Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Date

Accounting for Population Stratification in DNA Methylation Studies

By

Richard Thomas Barfield

MPH

Biostatistics

_____

Karen Conneely, Ph.D

Thesis Advisor

Accounting for Population Stratification in DNA Methylation Studies

By

Richard Thomas Barfield

B.A., University of North Florida 2010

Thesis Committee Chair: Karen Conneely, Ph.D

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health in Biostatistics [2012]

Abstract

Accounting for Population Stratification in DNA Methylation Studies

By Richard Thomas Barfield

DNA methylation is an important epigenetic mechanism that helps regulate gene expression and can be influenced by both the environment and the genome. DNA methylation has also been linked to some cancers, complex diseases, and transgenerational effects, and is thus of great interest to public health researchers as a potential link between genome, environment, and disease. In recent years there has been an increase in the number of genome-wide DNA methylation association studies due to a decrease in prices and improved technology. We can now perform DNA methylation association studies at the scale that genome wide association studies (GWAS) were performed a few years back. As with GWAS, problems such as population stratification will also need to be addressed in these DNA methylation studies. Failure to adjust for population stratification in genetic association studies can lead to potential false positives and erroneous results, but population stratification has yet to be accounted for in DNA methylation studies. To address this, we analyzed DNA methylation for association with race in two separate datasets, and identified widespread associations with race across the genome in both cases. We then performed principal components analysis on different forms of the data and included these principal components in the model to determine whether this approach would reduce the number of sites significantly associated with race. We examined principal components computed from data pruned based on correlation and principal components based on CpG sites within a certain distance of a SNP ("informed pruning"). We found that the principal components from the informed pruning performed the best in reducing the number of sites significantly associated with race (90.55- 97.82% reductions in the number of FDR-significant and 84.07-94.38% reductions in the number of Holm-significant sites); this approach was also less computationally intensive than approaches requiring correlation-based pruning. We have therefore developed an effective method to account for population stratification in DNA methylation studies that does not require the collection of data on genetic variants.

Accounting for Population Stratification in DNA Methylation Studies

By

Richard Thomas Barfield

B.A., University of North Florida 2010

Thesis Committee Chair: Karen Conneely, Ph.D

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health in Biostatistics [2012]

Acknowledgements

I would like to thank Dr. John Hanfelt for being a reader for this Thesis, along with being a great academic advisor. Ian Goldlust for providing useful comments. Thanks to Varun Kilaru for providing useful data on CpG sites and SNPs. My family for always being supportive and encouraging in everything I do. Also, thanks to Dr. Conneely for being a great advisor over the past two years.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Epigenetics and DNA Methylation

Epigenetics is defined as any heritable change in gene expression that is caused by mechanisms other than alterations to the DNA sequence. The term was first coined in 1942 by Conrad Waddington to define the processes that lead from the genotype to the phenotype, the “epigenotype” [1]. The epigenome is to epigenetics what the genome is to genetics. Epigenetics is of great interest to public health researchers, as it can be influenced by both the environment and the underlying genotype, and thus provides a bridge between the genome, the environment, and individual phenotype. [2, 3]. Its role as a mediator has led to the hypothesis that epigenetics could explain some of the missing heritability problem, in which genetic variation only accounts for a small fraction of the heritability observed in phenotypes [4, 5]. For example, genes associated with height explain a very small portion of the variability in height [6, 7]. The true determinants of these observed phenotypes may be due to other complex networks not yet understood. Epigenetics may help to explain a portion of the differences that arise naturally in the population. The two most studied epigenetic mechanisms are histone modifications and DNA methylation; this thesis focuses on the latter. The goals of this thesis are to 1) document the extent to which population stratification can impact DNA methylation studies and 2) establish a method to correct for population stratification that could be present in DNA methylation association studies.

DNA methylation typically occurs when a methyl group is added to a cytosine

base pair followed by a guanine on the genome. These cytosine-phosphate-guanine (CpG) sites can be measured by microarrays which give a count of methylated and unmethylated signals for each site [8]. Using the signal information, the proportion of methylation (the beta value) at each site can be calculated. This value is the total methylated signal divided by the total signal. Sites with small beta values are considered hypomethylated, while large values are hypermethylated. DNA methylation's main purpose is to regulate and maintain gene expression [9]. For example, a hypermethylated CpG site in the promoter of a gene can silence the gene, effectively turning it off [2]. The most well-known instance of this is the hypermethylation of one of the X chromosomes in females to prevent over-expression [9]. If a gene is improperly hypomethylated, it can lead to over-expression of the gene product. CpG sites are predominantly found in areas called CpG islands which are typically unmethylated [9]. Near these islands are areas called CpG shores which tend to be highly methylated [10]. It is believed that there used to be more CpG sites throughout the genome, but DNA methylation can increase the probability of DNA repair mechanisms mutating cytosine into thymine, and guanine into adenine [11]. The remaining CpG sites may therefore have some functionality that has prevented them from being selected out. DNA methylation is also involved in differentiating tissue gene expression [12, 13] and the process of imprinting [9, 14].

Advances in technology and rapidly decreasing costs of data generation have led to an increased focus on large-scale studies of DNA methylation among clinical and public health researchers. Through these studies DNA methylation has been linked to various diseases and environmental effects. For example, in carcinogenesis, DNA methylation can turn off certain genes responsible for repressing tumor growth [15, 16].

Further, it was discovered that the genome in tumor cells is generally hypomethylated except in areas for genes controlling tumor growth [10, 17]. In specific cancerous tumors, DNA methylation has also been linked to colon, stomach, prostate and lung cancers [16, 18-22]. Outside of cancer, associations have also been found between DNA methylation and complex diseases such as asthma and lupus [23, 24]. In addition to reported associations with disease, DNA methylation has been linked to environmental stressors in rats and humans [25, 26]. Rats with less nurturing mothers had increased methylation in the promoter region of genes responsible for dealing with stress [25]. This effect reversed when the rats were placed with a more nurturing mother. Another study found that in suicide victims, those with a history of abuse as a child or psychiatric diseases had increased levels of methylation [26]. The above findings support the effect of the environment on epigenetic modifications, with changes to the environment during childhood leading to possible permanent alterations in the epigenome [9].

DNA methylation also has a strong association with age as patterns of methylation change throughout life [27-32]. This can be seen in a monozygotic twin studies where there were more dissimilarities between older pairs as compared to younger pairs or where there was significant association between age and DNA methylation[28, 33]. The twin studies indicated the effect of not only the environment but also of age on DNA methylation. Other studies have observed genome-wide methylation differences with age that are consistent across multiple samples [32, 34-37].

DNA methylation patterns can be influenced by genetic variation [38-43]. Bell et al. identified 180 CpG sites that were associated with single nucleotide polymorphisms (SNPs) within a 5kb range [43], and larger studies are likely to identify many more

associations. A twin study found a similar result with strong association between SNPs and differences in DNA methylation levels around the polymorphism [33]. The relationship between SNPs and CpG sites suggests a possible explanation for individual differences in DNA methylation patterns.

Other studies have examined transgenerational effects and the possibility of epigenetic inheritance. A study in Sweden found that individuals whose grandparents lived through times of drought or famine and had good nutrition during that period in their slow growth period had an increased risk of mortality [44]. Similar findings have been reported in other transgenerational studies of humans [44-48] and mice [49-51]. A possible explanation for this observed effect is epigenetic modifications [48]. The reported transgenerational effects appear to be due to environmental effects which can alter the epigenome. Since surviving a drought does not alter the DNA sequence, these transgenerational effects suggest another mode of inheritance. This is an interesting phenomenon because epigenetic effects are thought to be reset from generation to generation [52]. For example, DNA methylation is reset through a process of demethylation that wipes the genome of the previous methylation patterns in the early stages of gestation (except in areas where imprinting occurs) [9, 51, 53]. After the genome is demethylated, a new pattern of methylation is initiated in the genome that will reset the 'default' pattern of methylation [53]. The exact mechanism which would allow methylation to persist after this reset is unknown. There is known epigenetic inheritance in plants, zebrafish, and yeast while there are few known examples in mammals [54-59].

The argument for the possibility of epigenetic inheritance in humans is based on findings that suggest epigenetic inheritance in mice [56, 57, 60-62]. The histone

modification patterns between mice and humans appear to be well-conserved, which could indicate a shared process yet undiscovered [63]. This however suggests a type of inheritance analogous to the one put forth by Lamarckian evolution, where the passing of phenotypic changes from generation to generation is the main mode of inheritance [64]. This method of evolution though was rebutted by August Weismann and supplanted by Darwin's theory of evolution by natural selection [64]. Another explanation could be that heritable factors are responsible for transgenerational effects [39]. One study found that DNA methylation on the IGF2/H19 locus had a high level of variation between the amount of methylation that is explained by heritable factors, ranging from 20 to 74% in H19 and 57 to 94% in the IGF2 differentially methylated regions [39].

1.2 Population Stratification

While most of the human genome is the same between individuals (99.9%), there are subtle genetic differences between populations. All humans descended from the same region of Africa, but as the species spread out genetic drift and the accumulation of mutations led to genetic differences between groups [65, 66]. Mutations were passed on in some populations through generations by natural selection or genetic drift until two groups have developed separate phenotypes. This phenomenon is known as population stratification. These differences in genomes between populations do not just occur between observably distinct populations, but can also occur in subtler forms within isolated and seemingly homogeneous populations such as Iceland [67]. Population stratification has been well documented over the years [65, 68-70].

The differences between genomes of populations can occur at any genetic marker.

The most studied place where it can occur is at SNPs – places on the genome where a single base pair differ. There are extensive libraries that are designed to reference known SNPs [69, 70]. The vast majority of known SNPs are nonfunctional, with only a small proportion of SNPs actually leading to observable phenotypes.

Population stratification led to potential problems in genetic association studies through inflated risk of false positive associations. One example of this was a study examining diabetes, which reported that Pima Indians had an increased risk for the disease due to certain SNPs [71, 72]. The study found that Pima Indians were more susceptible to Type II diabetes in the GM locus area, where Caucasians had the protective allele [71]. Later studies found that the association was driven by tribe members having various levels of Caucasian ancestry [71]. The study failed to take into account this ancestry [71]. Case-control studies were therefore avoided for a period of time in favor of family studies [72, 73]. These studies were hard to conduct though due to the difficulty of enrolling multiple family members in a study. There are advantages of doing family studies/linkage studies though over association studies. Linkage studies have the advantage of being well-powered to find rare and low frequency variants since these variants would tend to cluster in afflicted families [5]. They can also more easily detect Mendelian diseases and can increase the power in genome wide studies [5]. Association studies however have increased power, are more well-powered to detect common variants, and do not require family members; but they do not perform as well in detecting rare variants [5]. As technology improved, there was a move toward candidate gene studies and case-control studies, where certain genomic markers related to a gene or genes were studied [73, 74]. New methods were developed to account for the population

stratification in the data [73, 75-79]. One method was to genotype Ancestry Informative Markers (AIMs), SNPs previously identified as associated with specific populations, and use these genotypes as a proxy for individual ancestry; the proxy can then be included as a covariate in the analysis [75, 76]. Another popular method was genomic control, which involved multiplying the inflated test statistics (due to population structure) by the median test statistic to correct for the stratification [79-81].

These methods brought a debate about the actual risk of false positives in studies due to population stratification [74, 82-84]. Irreproducibility of certain candidate gene studies was believed to be due to unaccounted population stratification [82, 83]. Other studies reported that the existence of population stratification would not lead to false positives if the study was well defined [74, 84]. However, a paper in 2004 suggested that as sample sizes increased, so would the number of spurious findings due to an increase in statistical power that would that the previous established method of genomic control would be too conservative [85]. This is because these findings are artifacts created by the population structure in the data; they reflect population differences in allele frequency that are not meaningful and have no real function. With the increase in sample sizes, studies will have increased power to detect these allele frequency differences even if those differences are entirely due to population stratification. The sample sizes were starting to reach this size though as technology advanced and there was a move toward genome wide association studies (GWAS). The earlier methods were likely better-suited for small studies, and there was a need for new methods [85, 86].

A paper published in 2006 found a way to address population stratification in large-scale studies by using the principal components of genome-wide genotype data to

summarize the population structure present in the data and serve as proxies for individual ancestry [86]. This method provided a straightforward way to account for population stratification in the data that was also easy to visualize by plotting the components. Researchers could now include the top principal components as covariates in the model to correct for the population structure.

1.3 Population Stratification in DNA Methylation Studies

Microarray data for DNA methylation has become increasingly popular in studies of complex disease. It is now common to perform DNA methylation association studies in the same manner as GWAS studies; such studies led to the findings involving DNA methylation described earlier. Several DNA methylation studies have identified CpG sites where methylation levels differed by race [87-93]. These differences could arise from epigenetic inheritance, population-specific environmental factors, or genetic variation such as SNPs or copy number variants.

As discussed above, the possibility of epigenetic inheritance seems unlikely due to demethylation and the resetting of epigenetic effects that occurs in the germ line. Population differences in methylation could be due to differences in SNP allele frequencies between populations. Several studies have found relationships between SNPs and CpG sites [38-43]. A recent study examining twins found that the pattern of methylation in monozygotic twins was much more similar that that of dizygotic twins [94]. While these patterns may also be due to a shared environment in the womb, the patterns of DNA methylation are probably also determined by the shared genome. Populations will not share an identical genome or environment as monozygotic twins do,

but they still share some environmental and genomic factors. Genetic variation is probably the most important cause of the different DNA methylation patterns between populations. A mouse study found that in primordial germ cells, the cell with activation-induced deaminase (AID: an enzyme) deficiency were far more methylated that the wild type indicating that DNA methylation is possibly contributed to genetic factors [95]. A study conducted in 2010 found that there was an association between CpG sites and SNPs that differed by population [88]. Gene function and the sequence in the promoter have also been shown to be major predictors of methylation in the proximity of that gene [42]. Taken together with the studies finding strong relations *in cis* between CpG sites and SNPs mentioned earlier, this suggests that genetic factors are an important determinant of the presence of population stratification in DNA methylation.

Regardless of the mechanisms behind the observed differences in DNA methylation across populations, there are no established methods to account for population stratification in DNA methylation association studies. Population stratification is a known problem in GWAS studies, and can be expected to present a similar problem in DNA methylation studies. DNA methylation microarrays are continuing to expand, with the most recent having over 450k CpG sites [8]. In both GWAS and methylation studies, self-reported race may not be accurate or account for the actual ancestry. So far population stratification has been unaddressed in DNA methylation association studies, even though it presents a potential problem that could lead to spurious results. The goal of this thesis is to document the effects of population stratification in two large datasets, and to identify a method to account for population stratification in DNA methylation studies.

2. Methods

2.1 Assessing Population Stratification

To determine whether population stratification was present in DNA methylation data, we first analyzed a dataset collected from umbilical cord blood from newborns to see if there was an association between DNA methylation and race (self-reported) accounting for covariates (gender and chip). The data was collected as part of a study of pregnant women with a history of neuropsychiatric diseases at Emory University and DNA methylation was extract with HumanMethylation27 BeadChip [96]. Upon delivery of the baby, DNA was collected from umbilical cord blood. We used the R package CpGassoc [97] to fit a linear model where methylation was regressed on a categorical variable for race; we included sex and chip as covariates.

In addition, we had methylation data on a group of 423 individuals involved in a larger on Post-traumatic Stress Disorder. Patients were recruited at a public hospital in an urban area. The data was a small part of a larger study looking at stressful life events and genetic and environmental factors [98]. Whole blood was extracted at the Emory University Biomarker Service Core. DNA methylation was gathered on a subset of these individuals by using the Illumina Infinium Assay. We examined DNA methylation from chromosome one through twenty two.

The purpose of the analyses described below is to compare the ability of principal components from different sources in their ability to account for population stratification.

We collected principal components from pruned and unpruned methylation data. The ability of these principal components to adjust for population stratification will be based on the reduction in the number of CpG sites that are considered significant after these principal components are included as covariates in the analysis. Our goal is to identify a method that can then be used in DNA methylation studies to account for the presence of population stratification in the data.

2.2 Principal Component Analysis of Genome-wide Methylation Data

For all autosomal CpG sites, we computed β-values as the total methylated signal divided by the total signal:

$$\beta = \frac{M}{U + M}$$

where *M* is the total methylated signal for that site, and *U* is the total unmethylated signal for that site.

CpG observations with a detection p-value less than 0.001 were set to NA. Individuals that had a mean total signal that was less than half of the overall median of the mean signals or was less than 2000 were removed. When analyzing the methylation data, we considered both the logit-transformed and the non-transformed data. If β is the beta value at a given CpG site, the logit transformation is:

$$\log\left(\frac{\beta}{1-\beta}\right).$$

This transformation has the benefit of stabilizing the variance [99]. The quantile normalized data was also examined in a similar fashion to the unnormalized data. We quantile normalized the methylated and unmethylated signals together. We first order the

combined signals by individual from lowest to highest. We then replace each value by the respective mean for that ordered value across all individuals regardless if it is an unmethylated or methylated signal. So, if we have m sites, and n individuals and if $x_{(i),k}$ is the i^{th} ordered value for individual k, the quantile normalized value would be:

$$\frac{1}{n}\sum_{j=1}^{n} x_{(i),j}.$$

For example, if A is our signal matrix:

$$\begin{pmatrix} 124 & 588 & 544 & 412 \\ 515 & 712 & 398 & 651 \\ 671 & 423 & 645 & 516 \\ 782 & 814 & 743 & 687 \end{pmatrix},$$

we then order the values for each individual:

$$\begin{pmatrix} 124 & 423 & 398 & 412 \\ 515 & 588 & 544 & 516 \\ 671 & 712 & 645 & 651 \\ 782 & 814 & 743 & 687 \end{pmatrix},$$

take the average across the ordered values:

$$\begin{pmatrix} 339.25 & 339.25 & 339.25 & 339.25. \\ 540.75 & 540.75 & 540.75 & 540.75 \\ 669.75 & 669.75 & 669.75 & 669.75 \\ 756.5 & 756.5 & 756.5 & 756.5 \end{pmatrix},$$

and put the values back in their original order. Thus the quantile normalized data would be:

$$\begin{pmatrix} 339.25 & 540.75 & 540.75 & 339.25 \\ 540.75 & 669.75 & 339.25 & 669.75 \\ 669.75 & 339.25 & 669.75 & 540.75 \\ 756.5 & 756.5 & 756.5 & 756.5 \end{pmatrix}.$$

The new beta values are then calculated based on these quantile normalized

signals. In our example above, rows one and three are the unmethylated signal for sites 1 and 2, and rows two and four are the methylated signals for sites 1 and 2. Thus, the beta value for individual 1, site 1, would be computed as 540.75/(339.25+540.75) Quantile normalization has the benefit of giving each individual the same distribution for their signals and thus removing systematic differences between individuals.

When adjusting for population stratification in GWAS, it is common to work with a roughly independent set of SNPs that have been pruned to remove highly correlated SNPs. Thus, prior to performing principal component analysis, we tried pruning the methylation data in a variety of ways for both the quantile-normalized and unnormalized data, and both the logit-transformed and untransformed data. In each case, we performed principal component analysis on:

- unpruned (complete) data.
- data pruned to have only CpG sites with $r^2 < 0.25$. This process is explained in further detail below.
- data pruned to have only CpG sites with $r^2 < 0.10$.

After this was done, there were six different sets of principal components each for the quantile-normalized and unnormalized data.

We pruned the data separately by chromosome. If a chromosome had over 5000 CpG sites we divided it further into windows of 5000 CpG sites. We then performed the following process on each window:

1) Let β be our matrix of DNA methylation data, with each row representing a CpG site and each column an individual.

2) Set any missing values in β equal to the mean for that CpG site.

3) Let R be the absolute value of our correlation matrix:

$$\left|\begin{pmatrix} 1 & r_{2,1} & \cdots & r_{1,5000} \\ r_{2,1} & 1 & \cdots & r_{2,5000} \\ \vdots & \vdots & \ddots & \vdots \\ r_{5000,1} & \cdots & \cdots & 1 \end{pmatrix}\right|$$

where $r_{i,j}$ represents the correlation between the ith and jth cpg site. This matrix is symmetrical.

4) The diagonal is then set to zero:

$$\left|\begin{pmatrix} 0 & r_{2,1} & \cdots & r_{1,5000} \\ r_{2,1} & 0 & \cdots & r_{2,5000} \\ \vdots & \vdots & \ddots & \vdots \\ r_{5000,1} & \cdots & \cdots & 0 \end{pmatrix}\right|.$$

5) For each site we then calculate the number of connections, where a connection is defined as a correlation above 0.5 (corresponding to an r^2 of 0.25).

$$\nu = (v_1, v_2, ..., v_{4999}, v_{5000})$$

where :

$$v_i = \sum_{j=1}^{5000} I(r_{i,j} > .5)$$

and I is the indicator function.

6) The sites with a v_i equal to zero are set aside since they have no connections as we defined them. We then focus on the absolute value of the reduced correlation matrix, R*.

7) We then begin a loop removing the site with the most connections:

$$v^* = (v_1^*, v_2^*, ..., v_n^*)$$

$$r^* = (r_{.1}^*, r_{.2}^*, ..., r_{.n}^*)$$

where v^* represents the number of connections from R^*, n is the number of CpG sites with connections, and r^* is the column sums of R^*. We remove the CpG site that had the maximum number of connections: max(v^*). If there are two or more sites with that value, we remove the one with the higher r^* value.

8) Upon removing this CpG site, the row and column corresponding to it are set to zero in R^* and steps 7 and 8 are repeated until there are no more connections.
9) Once there are no more connections the matrix of CpG sites is reassembled to include the CpG sites set aside in step 6.

We repeated this process on new windows of 5000 CpG sites, until there were no longer any connections at the .25 level within each chromosome. For the .1 pruning, we used the pruned data set from the .25 pruning, and again pruned by chromosome using the method above, only using the square root of .1 as the cutoff. We therefore had principal components from the $r^2<0.1$ data ($PC_{r2<0.1}$), the $r^2<0.25$ ($PC_{r2<0.25}$), and the unpruned data set ($PC_{unprune}$). For each of these there is a principal component from the logit-transformed and untransformed data. Italicized will correspond to principal components from a logit-transformed data set. So if X is our matrix after pruning by $r^2<0.1$, $PC_{r2<0.1}$ is the principal components from X, and $PC_{\mathit{r2<0.1}}$ is the principal components from log(X/(1-X)).

We also performed "informed pruning" of the methylation data by incorporating data on genetic variation from the 1000 Genomes Project [70]. For each CpG site we

determined the proximity of a nearby SNP. We then created pruned data sets that only included CpG sites that were within a certain distance of a SNP. We created seven informed pruned data sets based on the following criteria: if a CpG sites was in the same spot as a SNP, within one single base pair, within two bases, within five bases, within ten bases, within fifty bases, or within one hundred bases. The purpose of this informed pruning was to exploit information on allele frequency differences due to population stratification in these nearby SNPs. We hypothesized that the principal components from these CpG sites would pick up on the population differences present in the SNPs and provide an improved adjustment for population stratification. For each of these sets we calculated principal components from both the logit-transformed and untransformed data, and both the quantile-normalized and unnormalized data. Thus, there are an additional seven sets of principal components for each: from those on the same spot (PC_{0bp}), within one base pair (PC_{1bp}), within two (PC_{2bp}), within five (PC_{5bp}), within ten (PC_{10bp}), fifty (PC_{50bp}), and one hundred base pair (PC_{100bp}). The logistic transformed principal components are calculated in a similar fashion. So, if Y is our matrix of beta values of CpG sites within one base pair of a SNP, PC_{1bp} are the principal components from Y, and PC_{1bp} are the principal components from log(Y/(1-Y)).

Finally, we considered standardizing the methylation data in a similar manner as put forth by Patterson et. al [100]. If C was our matrix of β-values, with each row representing an individual, our standardized data was:

$$M = \frac{C - columnmeans(C)}{column\,\mathrm{var}(C)}$$

We then calculated the principal components based on this standardized methylation data. To assess which principal components were significant, we used the

Tracy-Widom test [100]. For the Tracy-Widom test, let M be our standardized β-values with *m* rows (individuals) and *n* columns (CpGs). Then

1) Compute $X=M M^T$
2) Compute the eigenvalues of X and order them:

   $\lambda_1 > \lambda_{2>\ldots} > \lambda_{m'} > 0$

   where $m'=m-1$
3) Then we use these eigenvalues to estimate *n'*

$$n' = \frac{(m+1)(\sum_{i=1}^{m'} \lambda_i)^2}{((m-1)\sum_{i=1}^{m'} \lambda_i^2) - (\sum_{i=1}^{m'} \lambda_i)^2}.$$

4) Then we compute our test statistic:

$$l = \frac{m'\lambda_1}{\sum_{i=1}^{m'} \lambda_i}.$$

5) This test statistic is then standardized:

$$x = \frac{l - \mu(m,n')}{\sigma(m,n')}$$

   where:

$$\mu(m,n') = \frac{(\sqrt{n'-1} + \sqrt{m})^2}{n'}$$

   and

$$\sigma(m,n') = \left(\frac{\sqrt{n'-1} + \sqrt{m}}{n'} \right) \left(\frac{1}{\sqrt{n'-1}} + \frac{1}{\sqrt{m}} \right).$$

6) *x* then follows a Tracy- Widom distribution [101]. To compute the probability we

used the RMTstat package [102].

7) If this value proves to be significant, steps 4-6 are repeated with λ_2 , with l becoming:

$$l = \frac{(m'-1)\lambda_2}{\sum_{i=2}^{m'}\lambda_i}.$$

8) This process is repeated until there is an eigenvalue that is not significant. These eigenvalues correspond to the respective principal components. For a more detailed explanation of the algorithm see Patterson et al. [100].

2.3 Comparison of Methods to Adjust for Population Stratification

We next assessed how effective each of these sets of principal components was in accounting for the population stratification in the data. First, we performed a general analysis of the DNA methylation data modeled on race to get an idea of how many CpG sites associate significantly with race. We performed multivariate linear regression that modeled the beta values (proportion of methylation) on race, including covariates for sex, age, and categorical variables for chip and location on chip. These covariates were included to adjust for the known effects of sex and age along with the effects of chip and location on chip [27-33, 103]. To assess significance we used the Benjamini-Hochberg FDR method and the Holm method (a stepdown Bonferroni procedure) [104, 105]. Since in many studies the researchers will not have reliable information on the ancestral background of the participants and therefore will not know which principal components are good proxies for ancestry, we attempted to adjust for population stratification by

including the first ten principal components from each approach described above as covariates in the model. To assess the performance of each approach, we calculated the reduction in the number of CpG sites that were significantly associated with race according to FDR or Holm criteria. We next refit the model including just the principal components associated with race in the analysis. We used our publicly available R package CpGassoc [97] to perform all of the above analyses.

3. Results

After cleaning the data, the PTSD sample consisted of 417 individuals: 388 African Americans, twenty four Caucasians, four of mixed heritage, and one other (Table 1). The mean age was 42 with 118 females and 299 males. The neonatal data consisted of 303 babies, 251 which were Caucasian, 29 that were African-American, and the remaining were Hispanic, Asian, or other. The sex of the babies was roughly split half and half (Table 1).

In the neonatal sample the analysis of DNA methylation modeled on race with covariates for sex and chip was performed on 27,578 CpG sites. When the dependent variable was the untransformed β-values, 540 sites were significantly associated with self-reported race according to FDR criteria and 126 according to Holm criteria (Table 2). Using the logit-transformed β-values yielded fewer significant sites, with 473 FDR and 104 Holm-significant (Table 2).

In the PTSD sample modeling DNA methylation on race while adjusting for age, sex, chip, and location on chip, the analysis was done on 473,864 CpG sites. In the

unnormalized data there were 6895 FDR-significant sites and 1024 that were Holm-significant (Table 2). In the logit-transformed data there were 7216 FDR significant and 838 Holm significant (Table 2). For the quantile normalized data, there were 2123 FDR and 1012 Holm significant using the untransformed beta values (Table 2). Using the logit-transformed data there were 8357 FDR and 815 Holm significant sites (Table 2).

We next looked at the principal components. In the unnormalized data $PC_{r2<0.25}$ was from a pruned data set of 174,079 CpG sites and $PC_{r2<0.1}$ was from one of 40,572. For the quantile-normalized data $PC_{r2<0.25}$ was principal components taken from 256,976 CpG sites and $PC_{r2<0.1}$ was from 121,776. For the informed pruning there were 5,431 sites on a base pair, 11,749 within one base pair, 13,521 within two base pair, 19,213 within five, 46,822 within ten, 137,812 within fifty, and 223,333 within one hundred base pair. These informed pruned data sets were used to compute the following principal components: PC_{0bp}, PC_{1bp}, PC_{2bp}, PC_{5bp}, PC_{10bp}, PC_{50bp}, and PC_{100bp} respectively.

For the most part the principal components that were picking up on race appeared in the first ten principal components (see Table 3 for list of Holm significant principal components when looking at race in the non-informed). The exception to this was the unpruned logit-transformed unnormalized data, where the second most significant principal component was PC eleven (Table 3). As might be expected due to the removal of spurious CpG sites, as the data was pruned more the principal components associated with race moved forward (Table 3). An interesting observation was that the principal components picking up on race in the logit transformed data always involved higher-order principal components as compared to the untransformed data. For example, in $PC_{r2<0.25}$ principal component four is most associated with race, while in $PC_{r2<0.25}$ it is

principal component six.

Interestingly, when using the Tracy-Widom test to determine whether any of the principal components were significant, the method failed to detect any significant principal components [100]. However, plots of the principal components illustrate that some were indeed correlated with race (Figure 1). The skree plots of the principal components also did not reveal any ancestry-informative principal components as significant, instead suggesting that only the first principal component explained the majority of the genomic variation in DNA methylation (Figure 2). This contradicts what was observed in the data with higher-order principal components picking up on race (Figure 1). It is also important to note that when the methylation data was standardized in a similar fashion to SNP data, the first ten principal components failed to pick up on race at all (Figure 3) [100].

Using each set of principal components on both of these datasets by including the first ten principal components as covariates; we saw an almost uniform reduction in the number of significant sites (Table 4). Interestingly $PC_{unprune}$ actually increased the number of race-associated sites for both the logit transformed and untransformed data (Table 4). This is possibly due to the principal components associated with race not being in the first ten components (Table 3). However, when $PC_{unprune}$ was included in the analysis of the untransformed data there was an 88.2% percent reduction in the number of FDR-significant sites and a 64.2% reduction in the number of Holm-significant sites (Table 4). Looking at Table 3, we can see that $PC_{unprune}$ actually had principal components significant for race in the first ten principal components. As can be seen in Table 4, pruning the data before computing principal components did not lead to

substantially more reduction in the number of significant sites, resulting in a decrease of at most 25 sites ($PC_{r2<0.25}$ vs $PC_{r2<0.1}$ Table 4).

With the quantile normalized data a similar pattern could be seen (Table 5). One of the main differences was that $PC_{unprune}$ no longer led to an increase in the number of significant sites (Table 4 vs. Table 5). This could be due to race-associated principal components now being in the first ten principal components (Table 3). The inclusion of the first ten from $PC_{unprune}$ did not lead to as much of a reduction as the other principal components, however. Excluding this set of principal components the mean percent reduction in sites in the non logit data is 87.19% for FDR and 64.19% for the Holm (Table 5). For the logit transformed data it was 89.91% and 77.04% respectively (Table 5).

The informed pruning yielded similar results with a general reduction in the number of sites being quite strong (Table 4). PC_{5bp} did the best in reducing FDR significance in logit transformed data, while PC_{0bp} did the best for reducing Holm significance in both data types (Table 4). In the quantile-normalized data there was a similar reduction in the number of significant sites (Table 5). PC_{5bp} performed the best in reducing the number of FDR significant sites in the logit data (Table 5). In the non-logit transformed data PC_{1bp} reduced the number of FDR significant sites the most (Table 5).

There was one individual whose race was defined as “other.” This person seemed to be driving some of the significant sites (Figures 4-6). Removing this individual led to an increase in the significance of some principal components and race (Table 6). The above processes were thus repeated after excluding this individual. The informed pruning now performed remarkably well in removing the number of significant sites (Table 7).

The set of principal components that seemed to remove the most significant sites was PC_{5bp} (Table 7). This set removed 96.12% of the FDR-significant and 85.37% of the Holm-significant sites in the untransformed data, while in the logit-transformed data it led to decreases of 97.82% and 93.48% respectively (Table 7). PC_{1bp} removed the largest amount of Holm-significnat sites (Table 7). $PC_{r2<0.1}$ did perform well, but not as well as the informed pruning (Table 7). A similar result was observed in the quantile-normalized data (Table 8).

Given that we have the advantage of knowing the individuals races, we also did a "data-snooping" analysis where we just included the principal components found to be significant by race and with an F-statistic above 20 (Table 3). This was in the data with the individual with race other. Contrary to expectations, this actually led to an overall increase in the number of significant sites across the board both in quantile and non quantile normalized data (Tables 9 & 10). One exception was in the quantile normalized data with principal components two, three and four from $PC_{r2<0.1}$ (Table 10).

Using the same methods to assess reduction of population stratification in the 27k neonatal data, we see a somewhat similar result (Table 11). Using principal components from the untransformed pruned data did lead to a reduction in sites that were significantly associated with race, but principal components from the logit transformed data did not lead to a large reduction in the number of significant sites (Table 11). The reduction in number of significant sites in the analysis of the data after including $PC_{r2<0.1}$ in the neonatal data was not as great as it was in the 450k data, but that could also be due to the smaller number of sites being analyzed. We also tried using the informed method on this data and saw a reduction in the number of significant sites (Table 11), though principal

components from informed pruning did not do as well as those from correlation-based pruning ($PC_{r2<0.1}$, Table 11). This could be due to the smaller size of the data, since there were only 27,578 CpG sites on the 27k array. The principal components for informed pruning were thus based on only 116 CpG sites on a SNP, 278 within one base pair of a SNP, 329 within two, 520 within five, 1,282 within ten, 6,441 within fifty, and 10,581 within one hundred base pair. This is compared to 18401 with an $r^2<0.5$ by chromosome and 6929 with an $r^2<0.10$ by chromosome (Table 11).

4. Discussion

The method of using principal components from methylation data is a potential way to correct for population stratification in a study. Unfortunately, it does not completely remove all of the population stratification that is present; there were still some CpG sites significantly associated with race even with the addition of principal components. However principal components did reduce the number of significant sites substantially. Outside of $PC_{unprune}$, these principal components always lead to a larger percent reduction in the logit transformed beta values vs. the untransformed data. It was interesting that the logit-transformed data had more FDR-significant but less Holm-significant sites in comparison to the untransformed data (Table 4, 5, 7, & 8).

In the data that included the single individual from the "Other" category, the best set for reducing the number of FDR significant sites in the untransformed data was PC_{1bp} (89.01% reduction, Table 4). In the logit-transformed data the best FDR reduction was by the $PC_{r2<0.1}$ (93.24% reduction, Table 4). For reducing the number of Holm-significant

data it was PC_{0bp} (65.92% and 78.76% for untransformed and logit-transformed respectively Table 4). In the quantile-normalized data, PC_{1bp} did the best in reducing the number of FDR-significant sites in the untransformed data (by 88.02%; Table 5) while in the logit-transformed data PC_{5bp} performed the best (90.39% reduction Table 5). For Holm-significant sites in the untransformed data PC_{1bp} performed the best (66.7% Table 5), while for the logit transformed PC_{0bp} performed the best (75.46% Table 5).

Upon removing the one individual classified as other, almost all sets of principal components had a higher percent reduction in the number of significant sites. For reducing the number of FDR-significant sites PC_{5bp} did the best in the untransformed data (96.12% for untransformed, 97.82% logit-transformed Table 7). In the quantile-normalized data the $PC_{r2<0.1}$ had the most reduction (94.78% and 96.81% for untransformed and logit-transformed respectively Table 8). It is important to note though that this was only slightly better than PC_{5bp}: (94.69% and 96.63% Table 8). In terms of Holm reduction, PC_{0bp} did the best for the non-normalized transformed data (86.37% and 94.38% for non logit and logit respectively Table 7). For the quantile-normalized data it was PC_{1bp} (85.97% and 92.92% Table 8). In the data without the outlier the set of principal components that reduces the most sites was the same for the untransformed and logit-transformed data, this was not the case when the outlier was included in the data.

The non-informed pruning was adequate in reducing the population stratification. The method of pruning the data sufficiently to perform principal components is computationally intense, but the overall approach does help in removing a large amount of population stratification. The inability of the unpruned logit transformed data to truly remove any of the population stratification is probably due to the interesting phenomenon

that we observed where principal components related to race were in the later principal components compared to their untransformed counterparts (Table 3). This result in the unpruned data is probably due to the first ten principal components not picking up on race (Table 3). This was also seen once the individual with race “other” was removed (Table 6). Pruning the data helps to get rid of excess signals that principal components are picking up on.

The informed pruning also removed a substantial (if not greater) amount of the population stratification. Unlike pruning based on correlation, informed pruning offers a more intuitive type of pruning as it is hopefully picking up on population differences due to SNP data. It is interesting to note that restricting our analysis to CpG sites that were located on or near a SNP was not as successful in reducing the amount of FDR significant sites as including sites within five single base pairs. This is probably just due to the range of the effect of SNPs on nearby CpG sites. It was also interesting that in reducing the number of Holm-significant sites, principal components from CpG sites close to SNPs (either in the same spot or within one base pair) led to the greatest reduction. This was seen with and without the individual whose race was “other” (Table 4, 5, 7, & 8 bolded in Holm column). The reason for this is unknown. The more stringent cutoff for Holm significance could make it so that only CpG sites close to SNPs are being picked up. However when we plotted the density of the distance to SNPs for the Holm significant sites, it did not appear that they were close to SNPs (Figure 7). These sites could be Holm-significant because they are close to yet unknown SNPs which are influencing these CpG sites. For each of the analyses excluding the “other” individual, over half of the Holm significant sites were less than one hundred base pair away from a SNP. More

research will be needed to understand this relationship.

The improvement in the results from the informed pruned sets after the person with “other” race was removed is probably based on that individual’s CpG sites differing due to nearby SNPs. The “other” individual was probably a race different from the African American and Caucasian populations that were predominant in the study and thus had SNP genotypes from a different allele frequency distribution. The different SNPs in these regions probably contributed to the different methylation values *in cis* for that individual. This in all likelihood influenced the principal components from that data set. Upon removing this individual, the principal components would more likely pick up on the two major races in the group. With the outlier the informed pruning only did better than the non-informed pruning, but once the outlier was removed it did noticeably better.

It was interesting that the normalization method mentioned by Patterson et al. [100] possibly removed all variation in the methylation principal components. The method appears to not transfer from SNP data to methylation data. In SNP data we want to stabilize the variance so as to not place greater emphasis on common sites (which are more variable), while with methylation data we want to focus on the most variable sites. The normalization Patterson et al. [100] suggested for SNP data downweights the highly variable sites which are driving the patterns in the data. Also when the Tracy-Widom test was performed on the principal components from the normalized data it actually detected significant principal components. However, when these principal components were plotted, they did not seem to pick up on anything. The Tracy-Widom test may not be applicable to principal components from continuous data.

Trying these methods in the neonatal data, $PC_{r2<0.1}$ did the best in reducing the

number of significant sites (94.63% and 96.41%, Table 11). PC_{100bp} also did well (92.78 and 96.41 for FDR reduction in non logit and logit respectively Table 11). The pruning based on correlation is not as computationally intensive on the 27k data as the 450k due to the small amount of data. Also, the information on the proximity to SNPs was based on data from the 450k illumina data set, which does not have all of the 27k data (about 2000 were missing). While it is unlikely that those extra 2000 could have contributed something, it cannot be ruled out. The informed pruning may have been more successful if there had been more CpG sites in proximity to SNPs.

To summarize, it appears that for reducing the number of Holm-significant sites, the principal components that did the best was always from the logit transformation of one of the data sets of CpG sites within one base pair of a SNP, or in the same position of the SNP. More often than not, it was PC_{0bp}. If the researcher is focused on reducing the number of sites showing association with race at a stringent cutoff (i.e. Holm-significant or Bonferroni-significant sites), it would probably be best to use principal components from one of those data sets to account for population stratification in the data. For reducing FDR significance it is not as easy to make a suggestion. PC_{5bp} appears to do well overall, as does $PC_{r2<0.1}$ does. Either would probably do well, but PC_{5bp} does take less to time to calculate. Since a researcher will want to remove as much confounding as possible (i.e. the most population stratification) using PC_{5bp} would make the most sense.

In conclusion, we have developed an approach that can be used to account for population stratification in DNA methylation studies. Using informed pruning, a researcher can obtain principal components that will successfully remove a wide majority of the effects of population stratification. Computing naïve principal components where

the data are pruned based on correlation can also be performed but it did not perform as well, and is very computationally inefficient. The informed pruning exploits information on location of known SNPs, and its performance in removing the majority of significant associations with race suggests that these principal components are picking up on the population stratification present in these SNPs. Finally, this method allows researchers to address population stratification when SNP genotype data is not available for the individuals in their study, making it of great benefit to a wide range of clinical and public health studies of DNA methylation data.

5. Bibliography

1. Waddington, C.H., *The Epigenotype.* Int J Epidemiol, 2011.
2. Jaenisch, R. and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.* Nat Genet, 2003. **33 Suppl**: p. 245-54.
3. Petronis, A., *Epigenetics as a unifying principle in the aetiology of complex traits and diseases.* Nature, 2010. **465**(7299): p. 721-7.
4. Maher, B., *Personal genomes: The case of the missing heritability.* Nature, 2008. **456**(7218): p. 18-21.
5. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
6. Weedon, M.N., et al., *Genome-wide association analysis identifies 20 loci that influence adult height.* Nat Genet, 2008. **40**(5): p. 575-83.
7. Yang, J., et al., *Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits.* Nat Genet, 2012. **44**(4): p. 369-75.
8. Bibikova, M., et al., *High density DNA methylation array with single CpG site resolution.* Genomics, 2011. **98**(4): p. 288-95.
9. Tost, J., *DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker.* Methods Mol Biol, 2009. **507**: p. 3-20.
10. Irizarry, R.A., et al., *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nat Genet, 2009. **41**(2): p. 178-86.
11. Egger, G., et al., *Epigenetics in human disease and prospects for epigenetic therapy.* Nature, 2004. **429**(6990): p. 457-63.
12. Eckhardt, F., et al., *DNA methylation profiling of human chromosomes 6, 20 and 22.* Nat Genet, 2006. **38**(12): p. 1378-85.
13. Lieb, J.D., et al., *Applying whole-genome studies of epigenetic regulation to study human disease.* Cytogenet Genome Res, 2006. **114**(1): p. 1-15.
14. Lewin, B., *Genes IX*. 9th ed. 2008, Sudbury, Mass.: Jones and Bartlett Publishers. xvii, 892 p.
15. Suter, C.M., D.I. Martin, and R.L. Ward, *Germline epimutation of MLH1 in individuals with multiple cancers.* Nat Genet, 2004. **36**(5): p. 497-501.
16. Tekpli, X., et al., *DNA methylation of the CYP1A1 enhancer is associated with smoking-induced genetic alterations in human lung.* Int J Cancer, 2011.
17. Schuebel, K.E., et al., *Comparing the DNA hypermethylome with gene mutations in human colorectal cancer.* PLoS Genet, 2007. **3**(9): p. 1709-23.
18. Feinberg, A.P. and B. Tycko, *The history of cancer epigenetics.* Nat Rev Cancer, 2004. **4**(2): p. 143-53.
19. Breitling, L.P., et al., *Tobacco-smoking-related differential DNA methylation: 27K discovery and replication.* Am J Hum Genet, 2011. **88**(4): p. 450-7.
20. Das, P.M., et al., *Methylation mediated silencing of TMS1/ASC gene in prostate*

cancer. Mol Cancer, 2006. **5**: p. 28.
21. Albany, C., et al., *Epigenetics in prostate cancer.* Prostate Cancer, 2011. **2011**: p. 580318.
22. Fackler, M.J., et al., *Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence.* Cancer Res, 2011. **71**(19): p. 6195-207.
23. Durham, A., et al., *Epigenetics in asthma and other inflammatory lung diseases.* Epigenomics, 2010. **2**(4): p. 523-37.
24. Javierre, B.M., et al., *Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus.* Genome Res, 2010. **20**(2): p. 170-9.
25. Weaver, I.C., et al., *Epigenetic programming by maternal behavior.* Nat Neurosci, 2004. **7**(8): p. 847-54.
26. McGowan, P.O., et al., *Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse.* Nat Neurosci, 2009. **12**(3): p. 342-8.
27. Rakyan, V.K., et al., *Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains.* Genome Res, 2010. **20**(4): p. 434-9.
28. Fraga, M.F., et al., *Epigenetic differences arise during the lifetime of monozygotic twins.* Proc Natl Acad Sci U S A, 2005. **102**(30): p. 10604-9.
29. Numata, S., et al., *DNA methylation signatures in development and aging of the human prefrontal cortex.* Am J Hum Genet, 2012. **90**(2): p. 260-72.
30. Gonzalo, S., *Epigenetic alterations in aging.* J Appl Physiol, 2010. **109**(2): p. 586-97.
31. Irier, H.A. and P. Jin, *Dynamics of DNA Methylation in Aging and Alzheimer's Disease.* DNA Cell Biol, 2012.
32. Alisch, R.S., et al., *Age-associated DNA methylation in pediatric populations.* Genome Res, 2012. **22**(4): p. 623-32.
33. Boks, M.P., et al., *The relationship of DNA methylation with age, gender and genotype in twins and healthy controls.* PLoS One, 2009. **4**(8): p. e6767.
34. Teschendorff, A.E., et al., *Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer.* Genome Res, 2010. **20**(4): p. 440-6.
35. Maegawa, S., et al., *Widespread and tissue specific age-related DNA methylation changes in mice.* Genome Res, 2010. **20**(3): p. 332-40.
36. Piyathilake, C.J., et al., *Race- and age-dependent alterations in global methylation of DNA in squamous cell carcinoma of the lung (United States).* Cancer Causes Control, 2003. **14**(1): p. 37-42.
37. Murgatroyd, C., et al., *The Janus face of DNA methylation in aging.* Aging (Albany NY), 2010. **2**(2): p. 107-10.
38. Kerkel, K., et al., *Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation.* Nat Genet, 2008. **40**(7): p. 904-8.
39. Heijmans, B.T., et al., *Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus.* Hum Mol Genet, 2007. **16**(5): p. 547-54.

40. Zhang, D., et al., *Genetic control of individual differences in gene-specific methylation in human brain.* Am J Hum Genet, 2010. **86**(3): p. 411-9.
41. Schalkwyk, L.C., et al., *Allelic skewing of DNA methylation is widespread across the genome.* Am J Hum Genet, 2010. **86**(2): p. 196-212.
42. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.* Nat Genet, 2007. **39**(4): p. 457-66.
43. Bell, J.T., et al., *DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.* Genome Biol, 2011. **12**(1): p. R10.
44. Kaati, G., et al., *Transgenerational response to nutrition, early life circumstances and longevity.* Eur J Hum Genet, 2007. **15**(7): p. 784-90.
45. Pembrey, M.E., et al., *Sex-specific, male-line transgenerational responses in humans.* Eur J Hum Genet, 2006. **14**(2): p. 159-66.
46. Kaati, G., L.O. Bygren, and S. Edvinsson, *Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period.* Eur J Hum Genet, 2002. **10**(11): p. 682-8.
47. Bygren, L.O., G. Kaati, and S. Edvinsson, *Longevity determined by paternal ancestors' nutrition during their slow growth period.* Acta Biotheor, 2001. **49**(1): p. 53-9.
48. Nadeau, J.H., *Transgenerational genetic effects on phenotypic variation and disease risk.* Hum Mol Genet, 2009. **18**(R2): p. R202-10.
49. Pentinat, T., et al., *Transgenerational inheritance of glucose intolerance in a mouse model of neonatal overnutrition.* Endocrinology, 2010. **151**(12): p. 5617-23.
50. Franklin, T.B., et al., *Epigenetic transmission of the impact of early stress across generations.* Biol Psychiatry, 2010. **68**(5): p. 408-15.
51. Walker, D.M. and A.C. Gore, *Transgenerational neuroendocrine disruption of reproduction.* Nat Rev Endocrinol, 2011. **7**(4): p. 197-207.
52. Reik, W., W. Dean, and J. Walter, *Epigenetic reprogramming in mammalian development.* Science, 2001. **293**(5532): p. 1089-93.
53. Sasaki, H. and Y. Matsui, *Epigenetic events in mammalian germ-cell development: reprogramming and beyond.* Nat Rev Genet, 2008. **9**(2): p. 129-40.
54. Macleod, D., V.H. Clark, and A. Bird, *Absence of genome-wide changes in DNA methylation during development of the zebrafish.* Nat Genet, 1999. **23**(2): p. 139-40.
55. Cubas, P., C. Vincent, and E. Coen, *An epigenetic mutation responsible for natural variation in floral symmetry.* Nature, 1999. **401**(6749): p. 157-61.
56. Rakyan, V.K., et al., *Transgenerational inheritance of epigenetic states at the murine Axin(Fu) allele occurs after maternal and paternal transmission.* Proc Natl Acad Sci U S A, 2003. **100**(5): p. 2538-43.
57. Morgan, H.D., et al., *Epigenetic inheritance at the agouti locus in the mouse.* Nat Genet, 1999. **23**(3): p. 314-8.
58. Chandler, V.L., W.B. Eggleston, and J.E. Dorweiler, *Paramutation in maize.* Plant Mol Biol, 2000. **43**(2-3): p. 121-45.
59. Masumoto, H., et al., *The inheritance of histone modifications depends upon the location in the chromosome in Saccharomyces cerevisiae.* PLoS One, 2011. **6**(12):

p. e28980.
60. Chandler, V.L., *Paramutation: from maize to mice.* Cell, 2007. **128**(4): p. 641-5.
61. Richards, E.J., *Inherited epigenetic variation--revisiting soft inheritance.* Nat Rev Genet, 2006. **7**(5): p. 395-401.
62. Rakyan, V.K. and S. Beck, *Epigenetic variation and inheritance in mammals.* Curr Opin Genet Dev, 2006. **16**(6): p. 573-7.
63. Bernstein, B.E., et al., *Genomic maps and comparative analysis of histone modifications in human and mouse.* Cell, 2005. **120**(2): p. 169-81.
64. Jablonka, E. and M.J. Lamb, *Epigenetic inheritance and evolution : the Lamarckian dimension*. 1995, Oxford ; New York: Oxford University Press. x, 346 p.
65. Cavalli-Sforza, L.L., P. Menozzi, and A. Piazza, *The history and geography of human genes*. 1994, Princeton, N.J.: Princeton University Press. xi, 541, 518 p.
66. Cavalli-Sforza, L.L. and M.W. Feldman, *The application of molecular genetic approaches to the study of human evolution.* Nat Genet, 2003. **33 Suppl**: p. 266-75.
67. Helgason, A., et al., *An Icelandic example of the impact of population structure on association studies.* Nat Genet, 2005. **37**(1): p. 90-5.
68. Cavalli-Sforza, L.L. and A.W. Edwards, *Phylogenetic analysis. Models and estimation procedures.* Am J Hum Genet, 1967. **19**(3 Pt 1): p. 233-57.
69. Altshuler, D.M., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.
70. *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
71. Lander, E.S. and N.J. Schork, *Genetic dissection of complex traits.* Science, 1994. **265**(5181): p. 2037-48.
72. Weiss, K.M., *Genetic variation and human disease : principles and evolutionary approaches*. Cambridge studies in biological anthropology. 1993, Cambridge ; New York: Cambridge University Press. xxiv, 354 p.
73. Satten, G.A., W.D. Flanders, and Q. Yang, *Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model.* Am J Hum Genet, 2001. **68**(2): p. 466-77.
74. Ardlie, K.G., K.L. Lunetta, and M. Seielstad, *Testing for population subdivision and association in four case-control studies.* Am J Hum Genet, 2002. **71**(2): p. 304-11.
75. Pritchard, J.K. and N.A. Rosenberg, *Use of unlinked genetic markers to detect population stratification in association studies.* Am J Hum Genet, 1999. **65**(1): p. 220-8.
76. Reich, D.E. and D.B. Goldstein, *Detecting association in a case-control study while correcting for population stratification.* Genet Epidemiol, 2001. **20**(1): p. 4-16.
77. Nicholson, G., et al., *Assessing population differentiation and isolation from single-nucleotide polymorphism data.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 2002. **64**: p. 695-715.
78. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.

79. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.
80. Bacanu, S.A., B. Devlin, and K. Roeder, *The power of genomic control.* Am J Hum Genet, 2000. **66**(6): p. 1933-44.
81. Devlin, B., K. Roeder, and S.A. Bacanu, *Unbiased methods for population-based association studies.* Genet Epidemiol, 2001. **21**(4): p. 273-84.
82. Cardon, L.R. and L.J. Palmer, *Population stratification and spurious allelic association.* Lancet, 2003. **361**(9357): p. 598-604.
83. Freedman, M.L., et al., *Assessing the impact of population stratification on genetic association studies.* Nat Genet, 2004. **36**(4): p. 388-93.
84. Wacholder, S., N. Rothman, and N. Caporaso, *Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias.* J Natl Cancer Inst, 2000. **92**(14): p. 1151-8.
85. Marchini, J., et al., *The effects of human population structure on large genetic association studies.* Nat Genet, 2004. **36**(5): p. 512-7.
86. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
87. Richards, E.J., *Population epigenetics.* Curr Opin Genet Dev, 2008. **18**(2): p. 221-6.
88. Liu, J., et al., *Identification of genetic and epigenetic marks involved in population structure.* PLoS One, 2010. **5**(10): p. e13209.
89. Terry, M.B., et al., *Genomic DNA methylation among women in a multiethnic New York City birth cohort.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(9): p. 2306-10.
90. Adkins, R.M., et al., *Racial differences in gene-specific DNA methylation levels are present at birth.* Birth Defects Res A Clin Mol Teratol, 2011. **91**(8): p. 728-36.
91. Nielsen, D.A., et al., *Ethnic diversity of DNA methylation in the OPRM1 promoter region in lymphocytes of heroin addicts.* Hum Genet, 2010. **127**(6): p. 639-49.
92. Kwabi-Addo, B., et al., *Identification of differentially methylated genes in normal prostate tissues from African American and Caucasian men.* Clin Cancer Res, 2010. **16**(14): p. 3539-47.
93. Figueiredo, J.C., et al., *Global DNA hypomethylation (LINE-1) in the normal colon and lifestyle characteristics and dietary and genetic factors.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(4): p. 1041-9.
94. Kaminsky, Z.A., et al., *DNA methylation profiles in monozygotic and dizygotic twins.* Nat Genet, 2009. **41**(2): p. 240-5.
95. Popp, C., et al., *Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency.* Nature, 2010. **463**(7284): p. 1101-5.
96. Schroeder, J.W., et al., *Neonatal DNA methylation patterns associate with gestational age.* Epigenetics, 2011. **6**(12): p. 1498-504.
97. Barfield, R.T., et al., *CpGassoc: an R function for analysis of DNA methylation microarray data.* Bioinformatics, 2012.
98. Smith, A.K., et al., *Differential immune system DNA methylation and cytokine regulation in post-traumatic stress disorder.* Am J Med Genet B Neuropsychiatr Genet, 2011. **156B**(6): p. 700-8.
99. Du, P., et al., *Comparison of Beta-value and M-value methods for quantifying*

methylation levels by microarray analysis. BMC Bioinformatics, 2010. **11**: p. 587.
100. Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* PLoS Genet, 2006. **2**(12): p. e190.
101. Tracy, C.A. and H. Widom, *Level-Spacing Distributions and the Airy Kernel.* Communications in Mathematical Physics, 1994. **159**(1): p. 151-174.
102. Johnstone, M., et al., *RMT: Distributions, Statistics and Tests derived from Random Matrix Theory.* R package version 0.2, 2009.
103. Chen, Y.A., et al., *Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray.* Genomics, 2011. **97**(4): p. 214-22.
104. Benjamini, Y., *The control of the false discovery rate in multiple testing under dependency.* Annals of Statistics, 2001.
105. Holm, S., *A simple sequentially rejective multiple test procedure.* Scandinavian Journal of Statistics, 1979. **6**(2): p. 65-70.

6. Tables

Table 1: Phenotype Data

Neonatal Data		
Phenotype	**N(%)**	**Mean (STD)**
Sex		
Male	151 (49.83%)	
Female	152 (50.17%)	
Race		
White	251 (82.84%)	
African American	29 (9.57%)	
Asian	6 (1.98%)	
4	2 (0.66%)	
Multi-Ethnic	15 (4.95%)	
PTSD Data		
Race		
African Americans	388 (93.05%)	
Caucasian	24 (5.76%)	
Mixed	4 (0.96%)	
Other	1 (0.24%)	
Age		41.63 (12.67)
Sex		
Male	299 (71.7%)	
Female	118(28.3%)	

Table 2: Number of Significant Sites

	Non Logit Transformed		**Logit Transformed**	
Data Set	**# of FDR**	**# of HOLM**	**# of FDR**	**# of HOLM**
Neonatal Data Set	540	126	473	104
PTSD Data	6895	1024	7216	838
PTSD Data quantile Normalized	8123	1012	8357	815

Table 3: PTSD Significant Principal Components with Race

Un-Normalized Data			Quantile Normalized Data		
PCA	**PC**	**F-Stat (P-value)**	**PCA**	**PC**	**F-Stat (P-value)**
$PC_{unprune}$	6	86.50 (2.39E-42)	**$PC_{unprune}$**	7	231.50 (2.84E-84)
	8	55.78 (9.96E-30)		8	26.35 (1.86E-15)
	7	19.44 (1.02E-11)		10	7.98 (3.64E-05)
	9	17.47 (1.27E-10)			
	10	9.93 (2.60E-06)	**$PC_{unprune}$**	8	97.62 (1.80E-46)
				9	60.03 (1.35E-31)
$PC_{unprune}$	9	111.71 (2.12E-51)		10	39.53(3.72E-22)
	11	27.59 (4.11E-16)		161	9.82 (3.03E-06)
	10	21.64(6.32E-13)		12	9.01 (9.03E-06)
	13	17.70(9.45E-11)		130	8.96 (9.67E-06)
$PC_{r2<0.25}$	4	197.39 (3.31E-76)	**$PC_{r2<0.25}$**	4	89.15 (2.37E-43)
	5	35.71 (2.85E-20)		5	56.82 (3.44E-30)
	3	10.25(1.71E-06)		7	14.16 (9.40E-09)
				3	7.58 (6.24E-05)
$PC_{r2<0.25}$	6	281.73(7.73E-95)			
	7	14.48 (6.20E-09)	**$PC_{r2<0.25}$**	7	179.62 (1.17E-71)
	381	8.45 (1.91E-05)		6	53.29 (1.30E-28)
				13	11.23 (4.59E-07)
$PC_{r2<0.1}$	2	164.54 (1.39E-67)		9	7.35 (8.57E-05)
	3	92.34 (1.54E-44)			
	4	8.91 (1.03E-05)	**$PC_{r2<0.1}$**	2	96.27 (5.55E-46)
				3	75.94 (3.12E-38)
$PC_{r2<0.1}$	3	226.282(4.32E-83)		4	26.17 (2.33E-15)
	2	44.936(9.35E-25)			
	4	11.519 (3.11E-07)	**$PC_{r2<0.1}$**	6	50.51 (2.37E-27)
	238	7.210 (1.03E-04)		5	47.09 (9.05E-26)
				7	33.80 (2.60E-19)
				4	23.96 (3.49E-14)

Table 4: PTSD, number of significant sites (% reduction) with Race Other (bold indicates most reduction)

	Non Logit Transformed		Logit Transformed	
PCA	**# of FDR (%)**	**# of HOLM (%)**	**# of FDR (%)**	**# of HOLM (%)**
None	6895	1024	7216	838
No Logit				
$PC_{unprune}$	811 (88.24)	366 (64.26)	545 (92.45)	194 (76.85)
$PC_{r2<0.25}$	784 (88.63)	371 (63.77)	522 (92.77)	194 (76.85)
$PC_{r2<0.1}$	759 (89.00)	363 (64.55)	***488 (93.24)***	186 (77.80)
Logit Transformed				
$PC_{unprune}$	9756 (-41.49)	528 (48.44)	7695 (-6.64)	369 (55.97)
$PC_{r2<0.25}$	790 (88.54)	367 (64.16)	520 (92.79)	198 (76.37)
$PC_{r2<0.1}$	788 (88.57)	363 (64.55)	529 (92.67)	190 (77.33)
Informed Pruning:				
No Logit				
PC_{0bp}	985 (85.71)	349 (65.92)	694 (90.38)	179 (78.64)
PC_{1bp}	***758 (89.01)***	351 (65.72)	539 (92.53)	185 (77.92)
PC_{2bp}	762 (88.95)	349 (65.92)	512 (92.9)	182 (78.28)
PC_{5bp}	770 (88.83)	350 (65.82)	503 (93.03)	183 (78.16)
PC_{10bp}	760 (88.98)	356 (65.23)	506 (92.99)	189 (77.45)
PC_{50bp}	859 (87.54)	363 (64.55)	599 (91.7)	192 (77.09)
PC_{100bp}	771 (88.82)	365 (64.36)	519 (92.81)	190 (77.33)
Logit Transformed				
PC_{0bp}	808 (88.28)	***345 (66.31)***	552 (92.35)	***178 (78.76)***
PC_{1bp}	834 (87.9)	350 (65.82)	561 (92.23)	183 (78.16)
PC_{2bp}	846 (87.73)	355 (65.33)	562 (92.21)	183 (78.16)
PC_{5bp}	992 (85.61)	362 (64.65)	676 (90.63)	188 (77.57)
PC_{10bp}	841 (87.8)	364 (64.45)	578 (91.99)	186 (77.80)
PC_{50bp}	1067 (84.53)	372 (63.67)	773 (89.29)	190 (77.33)
PC_{100bp}	1042 (84.89)	372 (63.67)	757 (89.51)	194 (76.85)

Table 5: PTSD Quantile Normalized, number of significant sites (% reduction) with Race Other (bold indicates most reduction)

	Non Logit Transformed		Logit Transformed	
PCA	**# of FDR (%)**	**# of HOLM (%)**	**# of FDR (%)**	**# of HOLM (%)**
None	8123	1012	8357	815

No Logit				
$PC_{unprune}$	1032 (87.3)	359 (64.53)	910 (89.11)	223 (72.64)
$PC_{r2<0.25}$	1068 (86.85)	365 (63.93)	832 (90.04)	216 (73.50)
$PC_{r2<0.1}$	1059 (86.97)	361 (64.33)	829 (90.08)	214 (73.74)
Logit Transformed				
$PC_{unprune}$	1702 (79.05)	363 (64.13)	1414 (83.08)	219 (73.13)
$PC_{r2<0.25}$	1060 (86.95)	366 (63.84)	836 (90.00)	216 (73.50)
$PC_{r2<0.1}$	985 (87.87)	361 (64.33)	808 (90.33)	216 (73.50)
Informed Pruning:				
No Logit				
PC_{0bp}	1274 (84.32)	358 (64.62)	1109 (86.73)	218 (73.25)
PC_{1bp}	1088 (86.61)	345 (65.91)	826 (90.12)	204 (74.97)
PC_{2bp}	1094 (86.53)	346 (65.81)	865 (89.65)	203 (75.09)
PC_{5bp}	1019 (87.46)	347 (65.71)	***803 (90.39)***	215 (73.62)
PC_{10bp}	1056 (87.00)	354 (65.02)	823 (90.15)	215 (73.62)
PC_{50bp}	1100 (86.46)	353 (65.12)	896 (89.28)	227 (72.15)
PC_{100bp}	1041 (87.18)	357 (64.72)	871 (89.58)	224 (72.52)
Logit Transformed				
PC_{0bp}	1117 (86.25)	341 (66.30)	970 (88.39)	***200 (75.46)***
PC_{1bp}	***973 (88.02)***	***337 (66.7)***	819 (90.20)	203 (75.09)
PC_{2bp}	1055 (87.01)	338 (66.60)	882 (89.45)	205 (74.85)
PC_{5bp}	1159 (85.73)	351 (65.32)	956 (88.56)	212 (73.99)
PC_{10bp}	1106 (86.38)	362 (64.23)	965 (88.45)	218 (73.25)
PC_{50bp}	1183 (85.44)	358 (64.62)	976 (88.32)	218 (73.25)
PC_{100bp}	1138 (85.99)	363 (64.13)	964 (88.46)	217 (73.37)

Table 6: PTSD Significant Principal Components with Race without "Other" Individual

Un-Normalized Data			**Quantile Normalized Data**		
PCA	**PC**	**F-Stat (P-value)**	**PCA**	**PC**	**F-Stat (P-value)**
$PC_{unprune}$	6	130.93 (1.21E-43)	**$PC_{unprune}$**	7	342.75 (1.09E-84)
	8	82.12 (3.36E-30)		8	40.19 (1.65E-16)
	7	28.48 (3.17E-12)		10	10.92 (2.49E-05)
	9	27.39 (8.17E-10)			
	10	14.41 (9.49E-07)	**$PC_{unprune}$**	8	143.36 (9.82E-47)
				10	89.93 (1.65E-32)
$PC_{unprune}$	9	163.96 (1.32E-51)		9	60.34 (2.20E-23)
	11	42.96 (1.72E-16)		12	13.52 (2.17E-06)
	10	32.44 (1.06E-13)		11	8.22 (3.24E-04)
	13	26.38 (1.97E-11)		130	7.19 (8.63E-04)

$PC_{r2<0.25}$	4	296.31 (2.52E-77)	**$PC_{r2<0.25}$**	4	132.24 (5.62E-44)
	5	54.01 (2.76E-21)		5	86.22 (2.02E-31)
	3	14.33 (1.02E-06)		7	19.90 (6.23E-09)
				3	10.49 (3.71E-05)
$PC_{r2<0.25}$	6	417.3 (3.03E-95)			
	7	22.84 (4.48E-10)	**$PC_{r2<0.25}$**	7	272.8 (2.54E-73)
	380	9.51 (9.43E-05)		6	75.16 (4.39E-28)
				13	16.31 (1.64E-07)
$PC_{r2<0.1}$	2	242.85 (6.52E-68)		9	10.45 (3.87E-05)
	3	139.49 (8.72E-46)			
	4	13.25 (2.78E-06)	**$PC_{r2<0.1}$**	2	140.43 (5.14 E-46)
				3	114.16 (2.81E-39)
$PC_{r2<0.1}$	3	338.85 (4.28E-84)		4	39.27 (3.51E-16)
	2	65.59 (4.42E-25)			
	4	17.25 (6.9E-08)	**$PC_{r2<0.1}$**	6	74.1 (2.37E-25)
	237	10.6 (3.35E-05)		5	71.7 (5.19E-27)
				7	51.22 (2.40E-20)
				4	24.305 (9.91E-14)

Table 7: PTSD number of significant sites (% reduction) Minus Race Other (bold indicates most reduction)

	Non Logit Transformed		**Logit Transformed**	
PCA	**# of FDR (%)**	**# of HOLM (%)**	**# of FDR (%)**	**# of HOLM (%)**
None	8250	998	9586	890
No Logit				
$PC_{unprune}$	355 (95.70)	152 (84.77)	262 (97.27)	68 (92.36)
$PC_{r2<0.25}$	358 (95.66)	165 (83.47)	244 (97.45)	71 (92.02)
$PC_{r2<0.1}$	351 (95.75)	165 (83.47)	234 (97.56)	68 (92.36)
Logit Transformed				
$PC_{unprune}$	14590 (-76.85)	387 (61.22)	12111 (-26.34)	336 (62.25)
$PC_{r2<0.25}$	346 (95.81)	156 (84.37)	237 (97.53)	68 (92.36)
$PC_{r2<0.1}$	347 (95.79)	156 (84.37)	255 (97.34)	70 (92.13)
Informed Pruning:				
No Logit				
PC_{0bp}	542 (93.43)	148 (85.17)	469 (95.11)	60 (93.26)
PC_{1bp}	322 (96.10)	140 (85.97)	247 (97.42)	55 (93.82)
PC_{2bp}	324 (96.07)	140 (85.97)	228 (97.62)	59 (93.37)
PC_{5bp}	***320 (96.12)***	146 (85.37)	***209 (97.82)***	58 (93.48)
PC_{10bp}	327 (96.04)	150 (84.97)	219 (97.72)	64 (92.81)
PC_{50bp}	390 (95.27)	156 (84.37)	291 (96.96)	68 (92.36)

PC_{100bp}	328 (96.02)	153 (84.67)	238 (97.52)	64 (92.81)
Logit Transformed				
PC_{0bp}	360 (95.64)	***136 (86.37)***	269 (97.19)	***50 (94.38)***
PC_{1bp}	351 (95.75)	143 (85.67)	246 (97.43)	59 (93.37)
PC_{2bp}	368 (95.54)	145 (85.47)	258 (97.31)	62 (93.03)
PC_{5bp}	495 (94.00)	152 (84.77)	410 (95.72)	70 (92.13)
PC_{10bp}	380 (95.39)	154 (84.57)	289 (96.99)	64 (92.81)
PC_{50bp}	633 (92.33)	161 (83.87)	580 (93.95)	73 (91.80)
PC_{100bp}	645 (92.18)	159 (84.07)	570 (94.05)	75 (91.57)

Table 8: PTSD Quantile Normalized number of significance sites Minus Race Other (bold indicates most reduction)

	Non Logit Transformed		**Logit Transformed**	
PCA	**# of FDR (%)**	**# of HOLM (%)**	**# of FDR (%)**	**# of HOLM (%)**
None	10326	962	10815	833
No Logit				
$PC_{unprune}$	589 (94.30)	157 (83.68)	412 (96.19)	76 (90.88)
$PC_{r2<0.25}$	655 (93.66)	155 (83.89)	387 (96.42)	76 (90.88)
$PC_{r2<0.1}$	637 (93.83)	161 (83.26)	401 (96.29)	77 (90.76)
Logit Transformed				
$PC_{unprune}$	1271 (87.69)	162 (83.16)	967 (91.06)	75 (91.00)
$PC_{r2<0.25}$	589 (94.30)	155 (83.89)	357 (96.70)	73 (91.24)
$PC_{r2<0.1}$	***539 (94.78)***	155 (83.89)	***345 (96.81)***	73 (91.24)
Informed Pruning:				
No Logit				
PC_{0bp}	976 (90.55)	162 (83.16)	761 (92.96)	90 (89.20)
PC_{1bp}	642 (93.78)	142 (85.24)	384 (96.45)	62 (92.56)
PC_{2bp}	663 (93.58)	140 (85.45)	401 (96.29)	63 (92.44)
PC_{5bp}	548 (94.69)	142 (85.24)	364 (96.63)	72 (91.36)
PC_{10bp}	612 (94.07)	153 (84.10)	371 (96.57)	77 (90.76)
PC_{50bp}	642 (93.78)	157 (83.68)	437 (95.96)	78 (90.64)
PC_{100bp}	613 (94.06)	158 (83.58)	411 (96.20)	79 (90.52)
Logit Transformed				
PC_{0bp}	766 (92.58)	147 (84.72)	573 (94.70)	67 (91.96)
PC_{1bp}	547 (94.70)	***135 (85.97)***	371 (96.57)	***59 (92.92)***
PC_{2bp}	644 (93.76)	139 (85.55)	451 (95.83)	64 (92.32)
PC_{5bp}	662 (93.59)	148 (84.62)	459 (95.76)	70 (91.60)
PC_{10bp}	603 (94.16)	156 (83.78)	471 (95.64)	80 (90.40)
PC_{50bp}	640 (93.80)	160 (83.37)	455 (95.79)	71 (91.48)
PC_{100bp}	620 (94.00)	159 (83.47)	435 (95.98)	73 (91.24)

Table 9: PTSD Data Snooping, Number of Significant Sites

	Non Logit Transformed		Logit Transformed	
PCA	**# of FDR**	**# of HOLM**	**# of FDR**	**# of HOLM**
None	6895	1024	7216	838
Non Logit				
$PC_{unprune}$ 6,8	86169	3217	80819	2693
$PC_{r2<0.25}$ 4	188691	81349	184175	77289
$PC_{r2<0.25}$ 4,5	182107	87031	181954	85405
$PC_{r2<0.1}$ 2	232168	77259	238639	77366
$PC_{r2<0.1}$ 2, 3	150300	59882	149462	57383
Logit Transformed				
$PC_{unprune}$ 9	74247	3389	73555	3703
$PC_{unprune}$ 9,10,11	143569	9589	147174	10539
$PC_{r2<0.25}$ 6	167480	58982	163961	56022
$PC_{r2<0.1}$ 3	105487	4197	104619	3887
$PC_{r2<0.1}$ 3,2	132985	40896	129093	37051

Table 10: PTSD Data Snooping Quantile, Number of Significant Sites

	Non Logit Transformed		Logit Transformed	
PCA	**# of FDR**	**# of HOLM**	**# of FDR**	**# of HOLM**
None	8123	1012	8357	815
Non Logit				
$PC_{unprune}$ 7	74830	1136	74516	1008
$PC_{unprune}$ 7,8	126624	8925	126252	8988
$PC_{r2<0.25}$ 4,5	179359	47563	182213	46990
$PC_{r2<0.1}$ 2,3,4	1378	350	1222	193
Logit Transformed				
$PC_{unprune}$ 8,9,10	135694	19803	136715	19828
$PC_{r2<0.25}$ 7	146401	8904	147809	8803
$PC_{r2<0.25}$ 6,7	132799	19555	134417	19207
$PC_{r2<0.1}$ 4,5,6,7	212721	78159	216044	78766

Table 11: Neonatal Data, Number of Significant Sites (% reduction) (bold indicates most reduction)

	Non Logit Transformed		Logit Transformed	
PCA	**# of FDR (%)**	**# of HOLM (%)**	**# of FDR (%)**	**# of HOLM (%)**

None	540	126	473	104
No Logit				
$PC_{unprune}$	55 (89.81)	18 (85.71)	18 (96.2)	6 (94.23)
$PC_{r2<0.25}$	43 (92.04)	18 (85.71)	17 (96.41)	***4 (96.15)***
$PC_{r2<0.1}$	***29 (94.63)***	***16 (87.30)***	***12 (97.46)***	5 (95.19)
Logit Transformed				
$PC_{unprune}$	442 (18.15)	106 (15.87)	351 (25.79)	91 (12.5)
$PC_{r2<0.25}$	2292 (-324.44)	446 (-253.97)	2059 (-335.31)	344 (-230.77)
$PC_{r2<0.1}$	546 (-1.11)	26 (79.37)	468 (1.06)	14 (86.54)
Informed Pruning:				
No Logit				
PC_{0bp}	185 (65.74)	62 (50.79)	153 (67.65)	45 (56.73)
PC_{1bp}	294 (45.55)	86 (31.75)	252 (46.72)	73 (29.81)
PC_{2bp}	248 (54.07)	79 (37.30)	191 (59.62)	61 (41.35)
PC_{5bp}	341 (36.85)	84 (33.33)	274 (42.07)	70 (32.69)
PC_{10bp}	131 (75.74)	36 (71.43)	84 (82.24)	30 (71.15)
PC_{50bp}	43 (92.04)	19 (84.92)	24 (94.93)	7 (93.27)
PC_{100bp}	39 (92.78)	20 (84.13)	17 (96.41)	6 (94.23)
Logit Transformed				
PC_{0bp}	177 (67.22)	58 (53.97)	140 (70.40)	45 (56.73)
PC_{1bp}	232 (57.04)	66 (47.62)	190 (59.83)	56 (46.15)
PC_{2bp}	213 (60.56)	62 (50.79)	166 (64.9)	48 (53.85)
PC_{5bp}	212 (60.74)	58 (53.97)	156 (67.02)	45 (56.73)
PC_{10bp}	174 (67.78)	45 (64.29)	115 (75.69)	32 (69.23)
PC_{50bp}	90 (83.33)	27 (78.57)	65 (86.26)	10 (90.38)
PC_{100bp}	198 (63.33)	48 (61.90)	137 (71.04)	31 (70.19)

7. Figures

Figure 1: Principal Components color-coded by race in PTSD data. Red signifies African Americans, green is Caucaasians, teal is Mixed Race, and purple is Other. Plots on bottom row are from the same data excluding the individual with race Other.

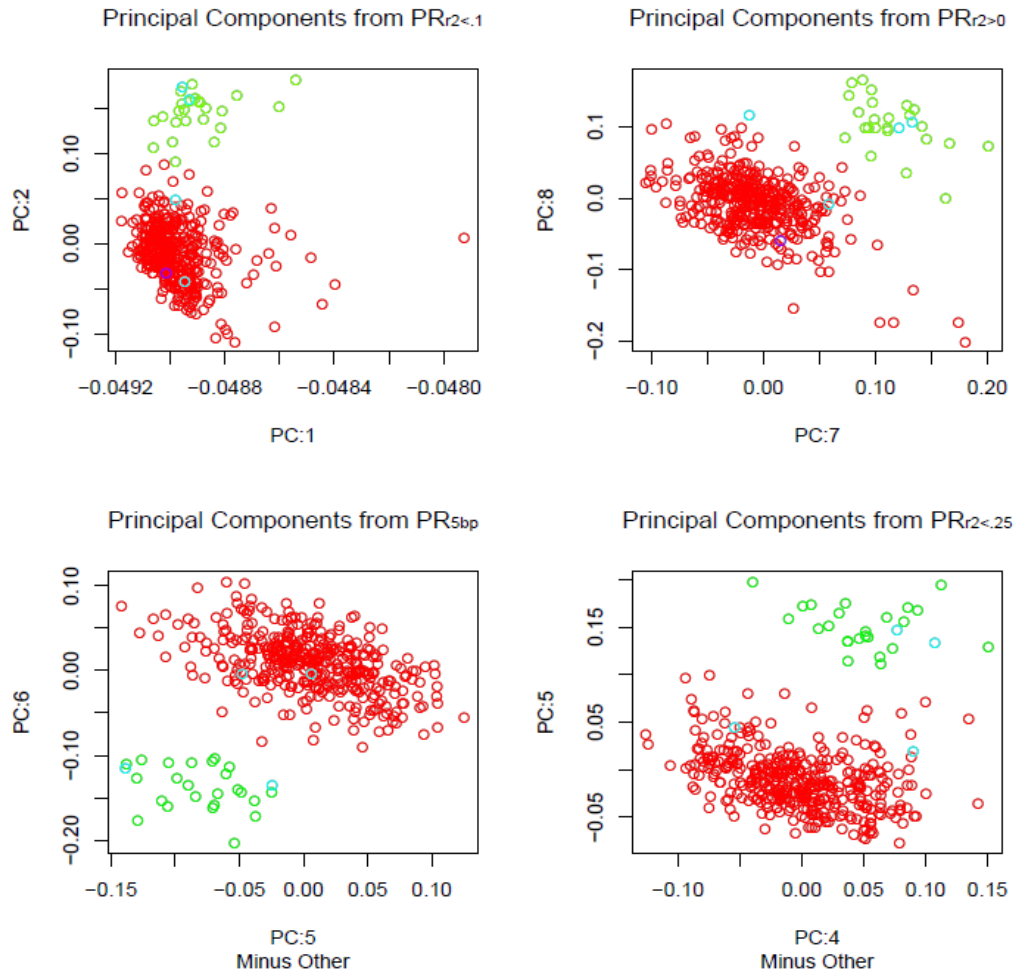

Figure 2: Skree plots of Principal Components in PTSD data from four different sets of Principal Components. Plots on bottom row are from the same data excluding the individual with race Other.

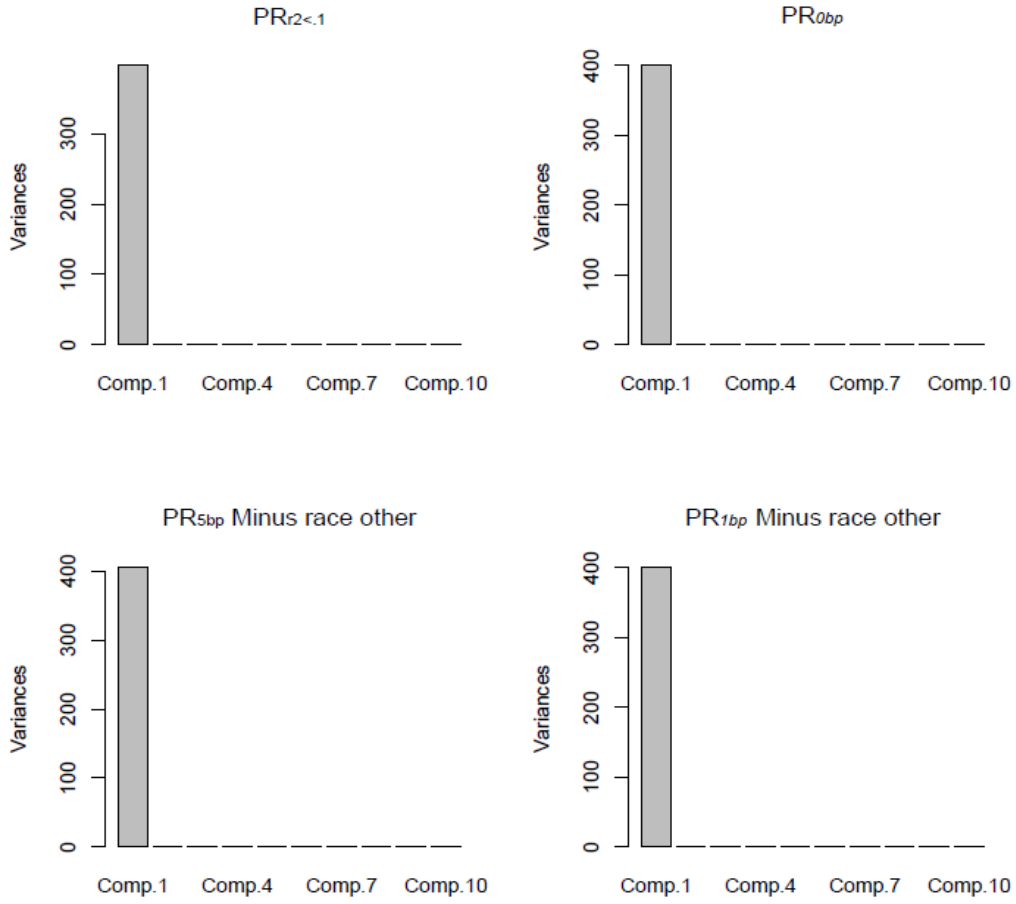

Figure 3: Normalized Principal Components from pruned dataset of r^2<0.25. From dataset including the individual with race Other. The colors by race: Red signifies African Americans, green Whites, teal is Mixed Race, and purple is Other. The different colors by Chip signify different chips, and the colors in by column location signify different locations on the chip.

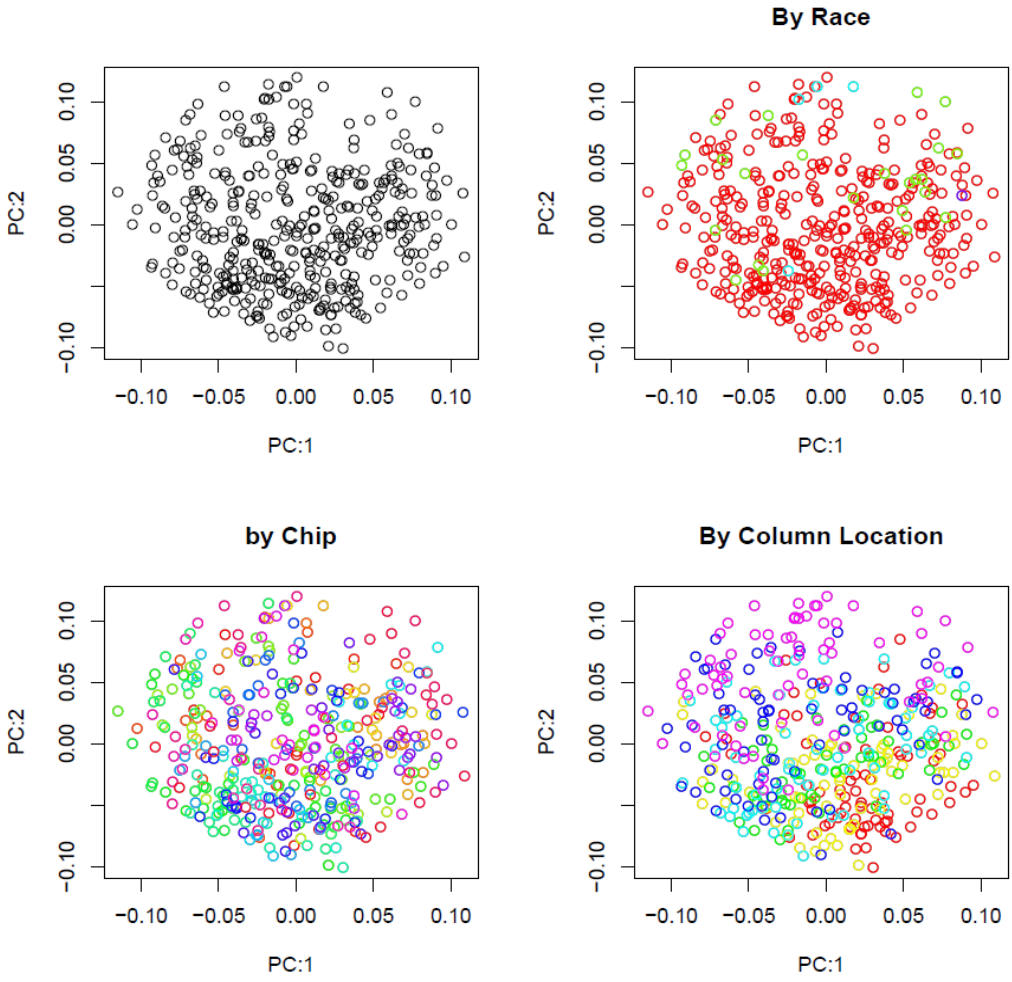

Figure 4: Boxplot of beta values from PTSD dataset (4 CpG Sites most associated with race).

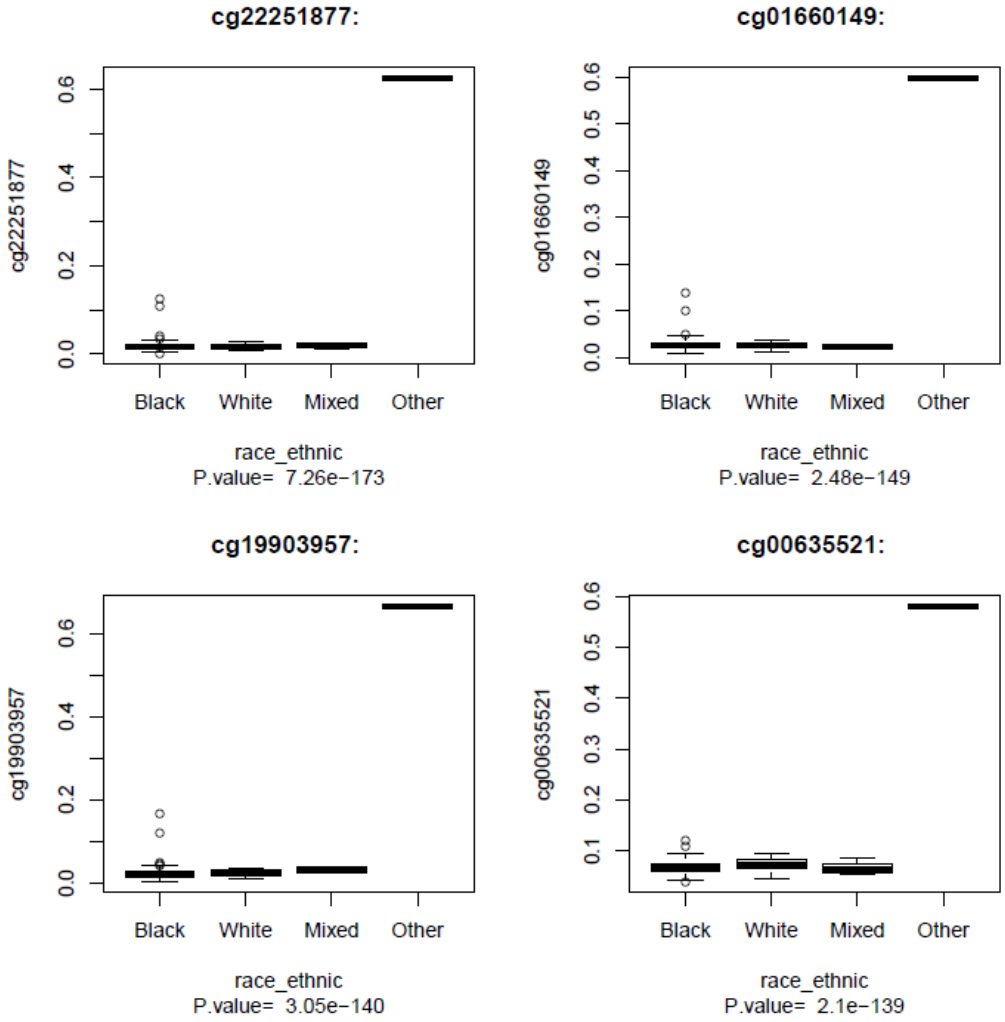

Figure 5: Boxplot of beta values from the CpG sites in Figure 4 (excluding individual with race Other).

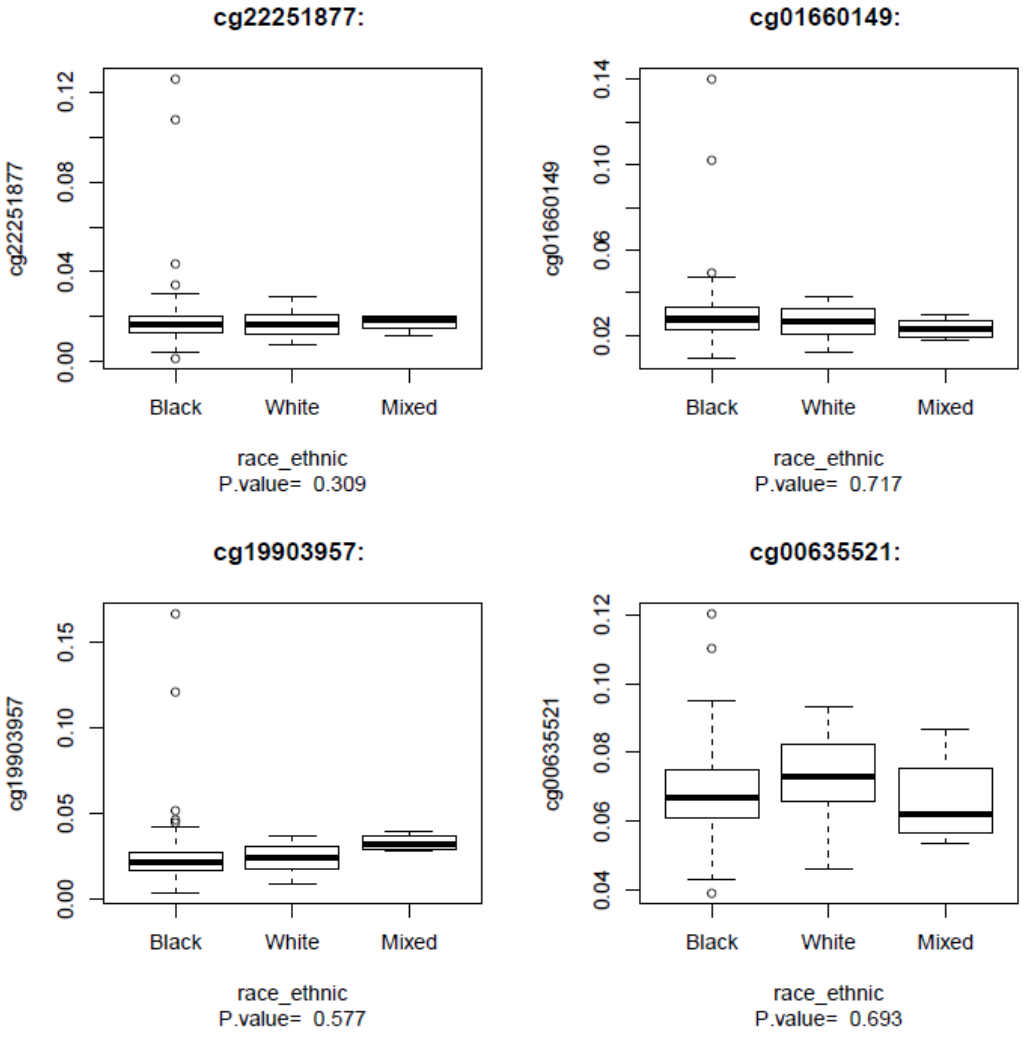

Figure 6: Boxplot of beta values of top four significant CpG sites with race (excluding individual with race Other).

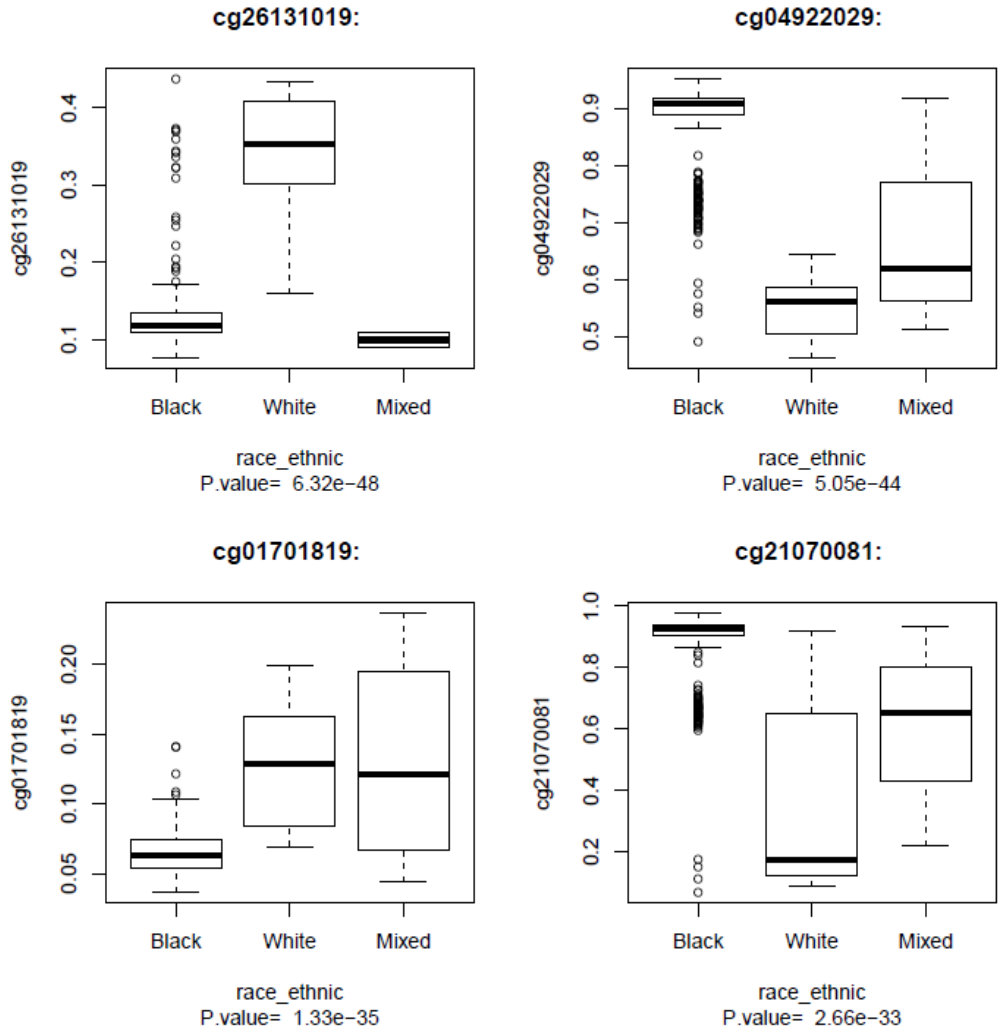

Figure 7: PTSD data, $\log_{10}$ of the distance from SNP of the Holm significant sites (excluding the individual with race Other). Distance is the distance from a known SNP of race-associated CpG sites in the four different analyses done on the data.