for Medical Research (Camden, NJ). To confirm the cell line identities, we genotyped 47 SNPs from the Sequenom iPLEX Sample ID Plus Panel (Sequenom, San Diego, CA) and compared them with the HapMap data (release 28) for 24 randomly selected samples. As an internal control, we selected 10 LCLs (4 CEU and 6 YRI samples), previously pelleted by the Pharmacogenetics of Anticancer Agents Research (PAAR) Group Cell Core at the University of Chicago, for DNA isolation and

array profiling. DNA was isolated using Qiagen Gentra Puregene Core Kit B (Qiagen, Germantown, MD).

The University of Chicago Genomics Core Facility was provided with 800 ng of DNA for bisulfite conversion with the Zymo EZ cytosine modification kit (Zymo Research, Irvine, CA). Cytosine modification levels were profiled using the Illumina 450K array (Illumina, San Diego, CA), according to the manufacturer's protocol. The efficacy of the bisulfite conversion was checked by PCR amplification and DNA sequencing of selected genomic regions. Approximately 150 ng of bisulfite-converted DNA from each sample was used for array hybridization. To limit the potential bias due to experimental batches, samples were randomized by population identity and hybridized in three batches.

Relative Epstein-Barr virus copy numbers

Quantitative real-time PCR (qRT-PCR) was performed to determine relative baseline Epstein-Barr virus (EBV) copy numbers in the DNA samples using a TaqMan assay, which interrogated a 66-bp fragment at the DNA polymerase locus. A 90-bp assay from the *NRF1* locus (encoding nuclear respiratory factor 1) was used as an internal reference (Supporting Information, Table S1). Genomic DNA template (0.5 ng) and Gene Expression master mix was used for each TaqMan reaction performed according to the manufacturer's protocol (Applied Biosystem, Foster City, CA). A standard curve method was implemented using quantities of total genomic DNA obtained from the IB4 cell line (gift from Janet Rowley's Laboratory, University of Chicago), which was run as a reference line on each plate for calculating the relative EBV copy numbers (Table S2). Any samples that did not achieve <20% standard deviation (in triplicates) were dropped from the analysis.

Cytosine modification data processing

To remove probes that potentially cross-hybridize, we aligned 482,421 probe sequences to the human genome (hg19) (Zhang *et al.* 2012). Both forward and reverse strands of the genome reference were bisulfite converted *in silico*, in which all cytosines within CpGs were converted to N's and all remaining C's to T's. All cytosines within CpG sites in the probe sequences were converted to N's. Probe sequences were aligned to the converted genome reference using Bowtie2 (v2.0.0 beta5) (Langmead and Salzberg 2012) in an end-to-end exhaustive alignment mode, with seed size of 16 bases, interval between seeds of 12 bases, allowing one mismatch in seeds. The scoring parameters were set so that the alignment allowing ≤4 ambiguous bases (*i.e.*, N's) and ≤2 mismatches in total. The alignment resulted in 340,658 perfectly matching unique probes (File S1). We further removed probes containing common SNPs within 20 bp of the interrogated CpG sites that had minor allele frequency (MAF) >0.01 from dbSNP v135 (Sherry *et al.* 2001) (Table S3).

Cytosine modification levels were summarized with the Illumina GenomeStudio-Modification Module v1.9 (GenomeStudio). The β-value was calculated as the proportion of

modified probe intensity over the sum of both modified and unmodified probe intensities; the *M*-value was calculated as the $\log_2$ ratio of the intensities of modified probe vs. unmodified probe. We detected and removed NA18862, which contained >5% of failed CpG probes with detection $P > 0.01$ (Fernandez *et al.* 2012). CpG probes with detection $P > 0.01$ in $\geq 5\%$ samples were also removed. Background correction (Borevitz *et al.* 2003) and quantile normalization (Bolstad *et al.* 2003) were then performed on the *M*-values or β -values of the remaining 290,577 probes across all samples. COMBAT (Johnson *et al.* 2007) was used to correct the batch effects of array hybridization. Comparison of the β -values between 10 DNA samples obtained from the Coriell Institute and 10 internal controls indicated a stronger correlation between sample pairs from the two sources than between random pairs of samples (r^2 intraquantile range: 0.94–0.98 vs. 0.80–0.86), confirming the stability of cytosine modification profiles.

Genic distributions of cytosine modifications

To examine the genic distributions of cytosine modifications within populations, population-specific CpGs, and CpGs correlated with gene expression, we aligned transcription start sites (TSSs) and end sites (TESs) across 18,036 genes. Regions 10 kb upstream of TSSs and 10 kb downstream of TESs were binned by 500 bp. Gene bodies were divided into 10 quantiles by length. Gene regions were therefore defined as regions covering 10 kb upstream of TSSs, 5'-UTRs (untranslated regions), gene bodies, 3'-UTRs, and 10 kb downstream of TESs.

Identification of differentially modified cytosines between populations

The final analysis set was composed of 283,540 autosomal CpG probes in 60 CEU and 73 YRI samples. Since the *M*-value was shown to provide better detection sensitivity at extreme modification levels (Du *et al.* 2010), we used the *M*-value in the statistical comparisons unless otherwise mentioned. A linear model: cytosine modification level $\sim$ population + gender + error, was used to identify differentially modified cytosines. False discovery rate (FDR) was estimated by 100 permutations across samples using an approximate test (Anderson and Robinson 2002). The results obtained by linear regression were strongly correlated with results obtained by the nonparametric Wilcoxon rank-sum test (Spearman's $\rho = 0.92$ across probe *P*-values). We further evaluated potential confounders including intrinsic growth rate (Im *et al.* 2012) and EBV copy number (Table S2) for their effects on the differential cytosines using multivariate linear regression models.

Genetic association of cytosine modification variation

SNP genotypes were obtained from the International HapMap Project (release 28). For SNP association tests within each population, SNPs that were called in at least 48 CEU samples or 48 YRI samples with $MAF > 0.05$ and that were not significantly deviated from the Hardy-Weinberg equilibrium

(HWE) ($P > 0.0001$) were selected. Cytosine modification levels of the 36,597 differential CpG sites detected at $FDR < 1\%$ were regressed on SNP allelic dosage using an additive model, with gender as a covariate. FDR was estimated by 100 permutations. For the local scan of SNPs within ± 100 kb of the target CpG sites, in total 1,530,751 SNPs and 36,577 CpG sites (6,227,494 associations) were tested in the CEU samples, resulting in 28,299 significant associations at 5% FDR. In the YRI samples, in total 1,696,808 SNPs and 36,577 CpG sites (6,843,483 associations) were tested, resulting in 19,701 significant associations. F_{st} values, a measure of population differentiation, were calculated for SNPs in dbSNP v135 (Sherry *et al.* 2001) with allele frequencies obtained from the unrelated individuals in the complete collection of CEU and YRI samples.

For SNP association tests across populations, SNPs that had genotypes in at least 96 samples with $MAF > 0.05$ across populations and that were not significantly deviated from the HWE ($P > 0.0001$) in either the CEU or YRI samples were selected. This resulted in 1,766,136 SNPs and 36,580 CpG sites, totaling 7,136,801 tests. Cytosine modification levels were then regressed on SNP allelic dosage using an additive model, with gender as a covariate. FDR was estimated by 100 permutations. For each significant association, Pearson's correlation r^2 with modification level was estimated for both SNP effect and population identity.

Correlation analysis of cytosine modification and gene expression

Gene expression data were previously generated by our group using the Affymetrix Human Exon 1.0ST Array (exon array) (Zhang *et al.* 2008a). To avoid sample mix-ups (Westra *et al.* 2011), we genotyped 47 SNPs from the Sequenom iPLEX Sample ID Plus Panel in 106 random samples previously profiled for gene expression and maintained at the PAAR group. All sample identities were confirmed by comparing the genotype calls with the HapMap data (release 28). The exon array data were background corrected (Borevitz *et al.* 2003) and quantile normalized (Bolstad *et al.* 2003) across 176 CEU and YRI samples (in parents-child trios) after removing probes containing common SNPs ($MAF > 0.01$) based on dbSNP v135 (Sherry *et al.* 2001). In total, 14,669 transcript clusters (gene level) mapped to unique Entrez Gene IDs were evaluated for correlation between gene expression and the *M*-values of 208,808 CpG sites located within gene regions, corresponding to 269,316 CpG-gene pairs. Correlation analyses were carried out across 58 CEU and 57 YRI samples that had both expression and modification data, as well as within each population, using a linear model: gene expression $\sim$ cytosine modification + error. FDR was estimated by 100 permutations.

Result validation

We performed bisulfite sequencing within four loci, *LIPH* (encoding lipase member H), *TBX21* (encoding T-box 21), *DENND2D* (encoding DENN/MADD domain containing 2D),

and *CR2* (encoding complement component receptor 2), that were found to be differentially modified and expressed between the two populations. Bisulfite sequencing at these loci was performed on at least 20 random samples from the original DNA collection used in the 450K array profiling. Zymo-Taq polymerase and bisulfite primers (Table S4) were used to amplify ~200 bp fragments that included the CpG sites interrogated by the 450K array as well as adjacent CpGs.

RT-PCR was performed to measure the expression levels of *TBX21*, *DENND2D*, *LIPH*, and *CR2*. Total RNA was extracted from 5 million cells using Qiagen RNeasy Plus Mini kit following the manufacturer's protocol. mRNA was reverse transcribed to cDNA using Applied Biosystems High-Capacity Reverse Transcription kit. qRT-PCR was performed for each gene, and *hUB2M* [beta-2-microglobulin (NM_004048.2)] was included as an endogenous control using TaqMan Gene Expression Assays. The TaqMan primers and probes (*DENND2D*, Hs00227687_m1; *TBX21*, Hs00203436_m1; *LIPH*, Hs00975887_m1; and *CR2*, Hs01079096_m1) were labeled with a FAM reporter dye and MGB quencher dye. The *hUB2M* primer/probe mixture was labeled with VIC reporter dye and MGB quencher dye. The fast thermocycler parameters were: 95° for 20 sec and 40 cycles of 95° for 1 sec and then 60° for 20 sec, with ramping speeds of 1.6°–1.9°/sec. Each sample was run on a minimum of two plates in triplicate on each plate with standard deviation ≤15%.

Results

Variation of cytosine modifications within populations

We profiled cytosine modification levels on the Illumina 450K array for 60 CEU and 73 YRI samples (HapMap 2003; HapMap 2005). A total of 283,540 autosomal CpG sites were selected for further analyses (Table S3). In general, cytosine modification levels showed a bimodal distribution, representing hyper- and hypomodification (Figure S1). About 82% of the analyzed CpG sites fell within gene regions. Cytosine modification levels were greater in gene bodies than in proximal regulatory regions (Figure 1A). In contrast, the variability of cytosine modifications was greater around the proximal regulatory regions than in the gene bodies (Figure 1B). To assess the extent of co-modification between CpG sites, we estimated Spearman's ρ for all CpG pairs <5 kb apart. As expected, co-modification decreased with distance (Figure 1C). Beyond 1 kb, co-modification between CpG sites decreased to about the background level, indicating a much shorter linkage range than genetic variations. CpG pairs that fell within gene bodies exhibited consistently stronger co-modification than those located in other regions. Overall, the CEU and YRI samples showed a similar co-modification pattern (Figure S2).

Population differences in cytosine modifications

Across the CEU and YRI samples, population identity was a prominent variable of cytosine modifications by the principal components analysis (PCA) (Figure S3). We regressed mod-

Figure 1 Variation of cytosine modification levels within populations. (A) Genic distribution of cytosine modification levels. The 232,148 CpG sites located in gene regions were binned based on their relative positions in 18,036 genes. The mean β -value across CpG sites was calculated for each gene. The median (dot) and intraquantile range (vertical line) across genes are plotted against the positional bin. (B) Genic distribution of the variability of cytosine modification levels. Coefficients of variation (CV) for the β -values across samples were estimated in each population. The mean CV across CpG sites was calculated for each gene. The median (dot) and intraquantile range (vertical line) across genes are plotted against the positional bin. In A and B, the black and orange dots denote the medians for the CEU and YRI samples, respectively. The blue box denotes the region from TSS to TES. (C) Cytosine co-modification in gene regions. Across the CEU samples, Spearman's ρ was calculated for all CpG pairs <5 kb apart. The CpG pairs were then grouped by their relative distance and according to whether both cytosines fell in specific gene regions. The median (bar) and intraquantile range (line) of the signed ρ^2 across the CpG pairs are plotted against the distance, for 10 kb upstream of TSSs (upstream), 5'-UTRs (UTR5), gene bodies, 3'-UTRs (UTR3), and 10 kb downstream of TESs (downstream).

ification levels on population identity for each CpG site, with gender as a covariate. A total of 36,597 (13%) differential CpG sites were detected at nominal $P < 0.001$, corresponding to FDR of 1% (Table S5). These sites were not biased to potential genetic hybridization differences between populations. Particularly, probes containing rare SNPs within 20 bp of the interrogated CpG sites were not overrepresented (13.4% among rare SNP-containing probes vs. 12.9% among all analyzed probes), nor were probes overlapping with known copy number variants (CNVs) (Redon *et al.* 2006) (13.3% among probes overlapping CNV regions vs. 12.9% among all analyzed probes).

We examined two confounding factors for the LCLs, the EBV copy number (Table S2), and the intrinsic growth rate

(Im *et al.* 2012). With these covariates being accounted for, population identity remained as the most significant variable for the differential CpG sites (Figure S4), suggesting that the observed population differences were not likely due to these confounding factors. We also compared our data with a published dataset that profiled cytosine modifications for the CEU and YRI samples using the Illumina 27K array (Fraser *et al.* 2012). Focusing on the same set of CpG sites and samples, the regression coefficients of between-population differences from the two studies were strongly correlated ($r^2 = 0.72$) (Figure S5), demonstrating that the observed population differences were robust against technical variations across experiments and profiling platforms.

Since the LCLs of the CEU samples have been in culture for a longer time than the YRI samples, we tested whether this confounded the observed population differences. We pooled our 450K array data with cytosine modification profiles of whole blood primary samples from healthy individuals of Dutch descent that were studied on the same platform or the Illumina 27K array (Horvath *et al.* 2012). Applying PCA, we found that the primary variations across the CEU, YRI, and Dutch samples were consistent with the gradient of population identity rather than the gradient of cell type or cell culturing time. Our differential CpG sites further separated the CEU, YRI, and Dutch samples along the primary gradient (Figure S6).

Distinct genic distributions of population-specific cytosine modifications

Among the 36,597 differential CpG sites, 27,051 fell within the gene regions of 12,485 genes. The average proportion of differential CpG sites was greater in gene bodies and UTRs than in the ± 10 -kb flanking regions (Figure 2A). Notably, CpG sites with greater modification levels in the CEU samples (CEU > YRI) were enriched in the regions from 1 kb upstream of TSS to the first gene body quantile (average proportion: 7.5%), but relatively depleted in the regions from the second gene body quantile to the 3'-UTR (average proportion: 3.9%). In contrast, CpG sites with greater modification levels in the YRI samples (YRI > CEU) were enriched in the regions from the second gene quantile to the 3'-UTR (average proportion: 8.6%), relative to the flanking regions (average proportion: 3.9%). An example was shown for gene *STK39* (encoding serine threonine kinase 39), which contained several CEU > YRI sites in the promoter region and YRI > CEU sites in its gene body (Figure 2B).

To characterize the differential CpG sites with respect to other chromatin marks, we mapped these population-specific CpG sites to DNase I hypersensitivity and histone modification peaks derived using a CEU sample (NA12878) from the ENCODE (Encyclopedia of DNA Elements) Project (Dunham *et al.* 2012). A substantial proportion of the CEU > YRI sites (44–66%) overlapped with histone modification peaks that mark active regulatory elements and transcription/elongation. The proportion of the YRI > CEU sites that coincided with these histone modifications was markedly

Figure 2 Distribution of population-specific cytosine modifications. (A) The genic distribution of the CpG sites differentially modified between the CEU and YRI samples. The proportion of differential CpGs was plotted against the positional bin in the gene region. About 87% of the 14,550 CEU > YRI (black) and 65% of the 22,047 YRI > CEU (orange) CpG sites were located within gene regions. (B) The t -scores of differential CpGs are plotted against the chromosomal position (hg19) for gene *STK39* (encoding serine threonine kinase 39). The analyzed CpG sites are denoted by dark gray vertical lines. The CEU > YRI (blue) or YRI > CEU (orange) CpG sites are marked by stepwise lines. Exons are denoted by black segments along the horizontal line. The direction of transcription is denoted by a black arrow. (C) The proportions of CpG sites that overlapped with DNase I hypersensitive regions and histone modification peak regions among all analyzed CpG sites (dark gray), the CEU > YRI CpG sites (blue), and the YRI > CEU CpG sites (golden) are shown.

lower, between 2 and 6% (Figure 2C). In contrast, the YRI > CEU sites were more frequently mapped to peak regions of H3K27me3 (60 vs. 39%), a histone modification that marks repressive domains and silent developmental genes.

Genetic regulation of population differences in cytosine modifications

The observed population differences in cytosine modifications can be caused by both genetic and environmental factors. To assess the genetic contribution, we tested SNP association with cytosine modification levels for the 36,597 differential CpG sites. Within each population, cytosine modification levels were regressed on SNP allelic dosage, with gender as a covariate. Because of strong enrichment of mQTL within the ± 100 kb of target CpG sites (Figure S7), we focused on these regions for greater detection power. At 5% FDR, significant modification-SNP associations were

Figure 3 Genetic regulation of population differences in cytosine modifications. (A) The number (blue) and the effect size (golden) of SNP-CpG associations are shown as functions of distance between mQTL and CpG sites. mQTL in strong linkage disequilibrium ($r^2 > 0.8$) were pruned, resulting in 5704 associations for the CEU and 6802 associations for the YRI samples. The ± 100 -kb regions flanking target CpG sites were binned by 5 kb. The median (golden point) and intraquantile range (golden line) of association r^2 are plotted for the CEU (open circle) and YRI (solid square) samples. (B) An example of common mQTL. The G allele of rs2776937 was associated with lower cytosine modification level of cg10312802 located in *NRP1* (encoding neuropilin 1) across the CEU (black) and YRI (orange) samples. (C) An example of population-specific mQTL. The G allele of rs28544087 was associated with lower cytosine modification level of cg09307883 in *ANAPC2* (encoding anaphase promoting complex subunit 2) in the CEU (black) but not YRI (orange) samples. In B and C, the positions of CpGs and mQTL are denoted by orange crosses and blue triangles, respectively.

detected for 1861 CpG sites (5.1%) in the CEU samples and for 2125 CpG sites (5.8%) in the YRI samples (Table S6). The median r^2 of associations was 0.31 and 0.30 in the CEU and YRI samples, respectively. The number of associations showed an inverse relationship with the distance between mQTL and their target CpG sites, consistent with a *cis*-acting mode (Figure 3A). In contrast, the strength of associations was less dependent on the distance in general (Figure 3A).

Comparison of the association r^2 values between the two populations indicated extensive population specificity of mQTL effects. For modification-SNP associations that were significant in either population and that had fairly common SNP allele frequency ($MAF > 0.1$), the correlation of SNP association r^2 values between the two populations was very low (Pearson's correlation coefficient = -0.03) (Figure S8), suggesting that genetic epistasis and/or genotype $\times$ environment interactions could be common. An example of common mQTL was shown for a CpG site within gene *NRP1* (encoding neuropilin 1) (Figure 3B), and population-specific mQTL for a CpG site within gene *ANAPC2* (encoding anaphase-promoting complex subunit 2) (Figure 3C).

A total of 5237 modification-SNP associations for 497 CpG sites were detected in both populations. As expected, these shared mQTL were enriched in SNPs with high F_{st} values, demonstrating the genetic contribution to the observed population differences in cytosine modifications (Figure S9). It also implied that at a given differential CpG site, the statistical power in detecting mQTL was likely different between the two populations. To estimate the proportion of differential CpG sites under genetic control, we regressed

modification levels on SNP allelic dosage across populations, with gender as a covariate. Here population identity was left out from the regression model, given that for mQTL underlying differential CpG sites, population identity will more or less correlate with genetic variation. At 5% FDR, local SNP associations were detected for 96% of differential CpG sites. We then assessed the relative significance of genetic effect vs. population identity for these associations by comparing their correlation r^2 with modification levels. For 19,651 differential CpG sites (54%), the r^2 of modification levels with at least one mQTL was greater than that with population identity, suggesting a primary genetic contribution.

The role of mQTL in complex traits

We hypothesized that mQTL that underlie differentially modified cytosines between CEU and YRI might have implications on complex traits with known racial disparities. We obtained the top complex-trait associated SNPs in the National Human Genome Research Institute (NHGRI) genome-wide association studies (GWAS) catalog (Hindorff *et al.* 2009) and compared them to the mQTL we identified in both CEU and YRI (Table S6). A number of mQTL were associated with complex traits with racial disparities (Table 1), including several metabolic disorders, cardiovascular diseases, autoimmune disorders, and neurological disorders. For instance, five SNPs associated with cholesterol levels and cardiovascular diseases were annotated as an mQTL for a CpG in the promoter region of *APOA5* (encoding apolipoprotein A-V), which plays an important role in regulating the plasma triglyceride levels, a major risk factor for

Table 1 mQTL colocalized with SNPs detected for complex traits with racial disparities

SNP	CpG	CpG host gene	CpG location	Complex trait	PubMed
Metabolic and cardiovascular traits					
rs6917603	cg02184540	<i>NCRNA00171</i>	Body	Lipid metabolism	22286219
rs10927875	cg02890259	<i>HSPB7</i>	Body	Dilated cardiomyopathy	21459883
rs10903129	cg06961873	<i>TMEM57</i>	3'-UTR	Total cholesterol	19060911
rs11820589	cg12556569	<i>APOA5</i>	Promoter	Metabolic syndrome	21386085
rs11823543	cg12556569	<i>APOA5</i>	Promoter	Triglycerides-blood pressure	21386085
rs12286037	cg12556569	<i>APOA5</i>	Promoter	Metabolic syndrome	21386085
rs12280753	cg12556569	<i>APOA5</i>	Promoter	Cardiovascular disease risk	20838585
rs964184	cg12556569	<i>APOA5</i>	Promoter	HDL cholesterol	20686565
				Hypertiglyceridemia	20657596
				Total cholesterol	20686565
				LCL cholesterol	20686565
				Coronary heart disease	21378990
Neurological disorders					
rs947211	cg06442372	<i>RAB7L1</i>	Promoter	Parkinson's disease	19915576
rs2395163	cg18698799	<i>C6orf10</i>	Body	Parkinson's disease	22451204
rs2373115	cg24769381	<i>GAB2</i>	Body	Alzheimer's disease	17553421
rs987870	cg09900440	<i>HLA-DPA1</i>	3'-UTR	Asthma	21814517
Autoimmune disorders					
rs987870	cg09900440	<i>HLA-DPA1</i>	3'-UTR	Systemic sclerosis	21779181
rs744166	cg11144103	<i>PTRF</i>	Body	Multiple sclerosis	20159113
rs6496667	cg22142142	<i>GABARAPL3</i>	Body	Rheumatoid arthritis	22446963
rs3117242	cg23333490	<i>HLA-DPB1</i>	Body	Antineutrophil cytoplasmic antibody-associated vasculitis	22808956
rs1610677	cg27230769	<i>LOC285830</i>	Body	Rheumatoid arthritis	21653640
rs2523393	cg27230769	<i>LOC285830</i>	Body	Multiple sclerosis	19525953
Cancer incidence					
rs2523395	cg27230769	<i>LOC285830</i>	Body	Prostate cancer	22219177
Other traits					
rs10903129	cg06961873	<i>TMEM57</i>	3'-UTR	Erythrocyte sedimentation rate	19060911
rs732505	cg24413826	<i>SAFB2</i>	3'-UTR	vWF and FVIII levels	21810271
rs11739663	cg15402732	<i>CEP72</i>	Body	Ulcerative colitis	23128233

APOA5, apolipoprotein A-V; C6orf10, chromosome 6 open reading frame 10; CEP72, centrosomal protein 72 kDa; GAB2, GRB2-associated binding protein 2; GABARAPL3, GABA(A) receptors associated protein like 3, pseudogene; HLA-DPA1 - major histocompatibility complex, class II, DP alpha 1; HLA-DPB1, major histocompatibility complex, class II, DP beta 1; HSPB7, heat shock 27-kDa protein family, member 7; PTRF, polymerase I and transcript release factor; RAB7L1-RAB7, member RAS oncogene family-like 1; SAFB2, scaffold attachment factor B2; and TMEM57, transmembrane protein 57.

coronary artery disease (Cullen 2000). All five SNPs showed a higher risk allele frequency in the YRI samples, which had greater levels of cytosine modification of the associated CpG (Figure 4). These results suggest that mQTL data generated from the LCL model could be used to functionally annotate complex trait-associated SNPs identified in previous GWAS, thus enhancing our understanding of the biological implications of these associations.

Correlation between cytosine modification and gene expression

We evaluated the correlation of cytosine modification levels with gene expression levels measured using the exon array (Zhang *et al.* 2008a), across 115 CEU and YRI samples. We examined all CpG-gene pairs for 208,808 CpG sites that were mapped to the gene regions of 14,669 genes. At 5% FDR, 5328 CpG-gene correlated pairs were detected for 13% of the analyzed genes. Positive correlations (3526) were enriched within the regions from the second gene body quantile to the 3'-UTR, while negative correlations (1802) peaked in the regions from 500 bp upstream of TSS to the first gene body quantile (Figure 5A). Overall, the gene

body/UTRs had a greater proportion of correlated CpG sites (3.1%) than upstream (0.87%) and downstream (0.65%) flanking regions.

Cytosine modification levels at 561 differential CpG sites between the CEU and YRI samples correlated with expression levels of 317 differential genes (10% of 3140 differentially expressed genes) (Table S7). Technical validation of the cytosine modification and gene expression levels for four such genes (*LIPH*, *TBX21*, *DENND2D*, and *CR2*) was performed by bisulfite sequencing and qRT-PCR (Figure S10). We found that in some cases adjacent CpGs captured by bisulfite sequencing were even more differentially modified between populations than the CpG measured by the 450K array (Figure S11). Examples of CpG-gene correlations were shown for gene *PLA2G4C* (encoding phospholipase A2 group IV C), for which CEU showed greater levels of modification for a CpG in the promoter region (Figure 5B) and lower levels of modification for a CpG in the gene body (Figure 5C). These modification patterns correlated with lower levels of *PLA2G4C* expression in the CEU samples.

To assess the genetic basis of CpG-gene correlations, we sought to detect modification-expression quantitative trait

Figure 4 SNPs associated with complex traits with known racial disparities are annotated by mQTL. (A) Five SNPs associated with cardiovascular traits such as cholesterol levels and cardiovascular diseases are mQTL for a CpG in the promoter of *APOA5* (encoding apolipoprotein A-V). The risk alleles for each trait have a higher frequency in the YRI samples compared with the CEU samples. The risk allele is denoted above each set of bars for each SNP. (B) Cytosine modification levels of the CpG in the *APOA5* promoter region are greater in the YRI samples.

loci (m-eQTL), which are SNPs contributing to both cytosine modification and gene expression variations. We focused on the CpG-gene correlated pairs for which modification and expression differed between populations. Across 115 CEU and YRI samples, we tested local SNP associations for the 561 differential CpG sites and 317 paired differential genes. Modification or expression levels were regressed on SNP allelic dosage, with population identity and gender as covariates. At 5% FDR, we detected 440 m-eQTL that were associated with 54 CpG sites and 24 genes (Table S8). For example, the T allele of m-eQTL rs10779587 was associated with a lower expression level of *FLVCR1* (encoding feline leukemia virus subgroup C cellular receptor 1) (Figure 5D) and greater cytosine modification level of cg01313622 (Figure 5E). The frequency of this T allele was higher in the CEU samples, consistent with the greater modification level of cg01313622 and lower *FLVCR1* expression in the CEU samples.

Discussion

In this study we detected abundant cytosine modification differences between the CEU and YRI samples. More than half of these differential CpG sites could be explained primarily by genetic variation through local mQTL, while population specificity of mQTL effects was also common. We found distinct correlations between gene expression and cytosine modification levels in proximal regions and gene bodies. Furthermore, mQTL underlying population-specific cytosine modifications can be colocalized with eQTL and GWAS SNPs for traits with racial disparities.

Based on several considerations, we concluded that the cytosine modification differences observed between the CEU and YRI samples in this study largely reflected true population variations rather than LCL cell culture or other experimental artifacts: (1) Our samples were randomized by population identity across batches, which eliminated confounding due to experimental batch effects. (2) The intrinsic growth rate and EBV copy number were unlikely to bias our results. (3) The consistency between this study with previous work using the Illumina 27K array on population-specific CpGs (Fraser *et al.* 2012) and mQTL associations

(Table S9) (Bell *et al.* 2011; Fraser *et al.* 2012) demonstrated the robustness of these results across experiments and profiling platforms. (4) PCA of our cytosine modification data and a Dutch whole blood dataset (Horvath *et al.* 2012) demonstrated that cell type or cell line culturing time were not the primary variables. (5) Approximately 54% of the differential CpG sites could be explained primarily by local genetic variation, which was unlikely to be confounded by any biases due to cell culture. (6) mQTL of differential CpG sites underlie several complex traits with known racial disparities between European and African descents.

Considering the overall similar genetic background but different environmental exposures of the two populations, we would expect a substantial proportion of differential modifications attributable to environmental effects, if such effects have been stably maintained in cell cultures. For example, it is unclear whether environment could partially explain the distinct genic enrichment patterns between CpG sites with greater modification levels in the CEU and those greater in the YRI samples. Since the variability of cytosine modifications was comparable between the two populations, our observed population specificity of mQTL effects suggested that genetic epistasis and/or genotype $\times$ environment interactions could be common. This is consistent with the findings from a recent study (Fraser *et al.* 2012). To dissect the genetic, environment, and interaction effects contributing to the population differences in cytosine modifications, carefully designed experiments that provide population and environment contrasts will be needed.

Transcriptional abundance has been related to cytosine modification in promoter regions and in gene bodies (Ball *et al.* 2009; Rauch *et al.* 2009; Bell *et al.* 2011). Our study further revealed interindividual correlations of gene expression with cytosine modification levels within these regions. A similar trend has been observed in *Arabidopsis* (Zhang *et al.* 2008b). Cytosine modification within proximal regions could interfere with the transcription initiation (Comb and Goodman 1990), which was also suggested by the very low modification levels in these regions. However, the interindividual variability of modification levels in these regions was relatively high, suggesting their importance in gene regulation (Figure 1B). Genic cytosine modifications were thought

Figure 5 Correlation between cytosine modification and gene expression levels. (A) The proportion of CpG sites for which modification levels correlated with gene expression levels is plotted against the positional bin across the gene region. The gray, orange, and black lines denote all significant, positive, and negative correlations, respectively. (B) The expression level of *PLA2G4C* (encoding phospholipase A2, group IVC) negatively correlates with the cytosine modification level of cg27270541 located in the promoter region, but (C) positively correlates with cg02380983 located in the gene body. The expression level of *PLA2G4C* is greater in the YRI (orange) relative to the CEU (black) samples. The positions of the CpG sites are denoted by an orange cross. (D) The T allele of m-eQTL rs10779587 is associated with lower expression level of *FLVCR1* (encoding feline leukemia virus subgroup C cellular receptor 1) and (E) greater cytosine modification level of cg01313622 located in the proximal regulatory region of *FLVCR1*, across the CEU (black) and YRI (orange) samples. The positions of CpG sites and m-eQTL are denoted by an orange cross and a blue triangle, respectively.

to prevent transcription from intragenic cryptic promoters (Zilberman *et al.* 2007). The greater levels of cytosine modification and the stronger co-modification within gene bodies further supported this model. Despite the relatively low variability of cytosine modification levels at individual CpG sites in gene bodies (Figure 1B), the cumulative effects of these modifications in a gene could be strong.

Investigation of the population differences in cytosine modifications may help elucidate molecular mechanisms of known health disparities. For example, genetic variants of *TBX21* have been implicated in childhood asthma, which disproportionately affects African American and Latino children compared with their Caucasian counterparts (Szefer 2010). We found that this gene had different cytosine modification and gene expression levels between the CEU and YRI samples, suggesting that epigenetic variation could be an additional regulatory component for *TBX21*. As demonstrated in this study (Table 1), the mQTL data could provide novel functional annotations to the SNPs previously discovered in GWAS. Considering the lack of data on minority populations in publicly available databases (Hindorff *et al.* 2009), the HapMap LCLs, which were derived from major global populations, would provide a useful resource for annotating the genomes of individuals of non-European descent.

Finally, it is important to note that because the current profiling platforms including the Illumina 450K and 27K arrays require bisulfite-treated DNA, they cannot distinguish between 5-mC and 5-hmC. A recent study demonstrated that the proportion of 5-hmC to 5-mC along the gene bodies was most predictive of gene expression levels in neurons (Mellen *et al.* 2012). Thus, future work utilizing base-pair resolution techniques for these two types of modifications will be needed to define the extent to which 5-mC vs. 5-hmC are involved in population-specific cytosine modifications.

Acknowledgments

The authors thank Pieter Faber and Jaejung Kim from the Genomics Core of the University of Chicago Comprehensive Cancer Center for performing the 450K array profiling assay; Nancy Cox for valuable discussions; and Won Huh for the assistance in RT-PCR. This work was supported by grants from the National Institutes of Health R21HG006367 (to W.Z., L.A.G., M.E.D.), R01CA136765 (to M.E.D.) and U01GM061393 (to M.E.D.).

Literature Cited

- Anderson, M. J., and J. Robinson, 2002 Permutation tests for linear models. *Australian and New Zealand Journal of Statistics* 43: 75–88.
- Ball, M. P., J. B. Li, Y. Gao, J. H. Lee, E. M. LeProust *et al.*, 2009 Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* 27: 361–368.
- Bell, J. T., A. A. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi *et al.*, 2011 DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* 12: R10.
- Bibikova, M., B. Barnes, C. Tsan, V. Ho, B. Klotzle *et al.*, 2011 High density DNA methylation array with single CpG site resolution. *Genomics* 98: 288–295.
- Bird, A., 2002 DNA methylation patterns and epigenetic memory. *Genes Dev.* 16: 6–21.
- Bock, C., J. Walter, M. Paulsen, and T. Lengauer, 2008 Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Res.* 36: e55.

- Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed, 2003 A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19: 185–193.
- Borevitz, J. O., D. Liang, D. Plouffe, H. S. Chang, T. Zhu *et al.*, 2003 Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* 13: 513–523.
- Comb, M., and H. M. Goodman, 1990 CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. *Nucleic Acids Res.* 18: 3975–3982.
- Cookson, W., L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, 2009 Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10: 184–194.
- Cullen, P., 2000 Evidence that triglycerides are an independent coronary heart disease risk factor. *Am. J. Cardiol.* 86: 943–949.
- Du, P., X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe *et al.*, 2010 Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11: 587.
- Dunham, I., A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis *et al.*, 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.
- Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink *et al.*, 2008 Genetics of gene expression and its effect on disease. *Nature* 452: 423–428.
- Fernandez, A. F., Y. Assenov, J. I. Martin-Subero, B. Balint, R. Siebert *et al.*, 2012 A DNA methylation fingerprint of 1628 human samples. *Genome Res.* 22: 407–419.
- Flanagan, J. M., V. Popendikyte, N. Pozdniakovaite, M. Sobolev, A. Assadzadeh *et al.*, 2006 Intra- and interindividual epigenetic variation in human germ cells. *Am. J. Hum. Genet.* 79: 67–84.
- Fraser, H. B., L. L. Lam, S. M. Neumann, and M. S. Kobor, 2012 Population-specificity of human DNA methylation. *Genome Biol.* 13: R8.
- Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861.
- Grewal, S. I., and D. Moazed, 2003 Heterochromatin and epigenetic control of gene expression. *Science* 301: 798–802.
- HapMap, 2003 The International HapMap Project. *Nature* 426: 789–796.
- HapMap, 2005 A haplotype map of the human genome. *Nature* 437: 1299–1320.
- Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta *et al.*, 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* 106: 9362–9367.
- Horvath, S., Y. Zhang, P. Langfelder, R. S. Kahn, M. P. Boks *et al.*, 2012 Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* 13: R97.
- Huang, R. S., E. R. Gamazon, D. Ziliak, Y. Wen, H. K. Im *et al.*, 2011 Population differences in microRNA expression and biological implications. *RNA Biol.* 8: 692–701.
- Huang, Y., W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu *et al.*, 2010 The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* 5: e8888.
- Im, H. K., E. R. Gamazon, A. L. Stark, R. S. Huang, N. J. Cox *et al.*, 2012 Mixed effects modeling of proliferation rates in cell-based models: consequence for pharmacogenomics and cancer. *PLoS Genet.* 8: e1002525.
- Ito, S., L. Shen, Q. Dai, S. C. Wu, L. B. Collins *et al.*, 2011 Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* 333: 1300–1303.
- Jin, S. G., S. Kadam, and G. P. Pfeifer, 2010 Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res.* 38: e125.
- Johnson, W. E., C. Li, and A. Rabinovic, 2007 Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118–127.
- Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9: 357–359.
- Mellen, M., P. Ayata, S. Dewell, S. Kriaucionis, and N. Heintz, 2012 MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* 151: 1417–1430.
- Rakyan, V. K., T. A. Down, N. P. Thorne, P. Flicek, E. Kulesha *et al.*, 2008 An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.* 18: 1518–1529.
- Rauch, T. A., X. Wu, X. Zhong, A. D. Riggs, and G. P. Pfeifer, 2009 A human B cell methylome at 100-base pair resolution. *Proc. Natl. Acad. Sci. USA* 106: 671–678.
- Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry *et al.*, 2006 Global variation in copy number in the human genome. *Nature* 444: 444–454.
- Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards *et al.*, 2005 An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37: 710–717.
- Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan *et al.*, 2001 dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29: 308–311.
- Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley *et al.*, 2007 Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315: 848–853.
- Szefer, S. J., 2010 Advances in pediatric asthma in 2009: gaining control of childhood asthma. *J. Allergy Clin. Immunol.* 125: 69–78.
- Tahiliani, M., K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala *et al.*, 2009 Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 324: 930–935.
- Westra, H. J., R. C. Jansen, R. S. Fehrmann, G. J. te Meerman, D. van Heel *et al.*, 2011 MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* 27: 2104–2111.
- Zhang, W., S. Duan, E. O. Kistner, W. K. Bleibel, R. S. Huang *et al.*, 2008a Evaluation of genetic variation contributing to differences in gene expression between populations. *Am. J. Hum. Genet.* 82: 631–640.
- Zhang, X., S. H. Shiu, A. Cal, and J. O. Borevitz, 2008b Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. *PLoS Genet.* 4: e1000032.
- Zhang, X., W. Mu, and W. Zhang, 2012 On the analysis of the illumina 450k array data: probes ambiguously mapped to the human genome. *Front Genet* 3: 73.
- Zilberman, D., M. Gehring, R. K. Tran, T. Ballinger, and S. Henikoff, 2007 Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.* 39: 61–69.

Communicating editor: D. W. Threadgill

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.151381/-/DC1>

Genome-Wide Variation of Cytosine Modifications Between European and African Populations and the Implications for Complex Traits

**Erika L. Moen, Xu Zhang, Wenbo Mu, Shannon M. Delaney, Claudia Wing, Jennifer McQuade,
Jamie Myers, Lucy A. Godley, M. Eileen Dolan, and Wei Zhang**

Figure S1 The density distribution of cytosine modification levels. The β -values of 290,577 CpG sites on autosomes and sex chromosomes across 144 samples are shown. The β -values were quantile normalized and adjusted for batch effects.

Figure S2 Co-modification of CpG pairs in the YRI samples. Spearman's p values of CpG pairs < 5 kb apart were calculated. CpG pairs in which both cytosines fell in the upstream flanking region, 5' UTR, gene body, 3' UTR, and downstream flanking region of a particular gene were grouped by the distance between cytosines within a pair. Bars represent the medians of the signed p^2 across the CpG pairs for each distance group (in bp). Lines represent the intra-quantile ranges of the signed p^2 .

Figure S3 Prominent variables detected by principal components analysis. The first two principal components for the M-values of the 290,577 CpG sites after quantile normalization are plotted, colored by bisulfite conversion batch (A) and array hybridization batch (B). The array hybridization batch appears to be the prominent variable. (C) The M-values of the 290,577 CpG sites were quantile normalized and adjusted for batch effects using COMBAT. Gender is the prominent variable. (D) The M-values of the 283,540 autosomal CpG sites after quantile normalization and batch correction. Population identity is the prominent variable.

Figure S4 The differential CpG sites between the CEU and YRI samples are unlikely biased by EBV transformation or intrinsic growth rate. For the 36,597 differential CpG sites, their M-values were regressed using a linear model: M-value $\sim$ population + gender + EBV + error, or M-value $\sim$ population + gender + iGR + error. The $-\log_{10}P$ -values for EBV (A) or iGR (B) effect are plotted against the $-\log_{10}P$ -values for the population effect. Among the 36,597 differential cytosines, only 146 CpGs showed EBV effect, and 246 CpGs showed iGR effect more significant than the population effect. iGR: intrinsic growth rate.

Figure S5 The 118 unrelated samples (60 CEU + 58 YRI) and 16,651 CpG sites that overlapped between this study and GSE27146 (FRASER *et al.* 2012) were re-processed by the same procedure. Each dataset was then re-analyzed with the linear model: cytosine modification level $\sim$ population + gender + error. The regression coefficients of the two datasets are shown as scatter plot. The x-axis: data from FRASER *et al.* 2012, y-axis: data from this study.

Figure S6 Principal components analysis on cytosine modification profiles of pooled CEU, YRI and Dutch samples. Upper panel: 133 samples in this study pooled with 88 healthy individuals of Dutch descent on the 27K array (GSE41037, Plate C); Lower panel: samples in this study pooled with 33 healthy individuals of Dutch descent on the 450K array (GSE41169). The cytosine modification levels were quantile normalized and adjusted by batch. The clustering patterns were inconsistent with the difference of tissue types (whole blood vs. LCL) or cell cultures (primary vs. cultured).

(A) and (D): The principal components of all CpG sites, colored by batch (1: Dutch study; 2, 3 and 4: this study).

(B) and (E): The principal components of all CpG sites, colored by population identity (1: YRI; 2:CEU; 3:Dutch).

(C) and (F): The principal components of differential CpG sites between the CEU and YRI samples, colored by population identity (1:YRI; 2:CEU; 3:Dutch).

Figure S7 mQTLs are highly enriched within 100 kb of the target CpG sites. The proportions of significant mQTLs (detected in the 1 mb regions) for the 36,597 differential CpG sites are shown. X-axis represents the distance between mQTLs and the target CpG sites. At 5% FDR, a total of 23,924 modification-SNP associations (1,354 CpGs) were detected in the CEU samples. 17,643 modification-SNP associations (1,918 CpGs) were detected in the YRI samples.

Figure S8 Population specificity of mQTLs. The SNP association r^2 in the CEU are plotted against those in the YRI, for mQTLs detected in either the CEU or YRI samples at 5% FDR and with MAF > 0.1 in each population.

Figure S9 Common mQTLs across the CEU and YRI samples are enriched in SNPs with higher F_{st} values. The sampling distribution of the proportions of random SNPs with F_{st} values >0.05 (left panel), >0.10 (middle panel), or >0.15 (right panel) was obtained by 10,000 random draws of 5,237 SNPs controlled by the MAF distribution of the 5,237 common mQTLs shared between the CEU and YRI samples. The MAF distribution of common mQTLs was based on the CEU samples. Random SNPs and F_{st} values were based on the CEU and YRI data in the dbSNP v135 database. The actual proportion of the SNPs with $F_{st} >$ given cutoff among the 5,237 mQTLs is shown as a solid circle. Common mQTLs refer to the mQTLs obtained from 100 kb local regions without linkage disequilibrium pruning.

Figure S10 Bisulfite sequencing and RT-PCR. Bisulfite sequencing of 4 selected population-specific CpG probes in at least 20 random samples confirms cytosine modification levels measured by the 450K array (upper 8 panels). RT-PCR confirms expression levels measured by the Affymetrix Human Exon array (lower 8 panels). (A) Probe cg19269039, located in the first exon of *DENND2D* (encoding DENN/MADD domain containing 2D). (B) Probe cg06959103, located within the promoter region of the *TBX21* (encoding T-box 21). (C) Probe cg19810536, located in the promoter region of *CR2* (encoding complement component receptor 2). (D) Probe cg02124892, located in the 5'-UTR of *LIPH* (encoding lipase, member H).

LIPH Bisulfite Sequencing

Figure S11 Potential importance of the CpG sites adjacent to the target CpG probes. Bisulfite sequencing allows for the measurement of CpG sites adjacent to the target CpGs interrogated by the 450K array. Bisulfite sequencing of *LIPH* (encoding lipase, member H) reveals that adjacent CpGs could be more differentially modified between populations than the target CpGs by the array. Each row represents a different sample. Each pie chart represents a CpG site, and the proportion of modified cytosines at each locus is quantified by the amount of circle filled in. The average percentage of modified cytosines at each locus is denoted above each column. The CpG measured by the 450K array probe is boxed in red. The CpGs adjacent to the boxed CpG are more variable. Particularly, the third CpG (marked by *) is significantly differentially modified between the CEU and YRI samples at $p < 0.05$ by Student's t -test, suggesting that the CpG outside of the probe on the array could also be informative for CpGs of a given genomic region.

File S1

List of perfectly matching unique probes

File S1 is available for download at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.151381/-/DC1>.

Tables S1-S9

Available for download at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.151381/-/DC1>.

Table S1 Primers and probes used in measuring relative EBV copy numbers.

Table S2 Relative EBV copy numbers of the HapMap samples.

Table S3 Statistics of the 450K array probes.

Table S4 Primers used in bisulfite sequencing.

Table S5 Differentially modified cytosines between the CEU and YRI samples (YRI as reference).

Table S6 mQTL detected at 5% FDR within CEU and/or YRI.

Table S7 Population-specific cytosine modification correlated with differential gene expression between the CEU and YRI samples.

Table S8 m-eQTLs underlying modification-expression correlations, for which both modification and expression levels differed between the CEU and YRI samples.

Table S9 The overlap of mQTL detected in this study with the mQTL reported by previous studies.