DNA methylation is the addition of a methyl group to a DNA nucleotide. Aberrant methylation patterns have been associated with a large number of human behaviours and diseases. As part of the Atherosclerosis Risk in Communities (ARIC) study, the Illumina Infinium HumanMethylation450K (HM450) BeadChip was used to measure DNA methylation in peripheral blood obtained from ~3000 African American participants. Over 480,000 cytosine-guanine ( CpG) dinucleotide sites were surveyed. In this project, I calculate short- and long-distance correlation between CpG sites and aim to understand the impact from shared environmental effects (e.g., age, sex, White Blood Cell types etc.).

### Methods

All data analysis was implemented by R and run on the clusters of Minnesota Supercomputing Institute (MSI). The computational challenges of this analysis are from three aspects: limit of memory, computing time, and read/write huge matrices (one correlation matrix is about 1TB).

**In order to utilize memory efficiently,** I used a R package called ff to build a connection between RAM and disk, saved data frames X and X’ into disk and loaded parts of data frame into memory for computation. **In order to reduce calculation time** I divided each data frame into 162 sub-sets, roughly equal sized. Then I calculated the correlation matrices within and between these 162 matrices, which included 13203 (162 x 163/2) correlation matrix calculations. I grouped them into 16 jobs, submitted them to MSI and every job called 24 cores. All these 13203 correlation matrices consist of the up-triangle part of correlation matrix of the data frame and they were saved in disk by the ff package. **In order to read and write these huge matrices conveniently,** I developed R functions to build connections between each other among these 13203 correlation matrices so that I can extract columns, rows and sub-matrix from the huge 483735 by 483735 matrix efficiently.

### Comparison of CpG sites Across Human Genome

I extracted the 99th percentile correlation of each CpG site in the correlation matrix (in step 1) of X, and generated a Manhattan plot in Figure 1 (left), where the 99th correlations are plotted by the location of CpG sites. It demonstrated that high correlations exist between CpG sites, even across different chromosomes. After shared environmental effects was removed by the Linear Mixed-effects models (in the step 4), the 99th percentile correlation of 483735 CpG sites (step 5) is shown in Figure 1 (right). Significant decrease in correlations is observed after shared environmental effects were removed.

### Comparison of Selected CpG sites

I investigated the correlations between a selected CpG site, ‘cg05575921’, and all other CpG sites before and after the shared environmental effect smoking being removed (Figure 2). This site has been previously reported being associated with smoking. After adjustment for smoking effect, most correlations with ‘cg05575921’ are below 0.2.

In addition, I randomly selected three other CpG sites: ‘cg02712183’, ‘00000924’ and ‘cg01629716’. The same analysis was repeated to these sites, with Manhattan plots of correlations in Figure 3,4,5, respectively. Similarly, the correlations significantly decreased after the environmental effects are removed. However, long-distance correlations are still observed for ‘00000924’ and ‘cg01629716’ after the LMM model, suggesting other environmental or biological factors are unaccounted for.