Supplementary Methods

Quantile normalization of methylation signal data

We quantile normalized the methylated and unmethylated signals together. We first ordered the combined signals from lowest to highest for each individual. We then replaced each value with the respective mean of that ordered value across all individuals. 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 of $x_{(i),k}$ would be:

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

For example, if A is our signal matrix:

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

we then order the values for each individual:

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

take the average across the ordered values:

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

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

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

We then calculated the new β values 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 β value for individual 1, site 1, would be computed as 540.75/(339.25+540.75)

Correlation-based pruning of methylation 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 B be our matrix of DNA methylation β values, with each row representing a CpG site and each column an individual.
2) Set any missing values in B equal to the mean for that CpG site.
3) Let R be the correlation matrix of B, with the diagonal set to zero:

$$\begin{bmatrix} 0 & r_{2,1} & \cdots & r_{5000,1} \\ r_{2,1} & 0 & \cdots & r_{5000,2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{5000,1} & r_{5000,2} & \cdots & 0 \end{bmatrix}$$

where $r_{i,j}$ represents the correlation between the i^{th} and j^{th} CpG site.

4) For each site we then calculate the number of connections, where a connection is defined as a squared correlation above 0.25.

$$v = (v_1, v_2, \ldots, v_{4999}, v_{5000})$$

where :

$$v_i = \sum_{j=1}^{5000} I(r_{i,j}^2 > .25)$$

and *I* is the indicator function.

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

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

$$v^* = (v_1^*, v_2^*, \ldots, v_n^*)$$
$$r^* = (r_1^*, r_2^*, \ldots, r_n^*)$$

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 the absolute values 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.

7) 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.

8) 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 r^2>.25 level within each chromosome. In a similar fashion, we also defined a set of CpG sites pruned at the r^2>.1 level.

Supplementary Table I: Total number of sites associated with race, before and after correction for population stratification with M-values. Principal components were computed based on the M-values.

Correction method used	# markers used to compute PCs	# FDR-significant CpG sites	# Holm-significant CpG sites	λ_{GC}
No correction	-	13440	856	2.10
GC	-	487	75	1
PC_{GWAS}	54,610	0	0	1
PC_{GWAS_TW}	54,610	0	0	1
$PC_{unpruned}$	469,142	52	2	1.33
$PC_{r2<0.25}$	225,440	0	0	1.10
$PC_{r2<0.1}$	121,855	0	0	1.06
PC_{0bp}	7,326	0	0	1.34
PC_{1bp}	17,105	1	1	1.33
PC_{2bp}	20,336	2	0	1.38
PC_{5bp}	31,178	6	0	1.45
PC_{10bp}	48,998	4	0	1.31
PC_{50bp}	174,510	0	0	1.15
PC_{100bp}	271,877	5	1	1.19

Supplementary Table II: Principal component (p-value, Adjusted R^2) most associated with the phenotype of interest. "None" indicates that no p-values in the top ten principal components were below 0.005.

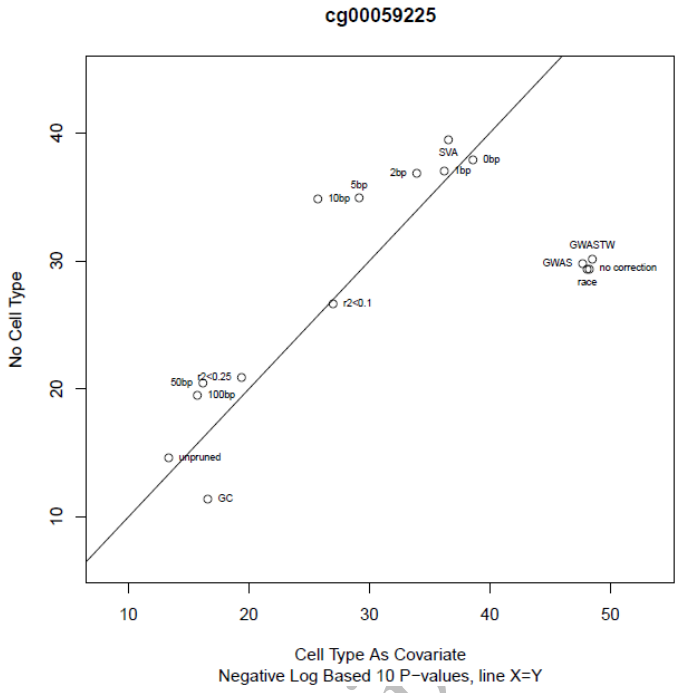
Note that a similar analysis was performed for smoking, but no PCs were significant.

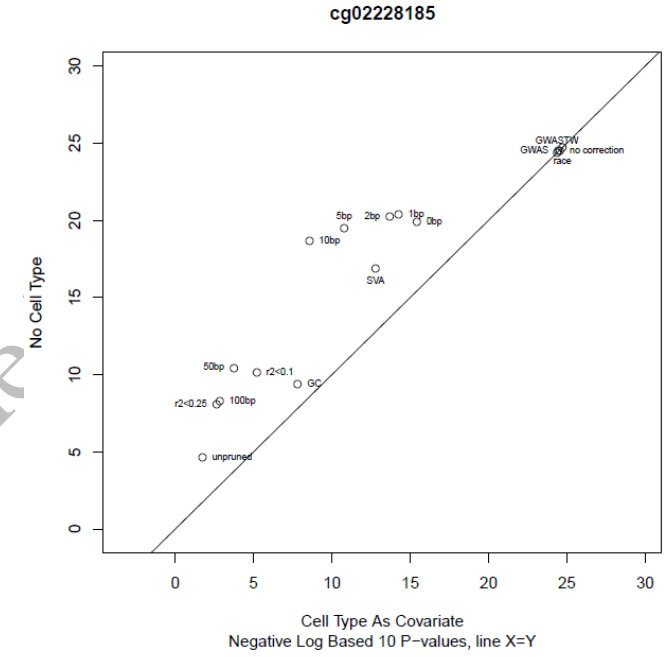
Approach	Race	Age	Row on chip	Chip	Six cell type proportions
$PC_{unpruned}$	7 (3.17×10^{-86}, 0.63)	4 (3.23×10^{-21}, 0.20)	3 (1.88×10^{-28}, 0.30)	6 (5.86×10^{-31}, 0.48)	2 (1.76×10^{-201}, 0.91)
$PC_{r2<0.25}$	4 (8.77×10^{-93}, 0.66)	7 (1.38×10^{-24}, 0.24)	2 (2.30×10^{-27}, 0.29)	10(3.02×10^{-20}, 0.39)	3 (8.03×10^{-115}, 0.75)
$PC_{r2<0.1}$	2 (2.02×10^{-39} ,0.36)	7 (5.75×10^{-26}, 0.25)	2 (7.87×10^{-25}, 0.27)	6 (1.78×10^{-21}, 0.40)	4 (3.05×10^{-58}, 0.51)
PC_{100bp}	6 (1.63×10^{-103},0.70)	4 (1.69×10^{-21}, 0.21)	3 (2.87×10^{-30}, 0.32)	7 (8.41×10^{-32}, 0.49)	2 (6.02×10^{-199}, 0.91)
PC_{50bp}	6 (3.72×10^{-40}, 0.37)	4 (6.72×10^{-15}, 0.14)	3 (9.76×10^{-32}, 0.33)	7 (7.24×10^{-32}, 0.49)	2 (7.5×10^{-199}, 0.91)
PC_{10bp}	3 (6.85×10^{-53} ,0.46)	5 (1.31×10^{-10}, 0.10)	3 (3.22×10^{-26}, 0.28)	8 (1.18×10^{-27}, 0.45)	2 (6.68×10^{-171}, 0.87)
PC_{5bp}	3 (9.68×10^{-68} ,0.54)	6 (1.09×10^{-12}, 0.12)	4 (2.00×10^{-18}, 0.21)	8 (4.18×10^{-25}, 0.43)	2 (3.33×10^{-96}, 0.69)
PC_{2bp}	2 (4.74×10^{-87} ,0.64)	6 (1.67×10^{-11}, 0.11)	4 (4.03×10^{-21}, 0.24)	4 (2.00×10^{-17}, 0.36)	3 (9.99×10^{-85}, 0.65)
PC_{1bp}	2 (1.57×10^{-108},0.72)	6 (5.26×10^{-10}, 0.092)	4 (6.74×10^{-22}, 0.25)	4 (8.23×10^{-18}, 0.36)	3 (5.36×10^{-107}, 0.73)
PC_{0bp}	2 (4.68×10^{-82} ,0.62)	6 (3.82×10^{-08}, 0.075)	4 (1.63×10^{-23}, 0.26)	4 (1.50×10^{-17}, 0.36)	3 (2.17×10^{-82}, 0.63)
PC_{GWAS}	1 (7.52×10^{-162},0.85)	None	1 (3.79×10^{-08}, 0.11)	None	None

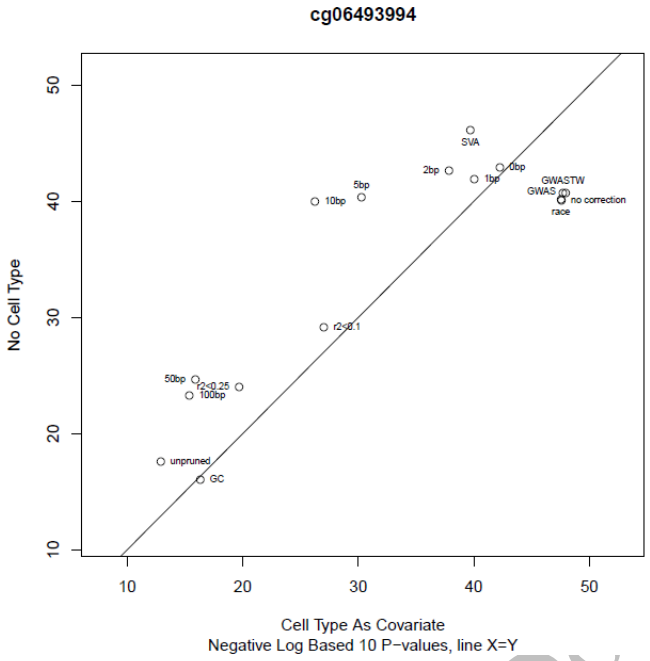
Supplementary Table III: Type I error rate and power for association of M-values with a continuous trait, by method of correction for population stratification

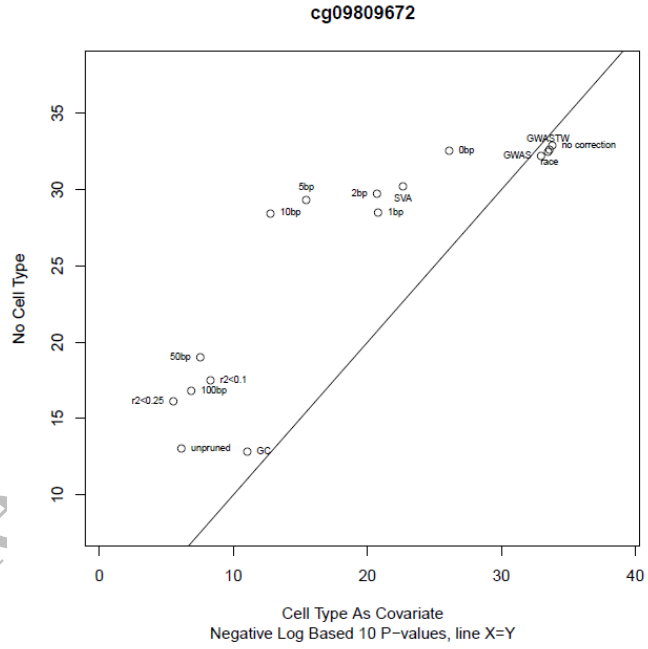
	Rate of type I error		Power	
Correction method	**No population stratification**	**Stratification present**	**No population stratification**	**Stratification present**
No correction	0.0394	0.2562	0.955	---
Race included as covariate	0.0346	0.0346	0.946	0.946
GC	0.0134	0.0742	0.910	0.656
PC_{GWAS}	0.0374	0.0358	0.867	0.855
PC_{GWAS_TW}	0.0350	0.0328	0.949	0.938
$PC_{unpruned}$	0.0484	0.0530	0.876	0.861
$PC_{r2<0.25}$	0.0514	0.0532	0.873	0.853
$PC_{r2<0.1}$	0.0524	0.0530	0.870	0.849
PC_{0bp}	0.0404	0.0472	0.836	0.828
PC_{1bp}	0.0416	0.0466	0.861	0.834
PC_{2bp}	0.0468	0.0454	0.861	0.840
PC_{5bp}	0.0494	0.0510	0.864	0.845
PC_{10bp}	0.0508	0.0498	0.868	0.856
PC_{50bp}	0.0518	0.0514	0.874	0.864
PC_{100bp}	0.0504	0.0510	0.875	0.863
SVA	0.0487	0.0497	0.769	0.737

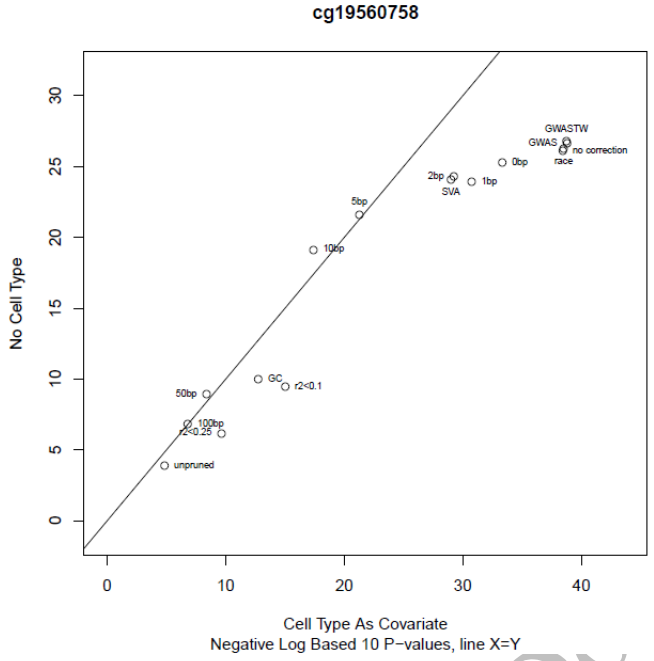
Supplementary Figure I: P-values shown for age-associated CpG sites with (X-axis) and without (Y-axis) cell type included as a covariate.

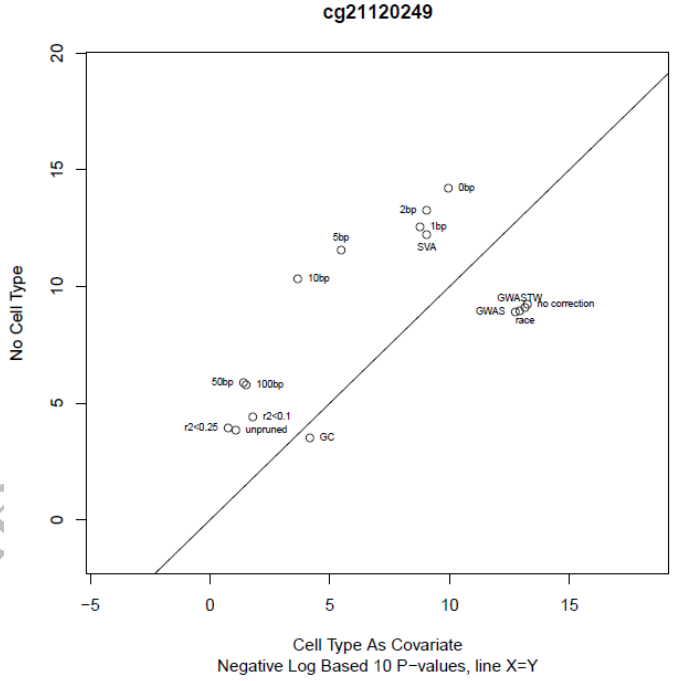


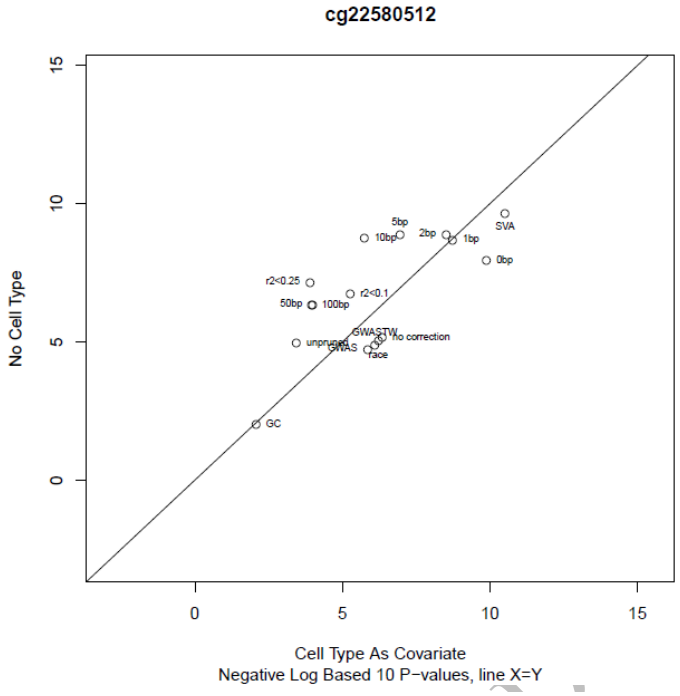


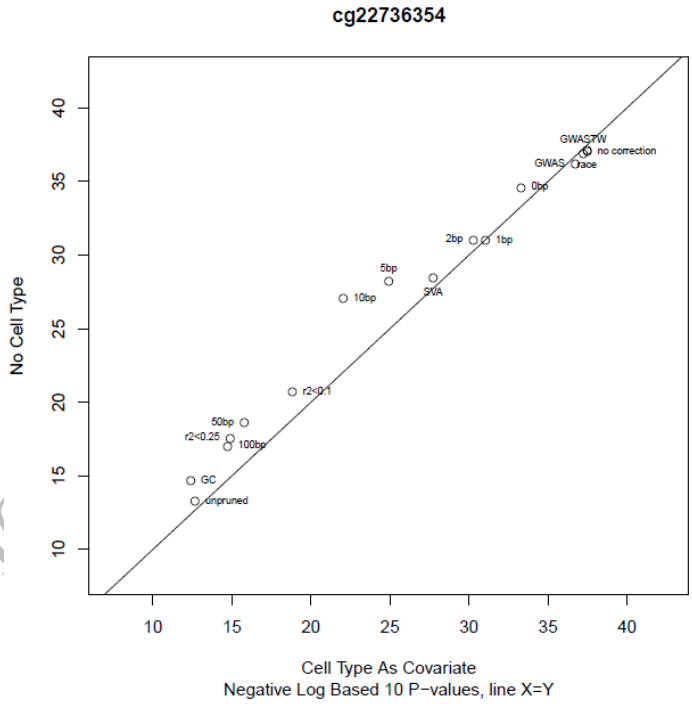




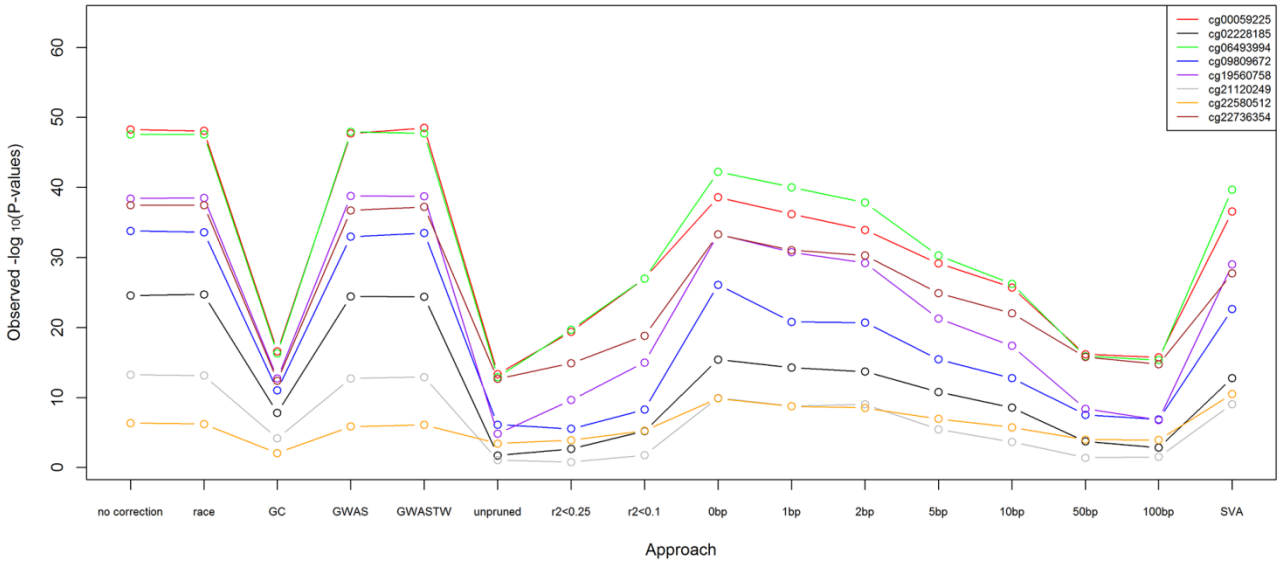








Supplementary Figure II: P-values for association between methylation and age, with estimated cell type proportions included as covariates.



Supplementary Figure III: P-values for association between methylation and smoking, with estimated cell type proportions included as covariates.