SUPPLEMENTARY FILE 3: DNA METHYLATION DATA ANLAYSIS

In what follows, we first provide a brief description of the data acquisition and preprocessing (batch effect removal, normalization, etc). Later in pages 3 and 4 we present two figures, showing boxplots of methylation values by exposure groups, in two clusters detected in the data analysis. In the paper we provide results of data analyses using two implementations of Aclust. The first figure shows a cluster detected by both implementations; the second figure shows a cluster detected by only one of them.

1. Data acquisition and preprocessing.
500ng of DNA was used to perform bisulfite conversion using the EZ-96 DNA Methylation Kit (Zymo Research, Orange, CA) following Illumina's protocol. Samples and controls for this study were dispersed in two plates, one plate with 24 samples (22 experimental and 2 control samples), and another plate with 60 samples (58 experimental and 2 control samples). The samples were balanced across the plates according to important covariates. Illumina BeadChips were scanned with an iScan and then analyzed by the GenomeStudio software. All experiments were conducted at the Genomics Core Facility of Northwestern University according to the manufacturer's protocol.

The data were normalized twice, once with the control samples, for the quality assessment step, and then time with the control samples excluded, for the main statistical analysis. We used the pipeline proposed by Touleimat and Tost (2012) that filters probes with less than beads, corrects for color bias using the Lumi package (Du at al., 2008) functionality, and matches the distribution of the beta values across the two probe types using subset quantile normalization, separately on each domain defined according to ``relation to CpG island". Probes with SNPs from the EUR population within a distance of 50 base pairs of the interrogated site and with MAF of at least 5% were removed. The SNP list was obtained from 1000 genome (Clarke et al., 2012). The threshold for detection P-value was set at 0.01. The threshold for removing sample was at least 80% of probes have detection P-value higher than this threshold (0.01) and all samples were retained. Consequent analysis used all data, including values with high detection value, and we verified at the end of the analysis that identified sites do not have high detection P-value. The data were corrected for batch effects using the empirical Bayes procedure ComBat (Johnson et al., 2007), for which the beta values were transformed to M-values, and later transformed back to beta values.

We estimated the differences between cell mixture distributions between the two exposure groups using the procedure described in Houseman et al. (2012), and their code and data. The results indicated that the high exposure group had on average 2.07% less CD8T cells in their blood. Using these results, we concluded that the minimal exposure effect size to consider is 0.02 (2%), representing change in proportion methylation.

REFERENCES

Touleimat, N., and Tost, J. (2012). Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, *4*(3), 325-341.

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, *24*(13), 1547-1548.

Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., ... and Flicek, P. (2012). The 1000 Genomes Project: data management and community access. *Nature Methods*, *9*(5), 459-462.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118-127.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., ... and Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, *13*(1), 86.

2. Figures.

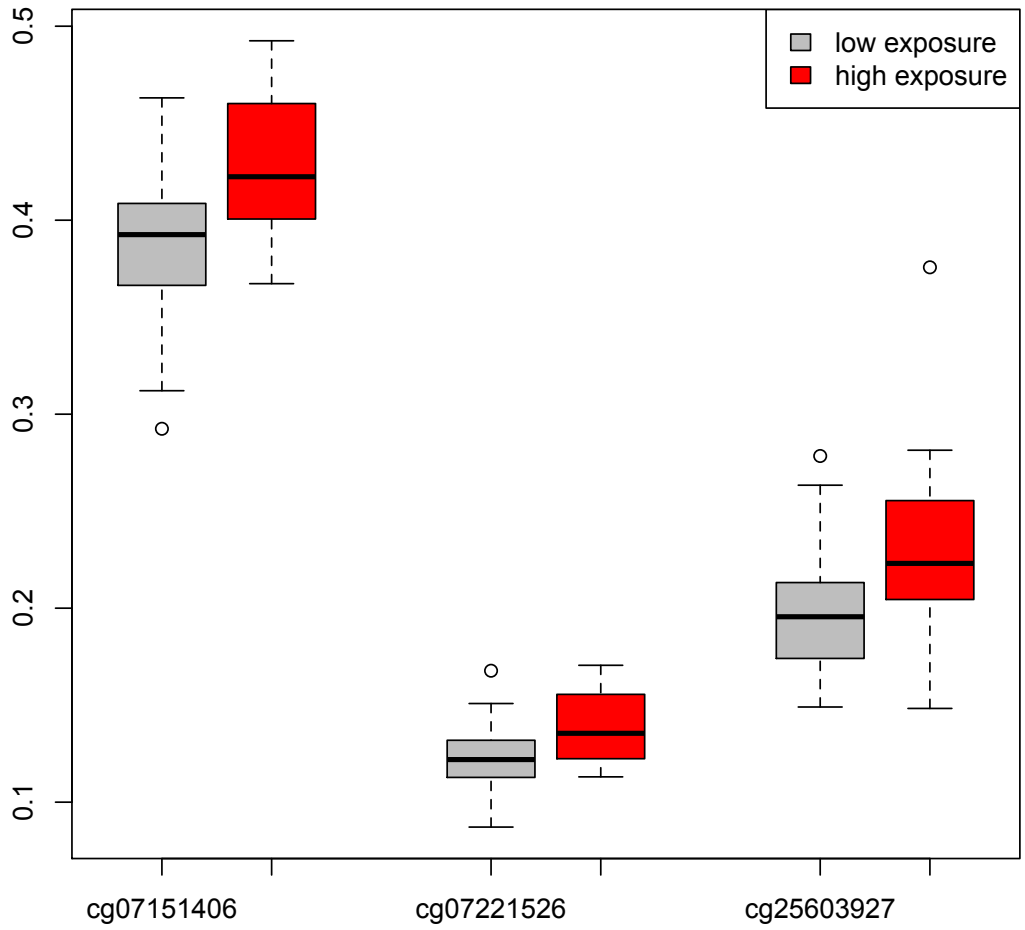

Figure 1: A detected cluster of three probes in a CpG island in the gene PGAM1. This cluster was detected by Aclust applied with distance threshold 0.2, as well as with 0.5 (see manuscript for more details of the two Aclust implementations).

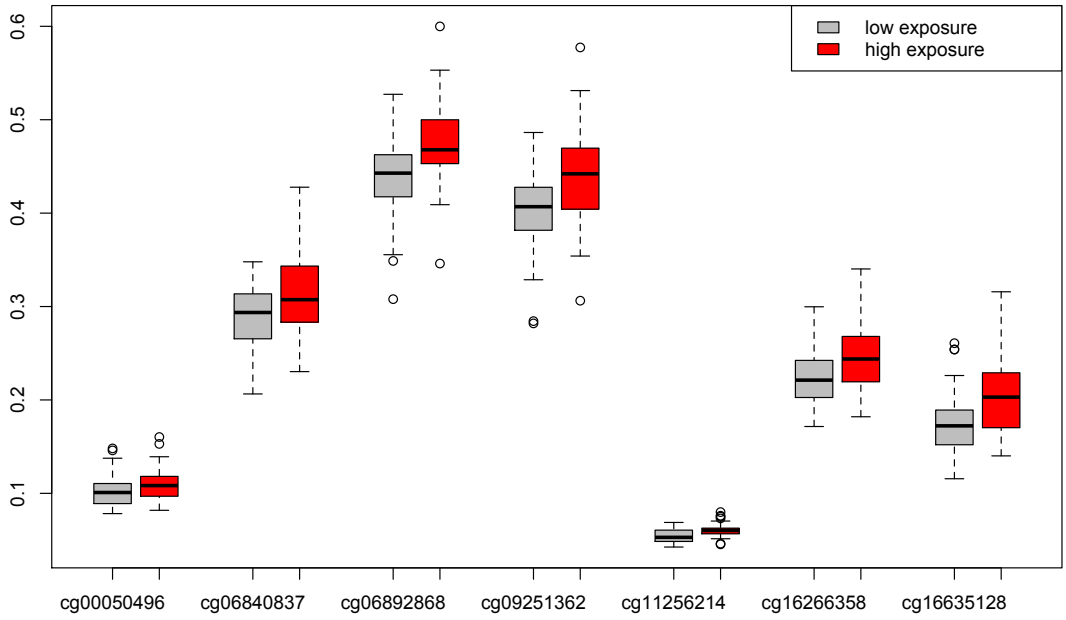

Figure 2: A detected cluster of seven probes in the TSS area of the MGC14436. This cluster was detected by Aclust applied with distance threshold 0.5, but was not detected by Aclust applied with distance threshold 0.2 (see manuscript for more details of the two Aclust implementations).