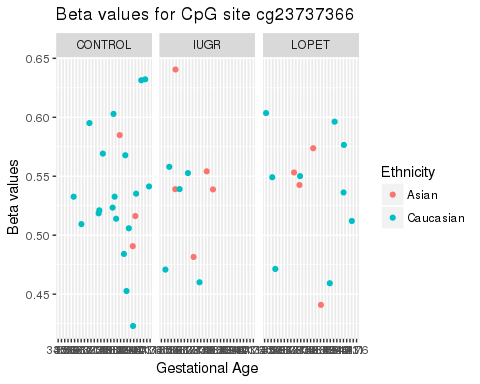
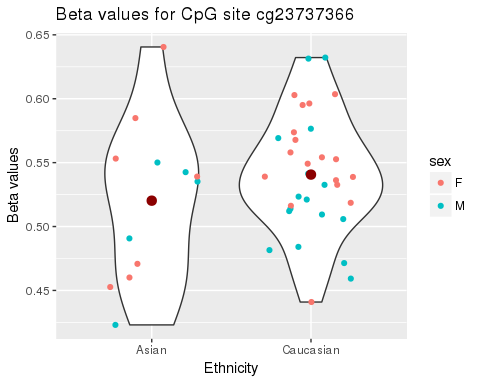
Exploratory Analysis of Processed Data

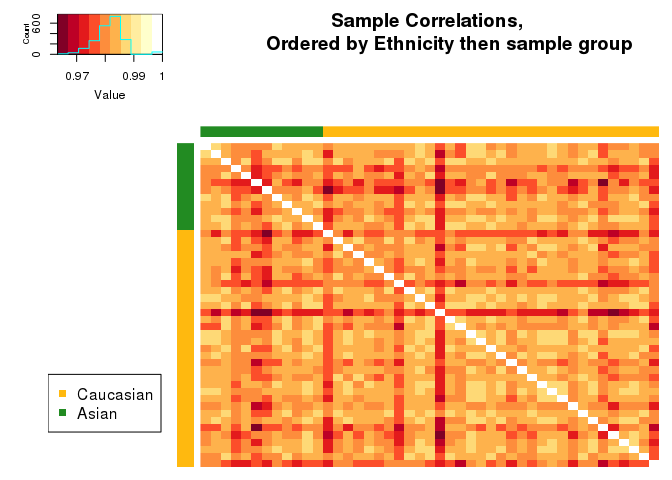
# Step 1: Exploratory Plots

### 1.1 Explore a random CpG Site



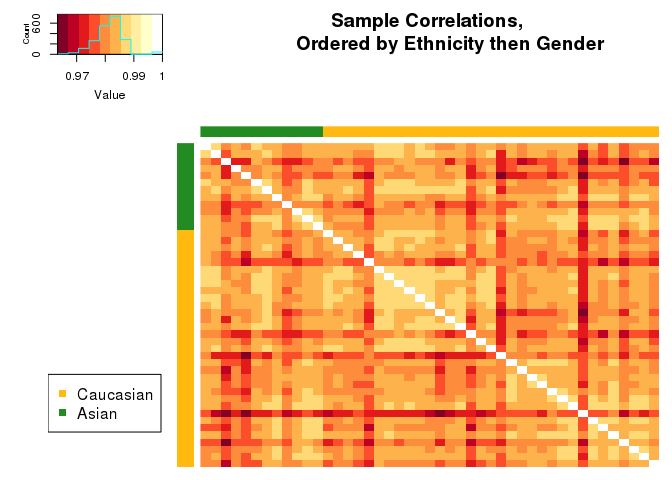


### 1.2 Sample to Sample Correlations



Sample group (Control, LOPET, IUGR) does not seem to affect the clustering of the samples. This fits with what it is known in the literature (control, LOPET, IUGR do not affect DNAm). Most of the samples seem to correlate evenly with each other.

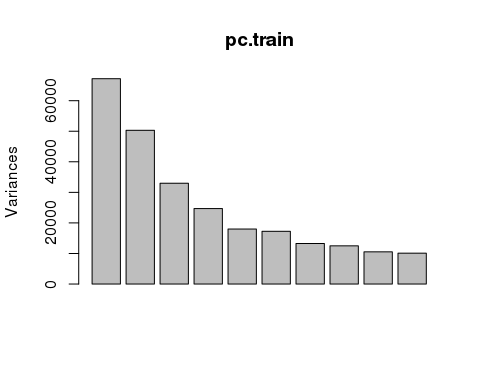
Now let's look at sample to sample correlation, ordered by ethnicity then gender.



It's difficult to tell if gender affects affects the clustering of the samples. We need to look at the data further.

# Step 2: Unsupervised clustering

### 2.1: PCA on training data:

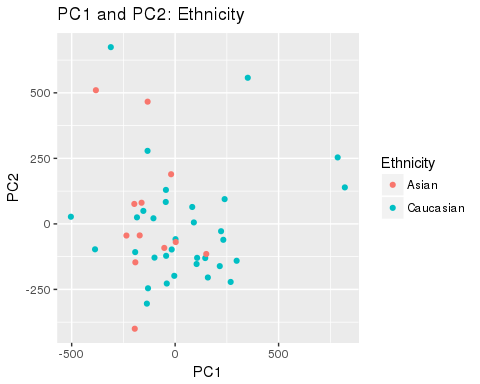


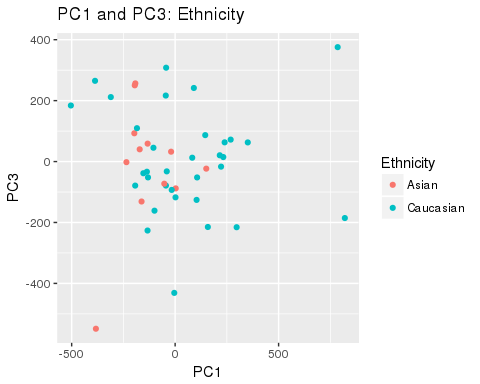
The plot() function returns a plot of the variances (y-axis) associated with the PCs (x-axis), which is useful to decide how many PCs to retain for further analysis.

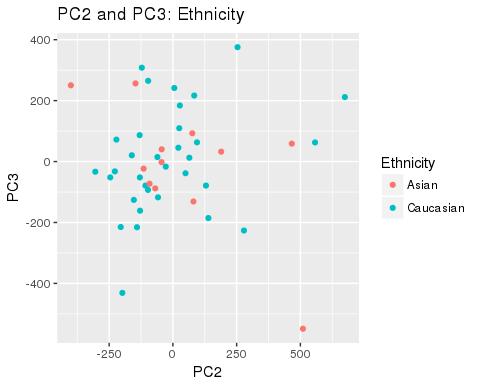
summary(pc.train)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 259.2041 224.2972 181.58732 157.16574 134.1441  
## Proportion of Variance 0.1445 0.1082 0.07092 0.05313 0.0387  
## Cumulative Proportion 0.1445 0.2527 0.32364 0.37677 0.4155  
## PC6 PC7 PC8 PC9 PC10  
## Standard deviation 131.3335 115.19038 111.78600 102.59258 100.45597  
## Proportion of Variance 0.0371 0.02854 0.02688 0.02264 0.02171  
## Cumulative Proportion 0.4526 0.48112 0.50800 0.53063 0.55234  
## PC11 PC12 PC13 PC14 PC15 PC16  
## Standard deviation 96.94551 95.14546 91.65346 90.93416 88.3696 87.3287  
## Proportion of Variance 0.02022 0.01947 0.01807 0.01779 0.0168 0.0164  
## Cumulative Proportion 0.57256 0.59203 0.61009 0.62788 0.6447 0.6611  
## PC17 PC18 PC19 PC20 PC21  
## Standard deviation 85.7165 83.17625 81.48632 81.13386 80.74645  
## Proportion of Variance 0.0158 0.01488 0.01428 0.01416 0.01402  
## Cumulative Proportion 0.6769 0.69176 0.70605 0.72021 0.73423  
## PC22 PC23 PC24 PC25 PC26  
## Standard deviation 79.31894 78.90655 77.92852 77.06075 76.8396  
## Proportion of Variance 0.01353 0.01339 0.01306 0.01277 0.0127  
## Cumulative Proportion 0.74776 0.76115 0.77422 0.78699 0.7997  
## PC27 PC28 PC29 PC30 PC31  
## Standard deviation 75.96452 75.71976 75.45761 75.41986 74.56784  
## Proportion of Variance 0.01241 0.01233 0.01225 0.01223 0.01196  
## Cumulative Proportion 0.81210 0.82443 0.83668 0.84891 0.86087  
## PC32 PC33 PC34 PC35 PC36 PC37  
## Standard deviation 73.59897 72.7979 72.30052 71.8361 71.36743 71.23120  
## Proportion of Variance 0.01165 0.0114 0.01124 0.0111 0.01096 0.01091  
## Cumulative Proportion 0.87252 0.8839 0.89517 0.9063 0.91722 0.92813  
## PC38 PC39 PC40 PC41 PC42  
## Standard deviation 70.42989 70.29780 69.93740 69.32554 68.82778  
## Proportion of Variance 0.01067 0.01063 0.01052 0.01034 0.01019  
## Cumulative Proportion 0.93880 0.94943 0.95995 0.97029 0.98048  
## PC43 PC44 PC45  
## Standard deviation 67.80424 66.91702 5.257e-13  
## Proportion of Variance 0.00989 0.00963 0.000e+00  
## Cumulative Proportion 0.99037 1.00000 1.000e+00

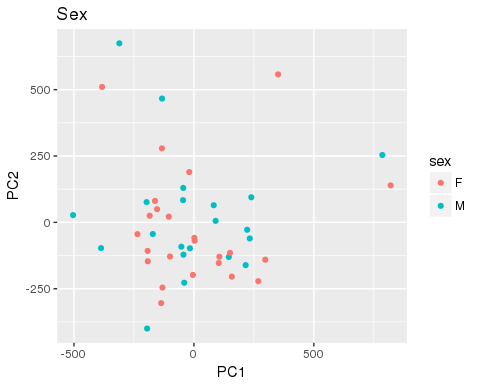
The summary() function describes the importance of the PCs. The first row describe again the standard deviation associated with each PC. The second row shows the proportion of the variance in the data explained by each component while the third row describe the cumulative proportion of explained variance.

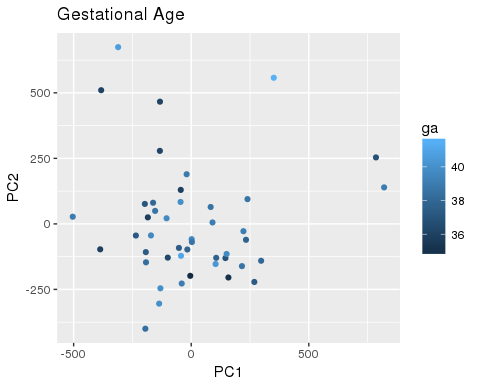


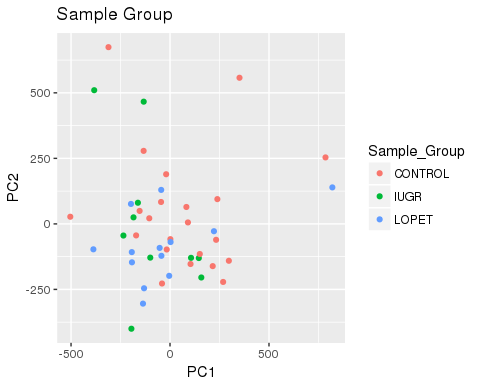




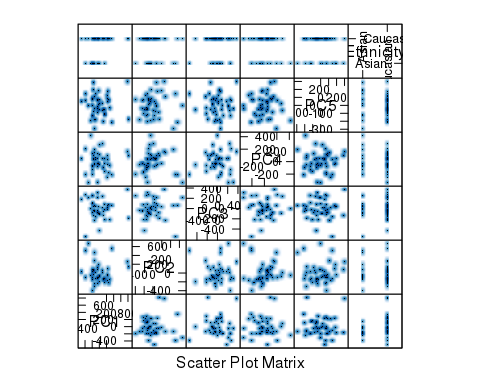
We can see from plotting the first three principal components that our groups (Asian, Caucasian) do not seem to separate. This indicates that the main drivers of the variance in the data is something else.







It's not clear that our other variables are driving the variance in the data (sex, gestational age, and sample group).



Plotting scatter plots of the top 5 PCs against ethnicity, none of the PCs can clearly separate samples by ethnicity, disappointing.