**Applications Note: Use on EPIC data**

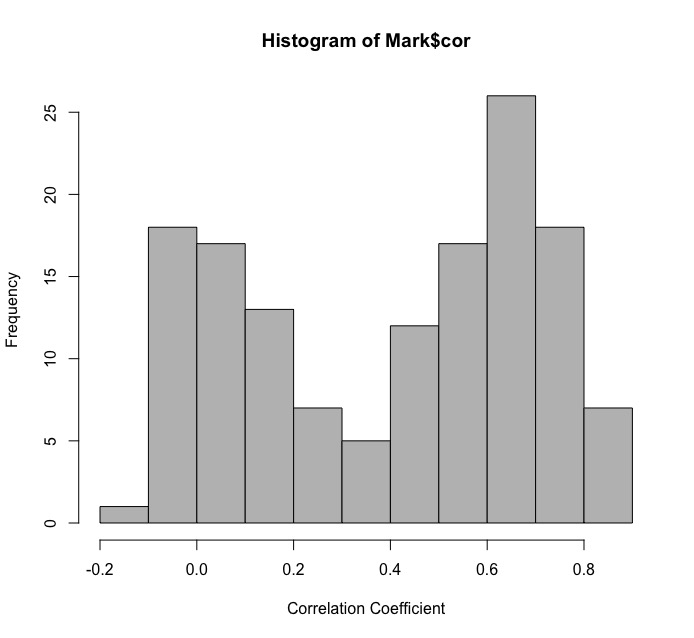


Figure 1: Correlations of the 141 CpG sites present on both the EPIC and 450K arrays, based on correlation coefficients from (1).

The newest array platform released by Illumina, the MethylationEPIC BeadChip (EPIC), assays DNA methylation at >850,000 CpG sites across the genome. More than 90% of CpG sites on the HumanMethylation45O BeadChip (450K) are present on the EPIC array. Although the overall correlation coefficient is >.98 for the two arrays, individual CpG sites can exhibit less correlation (1). Our predictor for DNAm GA uses 148 CpG sites present on the 450K array (2). Of those sites, 141 are present on the EPIC array. Figure 1 shows the correlations between samples of adult whole blood measures on the same person run on both arrays (1), with the mean correlation coefficient being .40.

In an effort to optimize this predictor for use with the EPIC array, we considered 2 possibilities:

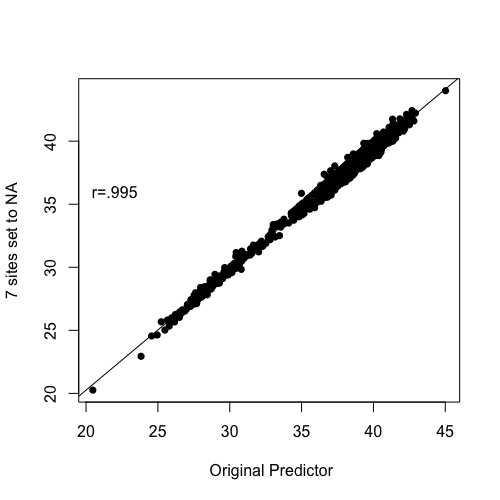
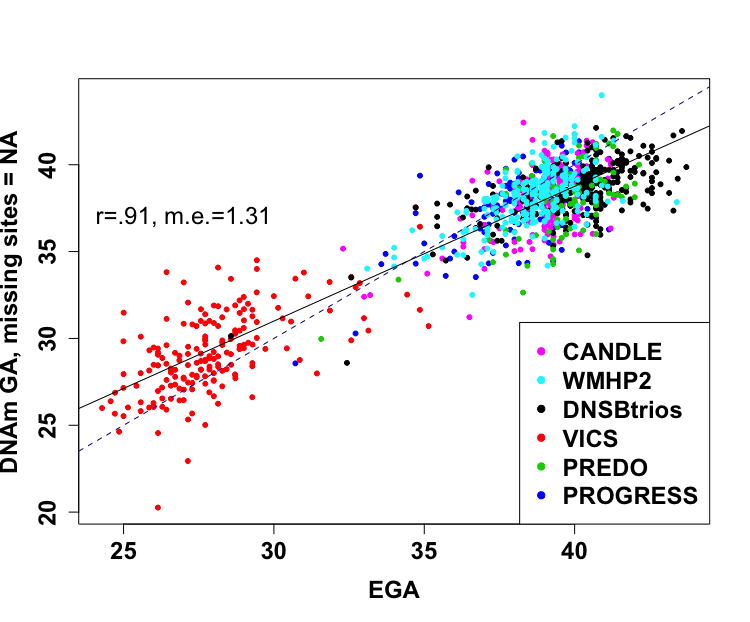
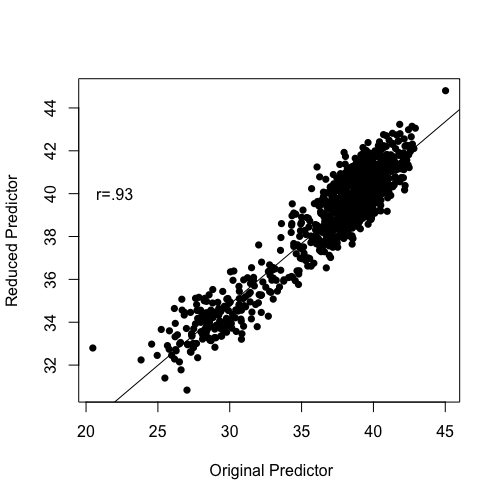
1. The most straightforward solution to apply the predictor to EPIC data is to set missing sites to NA. Under this approach, the predictor performs comparably to the predictor with the full set of CpG sites included (Figure 2). The full predictor has a correlation coefficient of .91 and a median error (m.e.) of 1.24. Estimates from this reduced predictor are also highly correlated with the original predictor (r=.995, Figure 3).

Figure 2: Correlation between EGA and DNAm GA, with the seven sites not present on the EPIC array having been assigned a value of NA.

Figure 3: Correlation between DNAm ages predicted by the original predictor compared to the predictor missing only the sites not present on the EPIC array.

1. We also tested the accuracy of this predictor after removing CpG sites with a correlation coefficient >.50 from the adult blood dataset described above. 69 CpG sites remained in the predictor (47%). Despite dropping a substantial number of CpG sites, the predictor still performed well, with an overall correlation coefficient of .86 (Figure 3). Note, this strategy was implemented by removing the poorly correlated probes from the “Clock” file. Estimates from this reduced predictor are less correlated than the predictor setting 7 CpG sites to NA (r=.93, Figure 4).



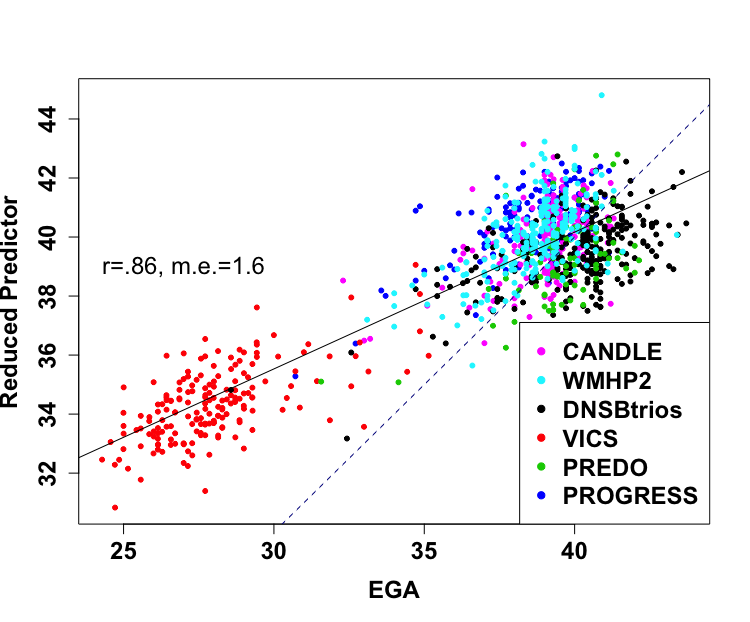


Figure 4: Correlation between DNAm ages predicted by the original predictor compared to the predictor containing only highly correlated sites between the EPIC and 450K arrays.

Figure 3: Correlation between EGA and DNAm GA, with the poorly correlated CpG sites according to Logue removed from the clock file entirely.

**Conclusions:** We recommend supplying a data frame of NAs for the 7 CpG sites missing from the EPIC predictor before estimating DNAm GA. This method is highly correlated with predictions from the original predictor and is a simple modification that yields high-quality data.

Currently, there is one publically available cord blood dataset that used both the 450K and EPIC arrays, GSE86829. We evaluated the association between predicted ages in the five samples available from this study using the method proposed above that sets missing CpG sites to NA on the EPIC array. Unfortunately, estimated gestational age at birth was not available, so we could not evaluate accuracy. However, samples measured on the two arrays do seem to be comparable. The median difference between the same person on the two arrays is .38, or about 2.6 days.

1. Logue MW*, et al.* (2017) The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics*.

2. Knight AK*, et al.* (2016) An epigenetic clock for gestational age at birth based on blood methylation data. *Genome biology* 17(1):206.