Supporting Information

Lam et al. 10.1073/pnas.1121249109

SI Materials and Methods

Sample and Design. Details about the community cohort used here have been published previously. Briefly, it included 94 adults recruited from Vancouver through postings in local media and public transit. Subjects were required to be 25–40 y of age and in good health, which was defined as being free of infection for the past 4 wk before blood draw and without a history of chronic disease. The small blood cell separation cohort consisted of five female volunteers recruited through advertisements in University of British Columbia buildings. This project was approved by the University of British Columbia's Research Ethics Board, and all subjects gave written consent before participating.

Peripheral Blood Mononuclear Cell Preparation and Blood Processing. Blood was drawn into Vacutainer Cell Preparation Tubes (Becton Dickinson) by antecubital venipuncture. The samples were processed immediately to isolate peripheral blood mononuclear cells (PBMCs) via density-gradient centrifugation, which were then lysed and homogenized in QiaShredder Spin Columns (Qiagen). Lysates were frozen at −80 °C until DNA was extracted using AllPrep DNA/RNA kits (Qiagen). As published previously, RNA was analyzed in a subset of 55 subjects by expression microarrays, whereas DNA was processed for DNA methylation measurements as described below. Furthermore, our study design included performing a detailed differential blood cell count using an Advia 70 (Siemens Medical) system at the time of the blood draw, thus enabling us to test whether the relative amount of either monocytes or lymphocytes was correlated with DNA methylation in PBMCs.

For the studies examining differential methylation in PBMC subpopulations, PBMCs were purified as before, followed by immunomagnetic positive selection to capture monocytes bearing CD14 and lymphocytes bearing CD3 (Miltenyi Biotec). Using these methods, we usually obtained purity of greater than 90% when evaluated by flow cytometry. Genomic DNA was subsequently extracted from all cells using DNAEasy Kits (Qiagen) and frozen at −80 °C until DNA methylation analysis.

Genome-Wide DNA Methylation Analysis. Genomic DNA was first bisulfite-converted using the EZ-096 DNA methylation kit (Zymo Research). This procedure basically transforms epigenetic information into DNA sequence-based information that can be interrogated with methods derived from allelic variation measurements. We used the HumanMethylation27 Bead Chip assay (Illumina), which enables the simultaneous quantitative measurements of 27,578 CpG sites in the promoters or first exons of 14,475 well-annotated human genes. For sample processing, we followed the experimental procedures specified in the manufacturer's manual. Each chip contained 12 arrays, and we initially ran samples from 94 subjects and six technical replicates. Our methodology was highly reproducible, because technical replicates derived from the same subject but run on different arrays had a mean correlation of $R^2 = 0.994$. Furthermore, we have previously shown an excellent correlation between CpG methylation values derived from IlluminaHuman DNA methylation arrays (Illumina) and those interrogated by pyrosequencing and traditional bisulfite sequencing (1).

DNA Methylation Array Quality Controls and Data Normalization. We used a multilayered approach to perform rigorous quality assessments of the array data to eliminate unreliable probes, followed by normalization across the different arrays to adjust for technical variability. Briefly, sample quality was assessed by the internal controls included on the array using Illumina GenomeStudio software. First, from the original set of 94 subjects, we removed two entire arrays that had poor bisulfite conversion (bisulfite control signal <4,000) or contained more than 1% bad data points, as defined by a detection P value <0.05 or less than five bead representations per data point. Further probe filtering of the remaining samples ($n = 92$) was done as follows. First, CpG sites on the X and Y chromosomes were removed due to our mixed sex population. Second, CpG sites with more than 5% bad data points were removed. Third, probes that were proposed to cross-hybridize to multiple genomic locations and/or to contain SNPs at the CpG site were eliminated from the analysis (2). Collectively, these procedures resulted in 22,922 CpG sites left for subsequent analysis.

Next, the Bioconductor package lumi was used to input the raw DNA methylation data files from GenomeStudio software into R, which were then processed using the build-in functions of the lumi package according to recommended settings (3). The raw methylation data were color-corrected to adjust for Cy3 vs. Cy5 intensity biases, background-corrected to account for local background signal present on the array, and quantile-normalized in each of the color channels individually to adjust for array-to-array variability. Potential batch effects between different chips were evaluated using clustering, correlation heat maps, and intercorrelations within batches. Using this approach, we did not find any batch effects present in our data.

β-values and M-values were calculated as per published procedures integrated in the lumi package. β-values represent the percentage of methylation calculated by M/(M + U), ranging from 0 to 1, with 0 being completely unmethylated and 1 being completely methylated. M-values represent the $\log_2$ ratio of M/U, with negative values indicating less than 50% methylation and positive values indicating more than 50% methylation. Analysis with hierarchical clustering showed no obvious outliers, leaving a total of 92 samples from our large cohort and 30 samples from a small cohort that passed all quality controls. Further filtering and removal of invariable sites for which all samples had a methylation level below 5% or above 95% resulted in 17,870 CpG loci remaining for subsequent correlation analyses.

Statistical Approach to Account for Interindividual Blood Composition Differences. Blood correction was performed to remove changes in DNA methylation due to the percentage of different blood cell components. This was done for every probe by a multiple linear regression with all the blood information (count and percentage for whole blood cells, neutrophils, lymphocytes, monocytes, basophils, and eosinophils). The residuals of each probe, adjusted to the mean value of the probe, were used as "corrected" data.

Determination of Variable CpG Sites. To examine variation of DNA methylation in our cohort, variability analysis was conducted on either blood-corrected or uncorrected data with 22,922 CpG sites. Invariable sites were not removed from this analysis because data containing those sites will be a better representation of sample variability throughout the entire DNA methylation range. Variability was assessed by computing the SD of each CpG site across all 92 samples. To gain further insight into the biology of DNA methylation variation, CpG loci were categorized using published definitions into low-density CpG (LC) regions, intermediate density CpG (IC) regions, and high-density CpG (HC) regions (4).

Statistical Tests for Correlation Between DNA Methylation and Variables. We used nonparametric statistical tests to determine the relationship between DNA methylation levels at all 17,870 CpG sites with biological, demographic, and psychosocial variables. Specifically, Spearman correlations for continuous variables and Wilcoxon rank sum tests for categorical variables were used to account for the nonnormal distribution of DNA methylation data. To assess the likelihood of type 1 errors, multiple testing adjustments were done using the Bioconductor qvalue package (5). This method was initially developed specifically for analysis of genomic data, such as gene expression arrays, and has been well-validated. It estimates the likelihood of multiple testing errors and calculates the false discovery rate (FDR), which is represented by q values. We reported both high-confidence sites ($q < 5\%$) and medium-confidence sites ($q < 25\%$), analogous to our published classification. Furthermore, we evaluated each correlated site for the magnitude of change in DNA methylation across the variable and separately reported the number of CpG sites with more than 5% change. For each variable, a P value distribution plot and a quantile-quantile plot were generated to inspect visually whether methylation showed statistical significance over the null distribution using the hist, qqnorm and qqline functions in R.

Principal Component Analysis. The covariance matrix for the dataset of $n = 17{,}870$ M-values for 92 individuals was constructed. The i,j element of the covariance matrix was given by

$$C_{i,j} = \frac{1}{(N-1)} \sum_{n=1}^{N} (x_{n,i} - \bar{x_i})(x_{n,j} - \bar{x_j})$$

where $x_{n,i}$ is the M-value for the nth probe in the ith individual and $\bar{x_i}$ is the average M-value for the ith individual. From this, the principal components are calculated by finding the eigenvalues/eigenvectors of the covariance matrix.

To correlate traits with the eigen-probes over the entire population, for quantitative traits (e.g., age, oligodeoxynucleotides, body mass index), we computed the Spearman correlation coefficient between each eigen-probe and the trait, and for binary valued traits, we computed the Pearson correlation coefficient to each probe. We used a 5% significance test corrected for multiple testing that gave a cutoff on significant P values <0.001.

To define subgroups of individuals who have covarying methylation from the eigen-probe profiles, we took all individuals who had a value greater than or less than $1 - \sigma$ variation ($= 0.1$) in the profile. We then asked if this group of individuals showed enrichment for any particular trait compared with all remaining individuals. For traits that were nonnumerical and had a binary assignment of values, we used a hypergeometric test for significance. For traits that had numerical values, we used the Wilcoxon rank sum test for significance between the two populations of individuals.

Correlation Between DNA Methylation and Gene Expression. We have previously published gene expression data from a subset of 55 subjects included in our DNA methylation cohort of 92 subjects, allowing us to test for the correlation between gene expression and DNA methylation in material derived from the very same PBMC sample. We only included CpG loci located in the vicinity or within clearly annotated genes, and we also removed those CpGs whose cognate genes were missing from the expression arrays. These eliminations resulted in a residual of 16,419 CpG sites located in the promoters or first exons of 10,577 genes. Log intensities of the expression data were correlated to nonblood corrected M-values using Spearman correlation, and FDR calculations were performed using the Bioconductor qvalue package as described previously. Enrichment of LC, IC, or HC region loci in highly correlated CpG sites was assessed by performing a hypergeometric test.

SI Results

To test more rigorously whether the differences in DNA methylation identified by the correlation analysis of lymphocyte and monocyte cell type percentages described in the main text (*Results*) were indeed caused by the specific cell types, we collected PBMCs and isolated T cells and monocytes from five subjects not included in our cohort at two different time points over the course of a day. Although individuals were not perfectly correlated between their two matched samples, we found no significant differences in DNA methylation between these samples, suggesting that DNA methylation remains constant during the day, at least for the loci interrogated here. Using ANOVA, we identified 1,208 CpG sites that were significantly different in one of the three cell fractions ($q < 0.05$ and β-average difference >5%) as clearly seen in a cluster heat map (Fig. S3). Given that T cells comprise the majority of cells in the PBMC fraction, it was not surprising that these two fractions clustered closely together and separate from the monocytes. Importantly, the set of 1,208 cell fraction-specific CpG sites included 93.2% (246 of 264) of the CpGs associated with lymphocyte percentage and 90.7% (225 of 248) of the CpGs associated with monocyte percentage in the community cohort analysis. There was a clear distinction for cell type-specific CpGs related to their genomic context, in that LC region loci were overrepresented, with an odds ratio of 4.1711 ($P < 0.00001$), whereas HC region loci were underrepresented, with an odds ratio of 0.1652 ($P < 0.00001$).

1. Fraser HB, Lam LL, Neumann SM, Kobor MS (2012) Population-specificity of human DNA methylation. *Genome Biol* 13:R8.
2. Chen YA, et al. (2011) Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray. *Genomics* 97:214–222.
3. Du P, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11:587.
4. Weber M, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 39:457–466.
5. Storey JD, Dai JY, Leek JT (2007) The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* 8:414–432.

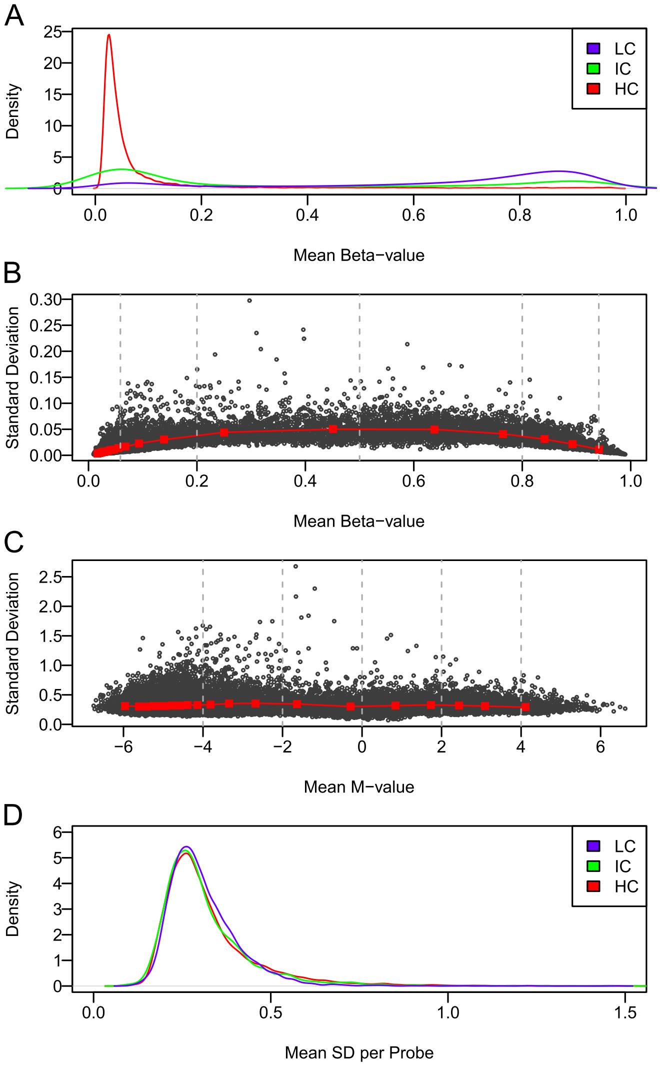

Fig. S1. Assessment of DNA methylation variation using different data representations. (*A*) Mean β-value representation of DNA methylation levels according to CpG density categories. Distinct distribution of mean DNA methylation levels was dependent on the context of CpG site. All CpG sites were classified into LC regions, IC regions, and HC regions. The SD of each CpG locus across the entire methylation range was assessed by β-values (*B*) or after transformation into M-values (*C*). Note the more uniform variability when using M-values. In both cases, the red line indicates an average SD of 20 bins, each created to represent an equal number of loci, and the gray dashed lines indicate the equivalent DNA methylation level between β-values and the corresponding M-values. (*D*) HC region loci were slightly more variable than those of IC and LC regions, with adjusted *P* values of 2.75E-6 and 2.20E-16 for IC and LC region loci, respectively.

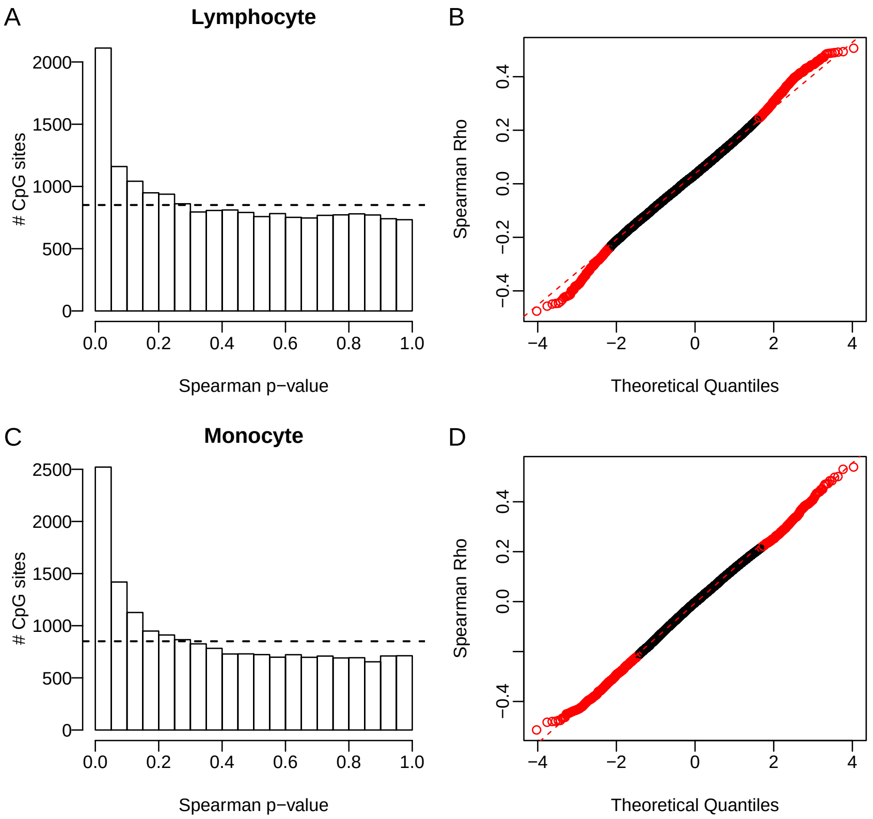

Fig. S2. PBMC DNA methylation was correlated with the relative amount of lymphocytes and monocytes present in leukocytes. Indicative of correlations, graphical representations showed skewing from what is expected by chance. In the *P*-value distributions, the dashed line represents the uniform distribution that was expected by chance, whereas skewed distributions with an enrichment of CpG sites having small *P* values suggest correlation with DNA methylation. The quantile-quantile (Q-Q) plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other. The departure from the reference line (red dashed line) representing a perfect fit of the test statistic with a null distribution provides evidence that the two datasets are from populations with different distributions. *P*-value distributions suggested that lymphocyte percentage was associated with PBMC DNA methylation (*A*), with Q-Q plots further supporting these findings (*B*). The same observation was seen in *P*-value distributions (*C*) and Q-Q plots (*D*) for monocyte percentage. Testing for correlations was done using Spearman ρ statistics. Each circle represents one CpG site. Red circles in the Q-Q plots indicate CpG sites that survive FDR correction at a q value <25%.

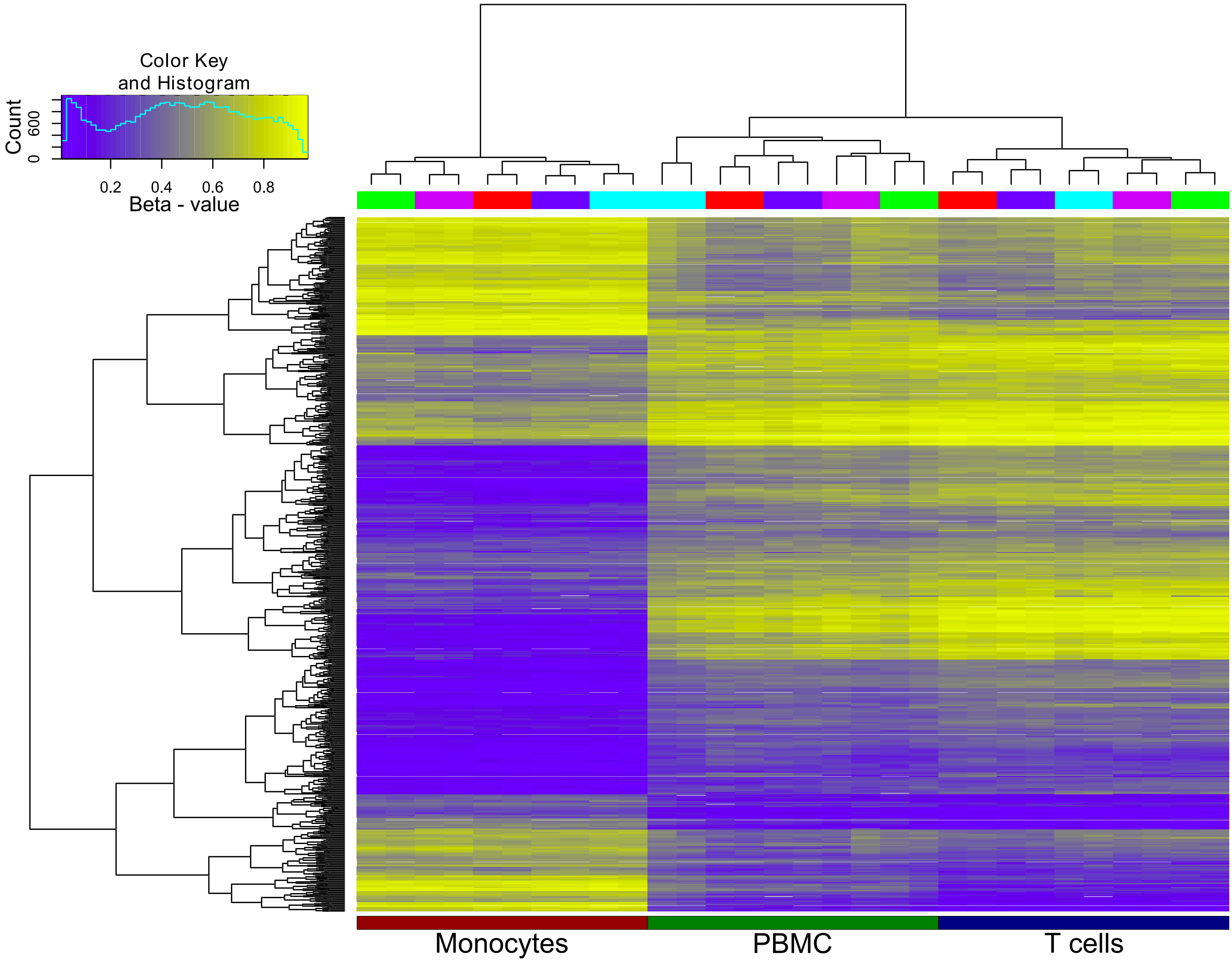

Fig. S3. PBMCs and T cells were most similar, with monocytes being more different. A heat map derived from unsupervised clustering of 1,208 high-confidence CpG sites that are consistently different between the blood subtypes in both samples collected from an individual is shown. After selecting CpG sites having an FDR of a q value <5%, further filtering was performed to include only loci with at least 0.2 differences in β-values between any two cell types. As shown, samples cluster by cell type, with PBMCs and T cells being closer to each other than to monocytes. This was consistent with the majority of cells in PBMCs being lymphocytes. (*Upper*) Color bar indicates each individual. (*Lower*) Color bar indicates the different subtypes.

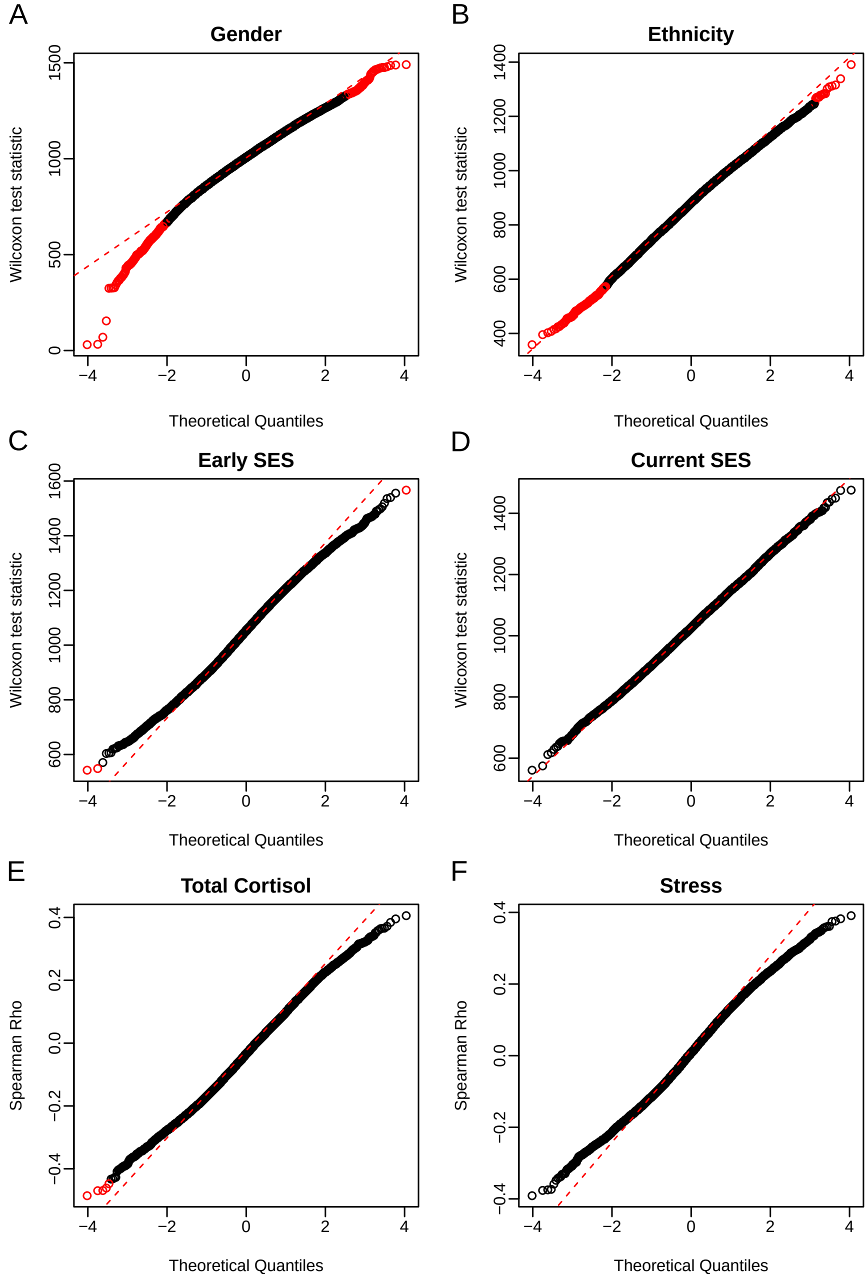

Fig. S4. Demographic and psychosocial factors were associated with DNA methylation. Deviation of quantile-quantile (Q-Q) plot patterns from the reference line (red dashed line) suggested that sex (*A*), ethnicity (*B*), and early-life socioeconomic status (SES) (*C*) but not current SES (*D*) were associated with DNA methylation. Furthermore, cortisol output (*E*) and perceived stress (*F*) were correlated with DNA methylation. Testing for correlations was done using either Wilcoxon tests (*A*–*D*) or Spearman ρ statistics (*E* and *F*). Each circle represents one CpG site. Red circles in the Q-Q plots indicate CpG sites that survive FDR correction at a q value <25%.

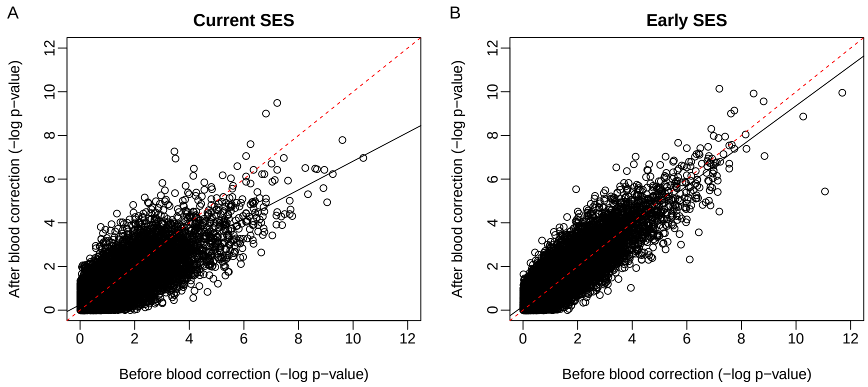

Fig. S5. Correlation of current socioeconomic status (SES) with DNA methylation was strongly affected by correction of blood composition. A comparison of current and early SES *P* values from Wilcoxon tests before and after blood correction, using the regression method developed by us, is shown. Current SES had stronger *P* values before correcting for blood composition (*A*), whereas early-life SES (*B*) was not strongly affected by the correction. The red dashed line indicates how data would look if there was no effect after blood correction. The black solid line indicates the best-fit line describing the relationship before or after blood correction.

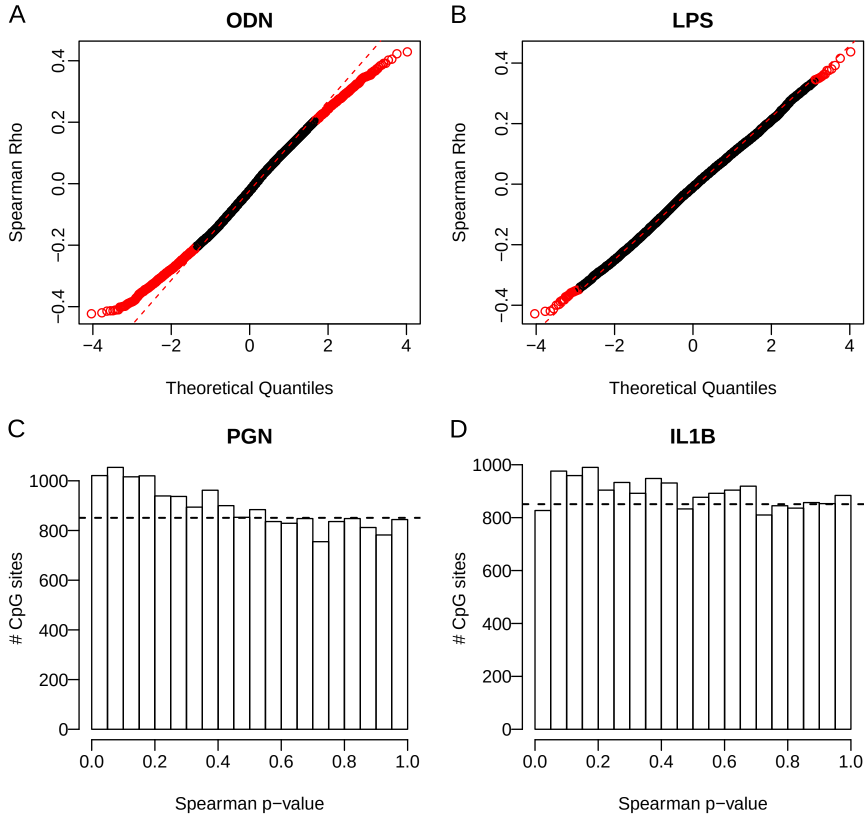

Fig. S6. DNA methylation was predictive of some PBMC ex vivo responses. As judged by the quantile-quantile (Q-Q) plots (*A* and *B*) deviating from the reference line (red dashed line) and skewed *P*-value distributions (*C* and *D*) differing from random (black dashed line), IL-6 production in PBMCs was associated with DNA methylation on ex vivo stimulation for 6 h by oligodeoxynucleotide (*A*), LPS (*B*), peptidoglycan (*C*), and IL-1B (*D*).

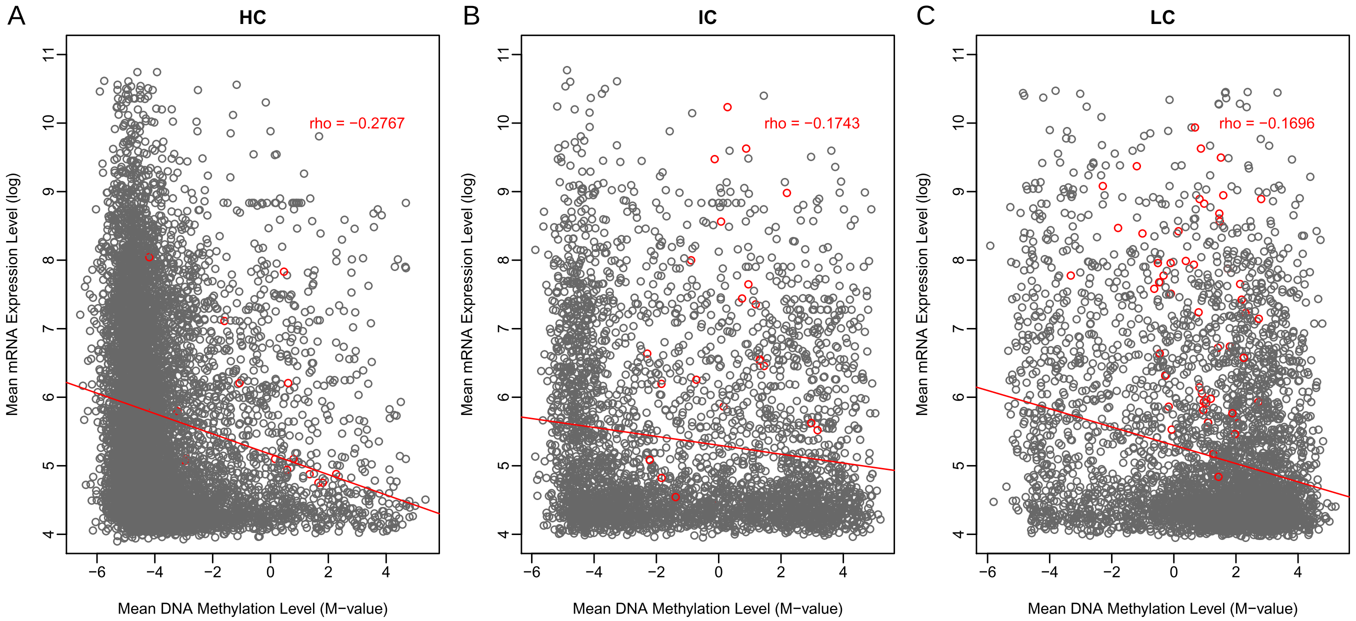

Fig. S7. DNA methylation and mRNA expression levels within an individual were generally negatively correlated regardless of genomic context. Correlation between DNA methylation and mRNA expression levels at HC (*A*), IC (*B*), and LC (*C*) regions shows a negative correlation (red line, best-fit line). Spearman correlation within each category is shown in red. Note that a substantial proportion of genes deviate from the general negative correlation between expression and promoter DNA methylation, however. For example, some lowly methylated promoters were associated with low expression levels and some highly methylated promoters were associated with high expression levels.

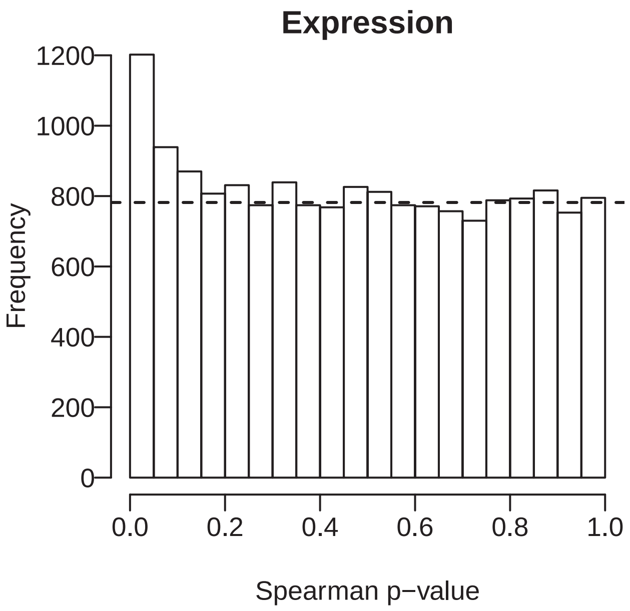

Fig. S8. DNA methylation and gene expression across individuals were weakly associated. The skewed Spearman *P*-value distribution supports a correlation of DNA methylation with mRNA expression levels of associated genes. The dashed line represents the uniform distribution that was expected by chance.

Table S1. Demographic, lifestyle, and psychosocial variables

			Delta > 0.05	
Variable	FDR, q < 5%	FDR, q < 25%	FDR, q < 5%	FDR, q < 25%
Sex	123 (↑107 ↓16)	487 (↑377 ↓110)	21 (↑21 ↓0)	27 (↑26 ↓1)
Age	2 (↑0 ↓2)	15 (↑6 ↓9)	2 (↑0 ↓2)	6 (↑0 ↓6)
Ethnicity	0 (↑0 ↓0)	299 (↑282 ↓17)	0 (↑0 ↓0)	21 (↑18 ↓3)
Early SES	0 (↑0 ↓0)	3 (↑2 ↓1)	0 (↑0 ↓0)	0 (↑0 ↓0)
Current SES	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
Birth control	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
BMI	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
Exercise	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
Stress	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
Depression	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
Sleep	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
PBImw	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
PBIfw	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
dslope	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
Total cortisol	0 (↑0 ↓0)	5 (↑0 ↓5)	0 (↑0 ↓0)	1 (↑0 ↓1)

All tables are organized to indicate the variables tested and the number of CpG loci whose correlation belonged to either the high-confidence group (q < 5%) or medium-confidence group (q < 25%). Furthermore, the number of CpG loci after filtering for methylation differences >0.05 in β-value is indicated, as is the number of sites whose methylation is either increasing or decreasing along the variable or between the groups tested. BMI, body mass index; dslope, daily slope; PBIfw, Parental Bonding Inventory–father's warmth; PBImw, Parental Bonding Inventory–mother's warmth.

Table S2. Ex vivo stimulation of PBMCs by indicated Toll-like receptor ligands

			Delta > 0.05	
Variable	FDR, q < 5%	FDR, q < 25%	FDR, q < 5%	FDR, q < 25%
Pam3CSK4	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
PGN	0 (↑0 ↓0)	2 (↑2 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
PIC	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
LPS	0 (↑0 ↓0)	52 (↑17 ↓35)	0 (↑0 ↓0)	8 (↑1 ↓7)
Flagellin	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
Zymosan	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
ssRNA	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
Imiquimod	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)
ODN	0 (↑0 ↓0)	2408 (↑838 ↓1570)	0 (↑0 ↓0)	364 (↑65 ↓299)
IL-1B	0 (↑0 ↓0)	1 (↑0 ↓1)	0 (↑0 ↓0)	0 (↑0 ↓0)
PMA	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)	0 (↑0 ↓0)

All tables are organized to indicate the variables tested and the number of CpG loci whose correlation belonged to either the high-confidence group (q < 5%) or medium-confidence group (q < 25%). Furthermore, the number of CpG loci after filtering for methylation differences >0.05 in β-value is indicated, as is the number of sites whose methylation is either increasing or decreasing along the variable or between the groups tested. ODN, oligodeoxynucleotide; PIC, Poly I:C; PGN, peptidoglycan; PMA, *phorbol*-12-myristate-13-acetate.

Table S3. Gene expression and DNA methylation

			Delta > 0.05	
Variable	FDR, q < 5%	FDR, q < 25%	FDR, q < 5%	FDR, q < 25%
Expression	97 (↑9 ↓88)	310 (↑72 ↓238)	22 (↑5 ↓17)	118 (↑26 ↓92)

All tables are organized to indicate the variables tested and the number of CpG loci whose correlation belonged to either the high-confidence group (q < 5%) or medium-confidence group (q < 25%). Furthermore, the number of CpG loci after filtering for methylation differences >0.05 in β-value is indicated, as is the number of sites whose methylation is either increasing or decreasing along the variable or between the groups tested.