Epigenomic Aging

Eric Kramer

1 June 2015

Introduction

DNA methylation is the binding of a methyl group to a guanine on DNA. Multiple studies have documented correlations between methylation in a gene's promoter and the expression of that gene; however, the exact causal relationship between these variables is unknown and likely depends on a variety of factors, such as the presence of eQTLs and the density of CpG sites in the region.

DNA methylation plays a crucial role in tissue differentiation. Studies have tracked increasing methylation as a cell proceeds down the differentiation pathway for innate immune cells, while the first principle component of methylation data will separate tissues, if there are multiple tissues present.

More recent work has shown that environmental factors also affect DNA methylation. Age, socio-economic status and ancestral background can all affect DNA methylation. [1] In general, studies have shown increased methylation with age and low socio-economic status. Nseuronal pathways appear to be the most highly methylated with agining; while stress and immune pathways demonstrate the most increase in methylation due to low socio-economic status.

Hovarth's seminal 2013 paper provided a predictive model of age based on DNA methylation. This model is able to predict a sample's age with astonishing accuracy. More importantly, the model is applicable to a wide-range of tissues. However, we noticed that this model performs worse on samples of African ancestry, likely due to the above-mentioned environmental effects. Because of this, we decided to recalibrate the epigenetic so that is provides improved performance in non-Europeans, while still maintaining performance in diverse tissues.

The following work is an attempt to rebuild the epigenetic clock of Horvath to ensure that the model performs equally well on European, African and Asian samples; while still demonstrating the cross-tissue validity of the original epigenetic clock.

This work proceeeds in four primary steps:

1. Preprocessing of publically available methylation data
2. Creation of predictive model for ethnicity
3. Promoter-level feature selection
4. Recalibrarting the epigenetic clock

[1] Disentangling the effects of ancestry and socio-economic presents problems in many datasets where only one of these variables is measured although they are strongly correlated

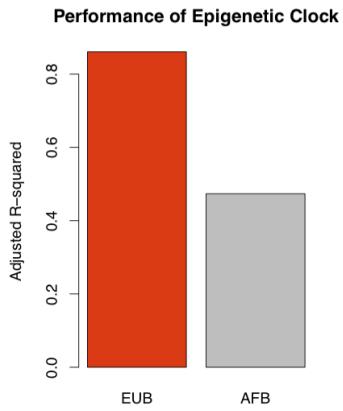

Figure 1: Horvath's epigenetic clock performs significantly worse on Belgians of african ancestry (AFB) than Belgians of European ancestry (EUB)

Preprocessing of publically available methylation data

In order to recalibrate the epigenetic clock so that is stable across ancestries and tissues, we require a training set with samples from multiple ancestries and multiple tissues.

Gene Expression Omnibus data

Gene Expression Omnibus (GEO) contains over 20,000 Illumina 450k DNA methylation arrays. Among these, approximately 7247 samples are annotated with an age. I downloaded and processed raw data for these GEO series and placed these data into a SQLite database (see *./data/BMIQ.db* in the project directory) [2]. These samples range in age from newborns to 105 years old.

[2] Recent work has allowed SQLite databases to be used within R as *data.frames*. See the *dplyr* package for more detail

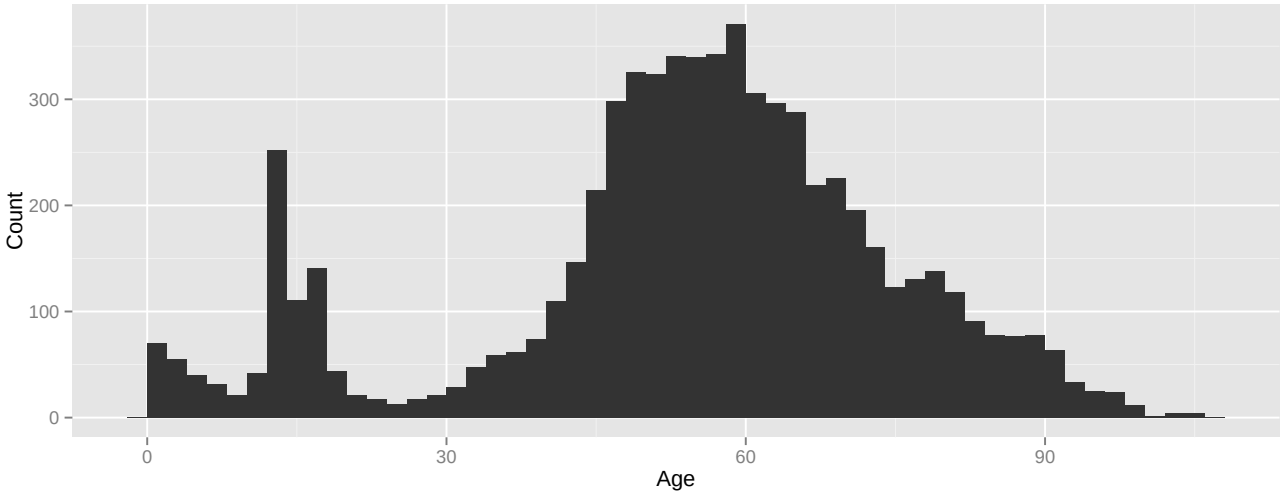

Figure 2: Age distribution of GEO samples assembled for this project

This database contains samples from 18 tissues. A majority of the samples are whole blood or other blood-based cell types (e.g. monocytes). While solid-tissue samples, such as samples from lung or muscle, are comparatively more rare. Because of the highly unbalanced nature of the dataset, careful statistics must be used to ensure that measured effects are not caused by confounding.

Table 1: Number of samples for each tissue

Tissue	Count
Adipose Tissue	174
Bone Marrow	72
Brain	832
Breast	14

Tissue	Count
Buccal	96
Colon	53
Leukocytes	923
Liver	231
Lung	11
Lymphoblasts	322
Lymphocytes	614
Monocytes	1202
Muscle	86
Neuronal	145
Pancreas	221
T-cells	262
Thyroid	82
Whole Blood	1907

Normalization

BMIQ normalization is one of the most common methods to normalize DNA methylation data. It uses a mixture of beta-distributions and quantile normalization to adjust for differences between the probe types on the chip. After normalization, we see that the PCA nicely separates samples based on tissue. The first principle component separates liquid tissues (whole blood, monocytes, T-cells, leukocytes, etc) from the solid tissues (lung, brain, muscle, pancreas, etc). The second principle component is correlated with age and could represent age-related changes in methylation.

If we color the plot according to tissue, rather that tissue group, we see that the two outlying clusters of solid tissues are samples from the cerebellum and the brain, excluding the cerebellum. Additionally, we see sub-structure within each of the clusters. For instance, monocytes in blue and T-cells in purple are clearly separated on the PCA.

Creation of a predictive model for ancestry

Unfortunately, most GEO entires are not annotated with ancestral background. Because of this, we decided to create a predictive model, which is able to impute a sample's ancestry based on it's DNA methylation levels

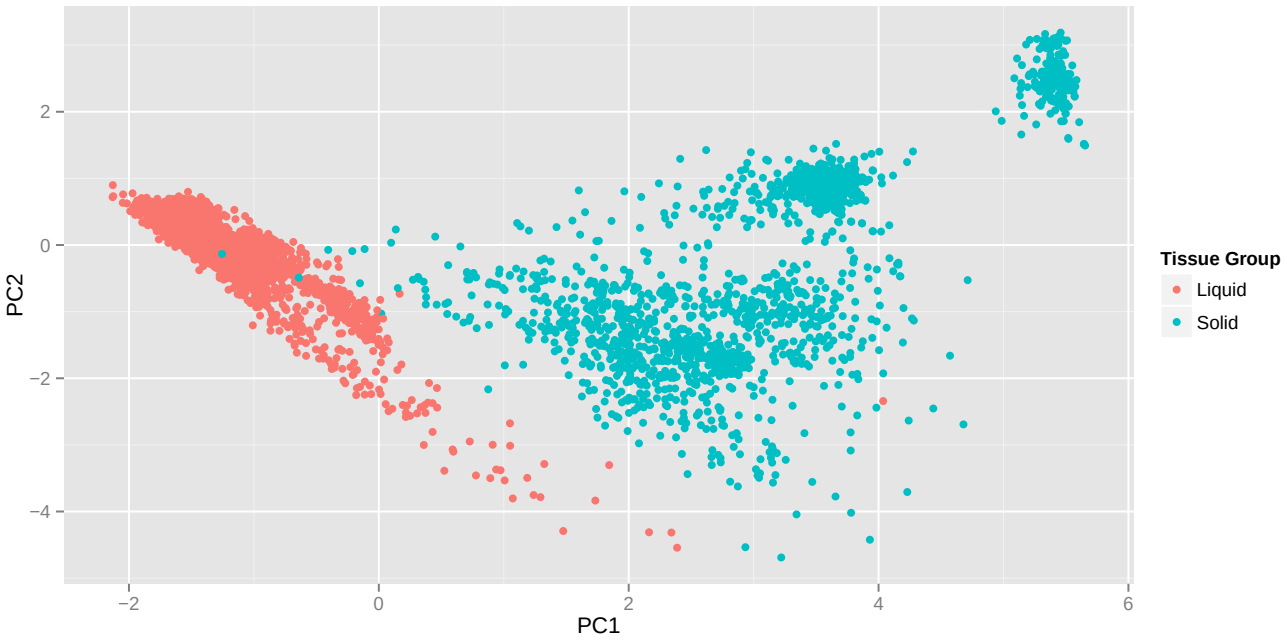

Figure 3: PCA of GEO samples colored according to tissue group

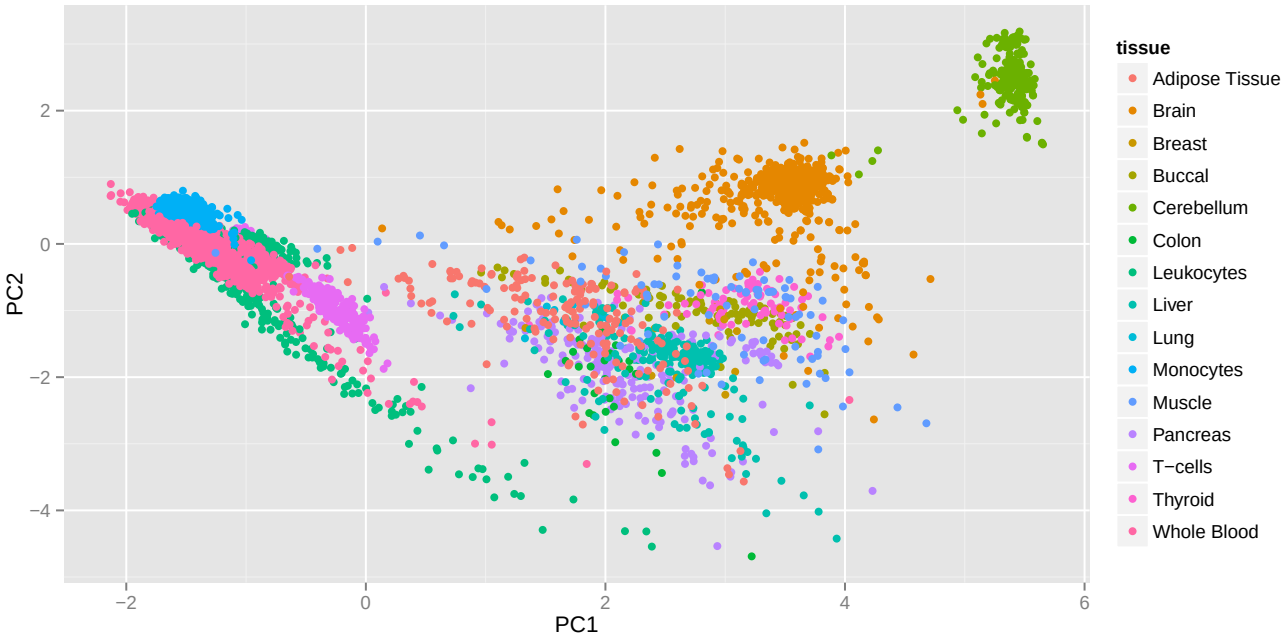

Figure 4: PCA of GEO samples colored according to tissue

Ancestry Informative Methylation Markers (AIMMs)

DNA methylation can only occur at guanines immediately preceeded by a cystine, known as CpG sites. SNPs in a CpG site will necessarily disrupt methylation. Conversely, we can use methylation at sites with known SNPs as a way to infer genotypes.

In order to create a set of ancestry informative methyaltion markers, I mapped common variants from the 1000 genomes project (>5% MAF in one population) onto the CpG sites on the Illumina 450k array. In total, I discovered 14226 methylation sites with a common SNP in the CpG site. [3]

[3] The ids for these sites are stored in *./data/aimms.Rdata* in the project directory.

Training the model

I used t