The haplotypes and their counts in all genomic regions assayed are reported. For selection of the marker set that can classify plasma samples, we define a methylated haplotype load (MHL) for each candidate region, which is the normalized fraction of methylated haplotypes at different length:

Where s the length of haplotypes, is the fraction of fully methylated and un-methylated haplotype with i loci. For a haplotype of length L, we considered all the sub-strings with length from 1 to L in this calculation. is the weight for i-locus haplotype. We typically used or to favor the contribution of longer haplotye. After calculating MHL for all candidate regions for all samples, we built a MHL matrix for feature selection, using standard machine learning approaches such as SVM and random forest.

MHL have the ability to select specific genome regions whose methylation status between adjacent CpG sites were high correlated (co-methylation or co-unmethylation).

Since we calculate the sum of mC and uC in the MHL, current MHL do not have the ability to measure the methylation level. Do we need add a variable to show the absolute methylation level, simultaneously?