Landes Bioscience www.landesbioscience.com

Supplemental Material to:

Degui Zhi, Stella Aslibekyan, Marguerite R. Irvin, Steven A. Claas, Ingrid B. Borecki, Jose M. Ordovas, Devin M. Absher and Donna K. Arnett

SNPs located at CpG sites modulate genome-epigenome interaction

Epigenetics 2013; 8(8)
http://dx.doi.org/10.4161/epi.25501

http://www.landesbioscience.com/journals/epigenetics/article/25501

Supplemental materials for

SNPs located at CpG sites modulate genome-epigenome interaction

Degui Zhi,[1] Stella Aslibekyan,[2]* Marguerite R. Irvin,[2] Steven A. Claas,[2] Ingrid B. Borecki,[3] Jose M. Ordovas,[4] Devin M. Absher,[5] Donna K. Arnett[2]

[1]Department of Biostatistics, University of Alabama, Birmingham, AL, 35294, USA

[2]Department of Epidemiology, University of Alabama, Birmingham, AL, 35294, USA

[3]Division of Statistical Genomics, Washington University, St Louis, MO, 63108, USA

[4]Jean Mayer USDA Human Nutrition Research Center on Aging at Tufts University, Boston, MA, 02111, USA

[5]Hudson Alpha Institute for Biotechnology, Huntsville, AL, 35806, USA

We enclosed the following supplementary materials:

- Four supplementary figures (in this file)
- Supplementary methods (in this file)
- One supplementary table (a separate excel file)

Figure S1. Distribution of methylation β scores by probe chemistry both before (A) and after (B) chemistry correction derived on a set of standards.

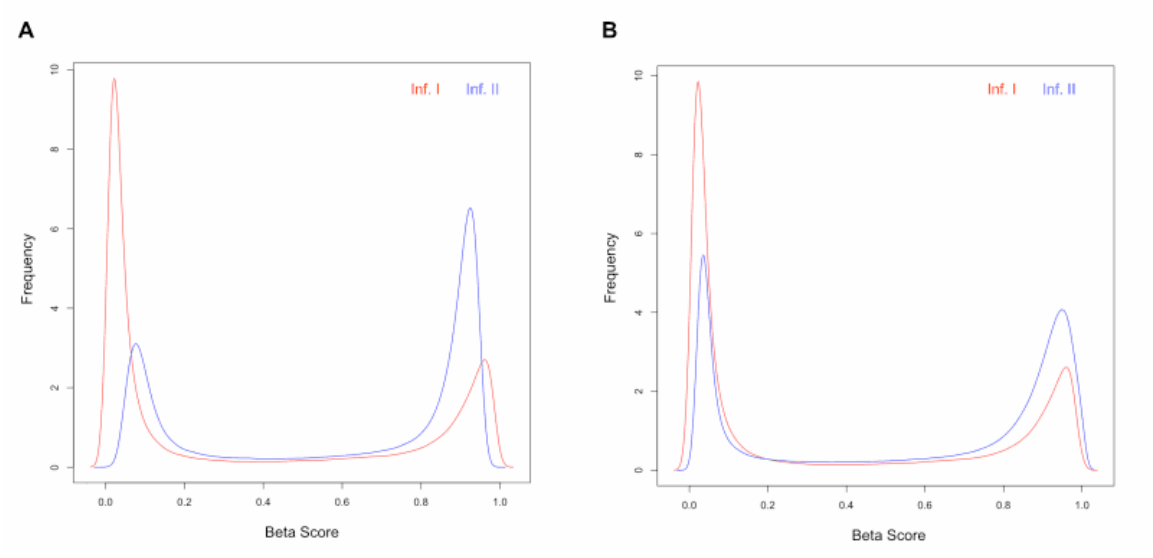


Figure S2. Demonstration of the probe hybridization effect of Infinium II probes of the Illumina M450K array. 50 bp probes from either the reverse or forward direction can be used to hybridize to the interrogated bisulfite treated DNA strand. 1 bp extension reaction, denoted by the asterisk, adds one nucleotide to the probe complement to the C methylated or the U unmethylated nucleotide on the interrogated DNA. A SNP on the hybridizing probe inflates the meQTL signal between the SNP and the interrogated CpG site. Such pairs were excluded from our analysis. However, SNPs on the 1bp extension site are not subject to the probe hybridization effect, and represent the focus of this investigation.

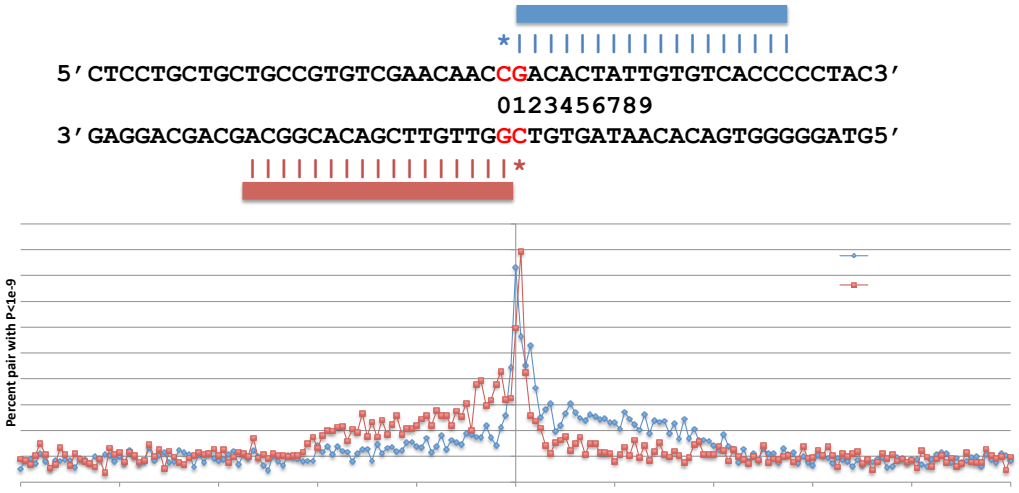
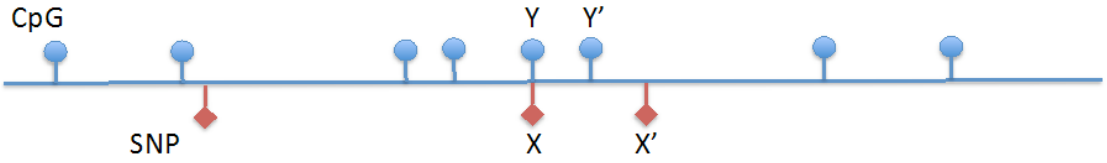


Figure S3. SNP classification by their relation to CpG sites in 1000 Genomes phase 1 (39,698,189 SNPs) and GOLDN (2,529,001 SNPs) data sets.

Figure S4. Diagram of mediation analysis. For a meSNP X-Y site, we regress the methylation status of a nearby CpG site, Y' to the methylation status, Y, at the meSNP site, and X', the genetic marker at site X'. Both X' and Y' are within 10kb of the meSNP site.



Supplementary methods:

Cell harvesting and sample preparation

CD4+ T-cells were harvested from frozen buffy coat samples isolated from peripheral blood using positive selection by antigen-specific magnetic beads (Invitrogen, Carlsbad, CA). Captured cells were lysed and DNA was extracted using DNeasy kits (Qiagen, Venlo, Netherlands).

Methylation450 assays, data QC and batch normalization

500ng of each DNA sample was treated with sodium bisulfite (Zymo EZ DNA) prior to standard Illumina amplification, hybridization, and imaging steps. To limit confounding from batch effects, we distributed SLE cases and controls equally among the 12 slots on each array. The samples were also grouped on the arrays by cell type. The resulting intensity files were analyzed with Illumina's GenomeStudio, which generated beta scores (proportion of total signal from the methylation-specific probe or color channel) and "detection p-values" (probability that the total intensity for a given probe falls within the background signal intensity). Beta scores were generated without background subtraction or Illlumina normalization options. Those beta scores with an associated detection p-value greater than 0.01 were removed and samples with more than 1.5% missing data points across ~470,000 autosomal CpGs were eliminated from further analysis. Furthermore, any CpG probes where more than 10% of samples failed to yield adequate intensity were removed.

The filtered beta scores were then subjected to non-parametric batch normalization with the ComBat package for R software (http:// http://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html). To parallelize this process on our computational cluster, normalization was performed on non-overlapping subsets of no more than 20,000 CpGs per job (randomly selected), and each array of 12 samples was used as a "batch". We also separately normalized probes from the Infinium I and II chemistries, as their beta score distributions are slightly different. For example, the 131,715 autosomal Infinium I CpGs were split into 6 randomly chosen sets of 20,000 CpGs each, plus one set of 11,715 CpGs, and each set was batch normalized in parallel. Linear regression tests for batch ID, with disease, age, gender, and ethnicity as covariates do not produce significant association after ComBat normalization. In addition, we compared our subsetting approach of 20,000 CpGs to similar ComBat runs with larger numbers of CpGs, but the efficacy of batch correction was virtually identical, while greatly reducing the computational time for normalization. Furthermore, our batch normalization process did not introduce any systematic bias into our data, as our disease-specific regression results applied before and after ComBat were highly similar. Data from the X chromosome was normalized separately for males and females due to the gender-specific effect of X-inactivation on the beta score distribution. Finally, we did not adjust for the position on the chip as a batch effect because (1) it is unlikely to confound the relationship between sequence variation and methylation as it is not associated with the genotype and (2) pilot studies demonstrated this effect to be minimal in our data.

After batch normalization, we further adjusted the beta scores for probes that utilized the Infinium II chemistry to better match the Infinium I chemistry using the equation $\beta' =$

$0.001514 + 0.3323 * \beta + 0.7411 * \beta^2$. This equation was derived from fitting a second order polynomial to the observed pairs of beta scores across all pairs of Infinium I-Infinium II probes (~10,000 pairs) located <50bp apart, with and both probes interrogating similar genomic regions. At this proximity, within-chemistry correlations are extremely high (R>0.99) due to locally correlated methylation patterns, and the non-linear relationship between the two chemistries is easily estimated. Our method differs from existing normalization pipelines as it relaxes the assumption that Infinium I and II probes should have a similar distribution, for which we found limited support (Figure S1).

Our dataset was further reduced by eliminating any CpGs where the probe sequence either mapped to a location in the genome that was different that the location found in Illumina's annotation file, or where the probe could potentially map to more than one locus. The list of these problematic CpGs was generated by re-aligning all probes (with unconverted Cs) to the human reference genome with BLAT. We also maintained a list of probes where known SNPs would fall within the probe sequence or at the CpG itself, but did not explicitly filter out these probes. There was no apparent enrichment for CpG probes that overlapped a SNP in dbSNP 135 among our most significant results. Annotations of the M450K chip were obtained from Illumina, with all coordinates sourced from hg19.

Genotyping, data QC, and imputation

All participants had previously been genotyped at 906,600 loci using the Affymetrix Genome-Wide Human SNP Array 6.0, as described in prior publications (references 9 and 10 in the manuscript). SNPs were removed from the analysis if they were

monomorphic, had a call rate below 96%, or exceeded a predefined threshold for the number of families with Mendelian errors, which was defined as follows: for minor allele frequency (MAF) ≥ 20%, removed if errors were present in >3 families (1,486 SNPs); for 20%>MAF≥10%, removed if errors were present in >2 families (1,338 SNPs); for 10%>MAF≥5%, removed if errors were present in >1 family (1,767 SNPs); for MAF<5%, removed if any errors were present (9,592 SNPs). In families with remaining errors, SNPs that exhibited Mendelian error were set to missing (31,595 SNPs). Also, SNPs were removed if they failed the Hardy-Weinberg equilibrium test at P-value $<10^{-6}$, had a minor allele frequency of <1%, were missing strand information, or had discrepancies with the *mlinfo* file. Additionally, 16 GOLDN participants were excluded from analyses because they had call rates of <96%. MACH software (Version 1.0.16) was used to impute non-genotyped SNPs using HapMap Phase II (release 22, Human Genome build 36, hg18) as a reference. We completed the imputation twice by using blocks of 30,000 SNPs, but with a 15,000 SNP shift for the second imputation. We combined the estimated gene dosages for the two imputations by switching between imputing approximately 5,000 SNPs before a 30,000 SNP break point in the first imputation. We then used the imputation results from the 15,000 SNP shift for the next 10,000 SNPs, and then switched back to the original imputation. By doing so, we kept contiguous blocks of imputed SNPs with the highest quality of imputation and removed those segments at block breakpoint after each run. The quality of imputation metrics for each of the 812 meSNPs are summarized in Table S1. UCSC Batch Coordinate Conversion (liftover) module was used to convert genome coordinates from hg18 to hg19.

Effect sizes of meQTL at meSNP as readout from M450K Infinium II chemistry

The probe effect is present for the 50 bp probe sequences that are hybridized to the interrogated DNA. However, it is important to note that a SNP at the C nucleotide of the interrogated CpG site is not subject to the probe hybridization effect for Illumina Infinium II chemistry.

In absence of a mutation, the unmethylated C becomes U by bisulfite conversion and 1 bp extension incorporates a ddATP. Methylated C has a 1 bp extension that incorporates a ddGTP. Illumina chemistry has ddATP and ddTTP with red florescent die and ddCTP and ddGTP with green die, and thus the intensities from different florescent color channel indicates the methylation level at the site. When a nucleotide other than C is present at the site, there are three scenarios:

(i) C-to-T SNP: The result would be equivalent to having an unmethylated C at that position. The reading would record a high intensity on the red channel and result in an effectively zero beta value, consistent with the fact that there is no methylation. In that case, a strong negative correlation in meQTL should be observed.

(ii) C-to-A SNP: the 1 bp extension should add a ddTTP, which has the same fluorescence as ddATP. As a result, we would expect a strong negative correlation in meQTL.

(iii) C-to-G SNP: The 1 bp extension should add a ddCTP nucleotide, which has the same fluorescence as ddGTP. Therefore, it will be read out as a fully methylated C, and beta-score close to 1. As a result, we would expect the methylation measurements to have a strong positive correlation.

All these expected behaviors are indeed observed on our meQTL data (Table 1).