Supplementary Notes and Figures

List of Figures

Supplementary Note 1: Algorithm for whole-sample ICC method based on linear mixed effects model

Suppose we have m independent samples measuring the methylation of p CpGs. Assume each sample replicates $n_i(i{=}1,\ldots,m)$ times, totaling $n{=}\sum_{i=1}^{m} n_i$ samples. Note in most studies, the majority of the samples are not replicated and the majority of $n_i{=}1$. Denote y_{ij} as the methylation M-value of a given CpG for ith sample and its jth replicate after data normalization. We model y_{ij} using a linear mixed effects model (LMM)

$$y_{ij} = \mu + \xi_i + \epsilon_{ij} \qquad i{=}1,\cdots,m \text{ and } j{=}1,\cdots,n_i, \tag{1}$$

where $\epsilon_{ij}{\sim}N(0,\sigma_\epsilon^2)$ represents technical variability and $\xi_i{\sim}N(0,\sigma_\xi^2)$ represents biological variability. Denote $\mathbf{y}_i = (y_{i1},\cdots,y_{in_i})$ and $Y = (\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_m)^T$, we then have

$$Y \sim MVN\left(\boldsymbol{\mu}, V\right),$$

with the mean $\boldsymbol{\mu} = (\mu,\mu,\cdots,\mu)^T$ and the covariance matrix V

$$V = \sigma^2 \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_m \end{pmatrix}_{n\times n} \text{ and } A_i = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}_{n_i\times n_i},$$

where $\sigma^2{=}\sigma_\epsilon^2{+}\sigma_\xi^2$ is the total variance and $\rho{=}\frac{\sigma_\xi^2}{\sigma_\epsilon^2+\sigma_\xi^2}$ is the intra-class correlation coefficient (ICC). The score function of the log likelihood is given by

$$U_\rho = -\frac{1}{2}tr(V^{-1}\frac{\partial V}{\partial \rho}) + \frac{1}{2}(Y-\mathbf{1}\hat{\mu})^T V^{-1}\frac{\partial V}{\partial \rho}V^{-1}(Y-\mathbf{1}\hat{\mu}),$$

where $\mathbf{1}$ is a column vector of 1's and

$$\hat{\mu} = (\mathbf{1}^T V^{-1}\mathbf{1})^{-1}(\mathbf{1}^T V^{-1} Y)$$

is a function of ρ. The maximum likelihood estimate (MLE) of ρ is given by equating the score function with 0.

As we have far more independent samples than technical replicates (e.g. 1,000 independent samples, 20 technical replicates), $\hat{\mu}$ can be approximated

and treated as free of ρ. To see this, we can study the range of $\hat{\mu}$ at different ρ's. By simple algebra, we get

$$\hat{\mu} = \begin{cases} \frac{\sum_{i=1}^{m} y_{i1}}{m} & \text{if } \rho = 1 \\ \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_j} y_{ij}}{n} & \text{if } \rho = 0. \end{cases}$$

Since the number of replicates is small, $\hat{\mu}$ at the two extremes (perfectly correlated and no correlation) is very close and we can use either one as the estimate of μ and substitute it into the score equation. In fact, both of $\hat{\mu}$ are working independence estimates and are consistent. Similarly, we can estimate the total variance as

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{m}(y_{i1} - \hat{\mu})^2}{m}.$$

The inverse of V can be obtained analytically as

$$V^{-1} = \frac{1}{\sigma^2}\begin{pmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_m \end{pmatrix}_{n\times n}, \quad B_i = \begin{pmatrix} b_{i0} & b_{i1} & \cdots & b_{i1} \\ b_{i1} & b_{i0} & \cdots & b_{i1} \\ \vdots & \vdots & \ddots & \vdots \\ b_{i1} & b_{i1} & \cdots & b_{i0} \end{pmatrix}_{n_i\times n_i,}$$

where

$$b_{i0} = \frac{1 + (n_i - 2)\rho}{(1-\rho)\{1 + (n_i - 1)\rho\}} \quad \text{and} \quad b_{i1} = -\frac{\rho}{(1-\rho)\{1 + (n_i - 1)\rho\}}.$$

The trace part can be simplified as

$$-\frac{1}{2}tr(V^{-1}\frac{\partial V}{\partial \rho}) = -\frac{1}{2}\sum_{i=1}^{m}\{n_i(n_i - 1)b_{i1}\},$$

and

$$V^{-1}\frac{\partial V}{\partial \rho}V^{-1} = \frac{1}{\sigma^2}\begin{pmatrix} C_1 & & & \\ & C_2 & & \\ & & \ddots & \\ & & & C_m \end{pmatrix}_{n\times n}, \quad C_i = \begin{pmatrix} c_{i0} & c_{i1} & \cdots & c_{i1} \\ c_{i1} & c_{i0} & \cdots & c_{i1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{i1} & c_{i1} & \cdots & c_{i0} \end{pmatrix}_{n_i\times n_i,}$$

where

$$\begin{aligned} c_{i0} &= (n_i-1)(n_i-2)b_{i1}^2 + 2(n_i-1)b_{i0}b_{i1} \\ c_{i1} &= \{(n_i-1)(n_i-2)+1\}b_{i1}^2 + 2(n_i-2)b_{i0}b_{i1} + b_{i0}^2. \end{aligned}$$

Finally, the score function can be simplified as

$$U_\rho = -\frac{1}{2}\sum_{i=1}^{m}\{n_i(n_i-1)b_{i1}\} + \frac{1}{2}\sum_{i=1}^{m}(\frac{\mathbf{y}_i - \mathbf{1}\hat{\mu}}{\hat{\sigma}})^T C_i(\frac{\mathbf{y}_i - \mathbf{1}\hat{\mu}}{\hat{\sigma}}).$$

For restricted maximum likelihood estimate (REML), the score function is given by

$$U_\rho^R = -\frac{1}{2}tr(P\frac{\partial V}{\partial \rho}) + \frac{1}{2}(Y-\mathbf{1}\hat{\mu})^T V^{-1}\frac{\partial V}{\partial \rho}V^{-1}(Y-\mathbf{1}\hat{\mu}),$$

where $P = V^{-1} - V^{-1}\mathbf{1}(\mathbf{1}^T V^{-1}\mathbf{1})^{-1}\mathbf{1}^T V^{-1}$. The only difference is the trace part, which can be simplified as

$$-\frac{1}{2}tr(P\frac{\partial V}{\partial \rho}) = -\frac{1}{2}\sum_{i=1}^{m}\{n_i(n_i-1)b_{i1}\} + \frac{1}{2d}\sum_{i=1}^{m} n_i\{c_{i0} + (n_i-1)c_{i1}\},$$

where $d = \sum_{i=1}^{m}\{n_i(b_{i0}-b_{i1}) + n_i^2 b_{i1}\}$.
For both MLE and REML, since most of the terms in the summation are 0's (independent samples), it only involves samples with replicates. The computation can be vectorized and does not require any matrix multiplication and inversion. Simple uniroot finding algorithm such as Newton method can be applied to get the solution. It usually involves less than 10 iterations to get the solution of desired numerical precision. Assume the number of technical replicates is small compared to the sample size, the computational cost for each iteration is only $O(np)$, which is linear time in sample size and CpG number.

Supplementary Note 2: Effects of ICC-based filtering on type I error and power

Independent filtering is a procedure that filters out those tests that are less likely to show statistical significance, without even looking at their test statistic. Typically, this results in increased statistical power at the same type I error level. A good choice of independent filtering criterion is the one that is (1) statistically independent from the test statistic under the null hypothesis, and (2) correlated with the test statistic under the alternative hypothesis. Statistical validity (controlled type I error) relies on the first property, while increased power is the result of the second property (Bourgon *et al.*, 2010). ICC-based filtering has both properties. Under the null (non-differential CpG), the p-value is uniformly distributed in [0, 1] regardless of the ICC value. Under the alternative (differential CpG), however, the p-value is correlated with the ICC value. This is because, given a fixed effect size, differential CpGs with larger ICCs will appear more significant while those with small ICCs look more like the non-differential CpGs. Therefore, the ICC filter results in increased detection power by removing CpGs whose p-values are distributed more or less uniformly in [0, 1].

To demonstrate the above theoretical properties, we conducted additional simulations that mimic the real CpG methylation data. Based on the observation from Meng *et al.*, 2010 as well as our own data, the methylation array consists of a mixture of 'good' (high ICC) and 'bad' (low ICC) CpG probes, that is, probes can either measure methylation accurately or not at all. According to Meng *et al.*, 2010, the proportion of the 'bad' probes can be up to 60%. Therefore, we simulated the ICCs from a mixture of two components (low and high ICCs) with a mixing proportion of 0.5 (Figure S2A). We simulated 1,000 independent samples from two groups (500 each), and 10,000 CpGs, among which 10% were differentially methylated between the groups. The differential CpGs were randomly distributed among the 10,000 CpGs. Different numbers of technical replicate pairs (4, 6, 8, 10, 12) were simulated to investigate the effects of the number of replicate pairs. Simulation were repeated for 200 times. The detailed simulation setting was shown in R code at the end.

We estimated the ICCs based on the whole-sample or replicate-sample approach, and performed association tests using a simple linear model. We then used the estimated ICCs to filter the CpGs at different quantile cutoffs,

and applied both Bonferroni correction (BF) and false discovery rate (FDR) control (Benjamini-Hochberg (BH) procedure) to identify significant CpGs at the adjusted p-value cutoff of 0.05. Type I error (Family-wise error rate (FWER) for BF, and observed FDR for BH) and power (the number of true positives identified) were then calculated.

The distribution of the estimated ICCs had a spike at 0, and a bump near 1 (Figure S2B), which was similar to what we observed from the Normative Aging Study data set. Since MLE constrained the estimate to be non-negative, many of the estimates for low ICC probes were exact 0's, forming a spike at 0 (Figure S2B). Looking closely, the proposed method estimated the ICCs quite accurately for the high-ICC component (Figure S2C) while the estimates of low-ICC component had much larger variability, and the spike at 0 mainly came from this component (Figure S2D). For these Bonferroni-significant CpGs, they usually had high estimated ICCs (Figure S2E). For these non-significant CpGs, they could come from either high-ICC component or low-ICC component, resulting in a spike at 0 and a bump near 1 (Figure S2F).

As expected for an independent filtering procedure, the ICC-based filtering controlled the type I error at the desired level for both Bonferroni correction and false discovery control, and for both the whole-sample and replicate-sample ICC method, at different numbers of technical replicate pairs (Figure S3). Therefore, using the ICC criteria will not increase the chance of false findings.

By filtering CpGs that were less likely to be significant, we achieve a gain in power (Figure S4A). The power gained with the whole-sample ICC method was consistently higher than the replicate-sample ICC method, at all numbers of replicates, suggesting better ICC estimation by pooling all the samples. Bonferroni correction yielded larger increases in power than FDR control (15% vs 6%, whole-sample ICC, 12 technical replicate pairs). The largest gain in power for whole-sample ICC method was achieved between 0.4-0.5 quantile cutoff, which was consistent with the results from the Normative Aging Study. As the number of replicates decreased, the increase in power was also reduced. For Bonferroni correction, we achieved a good improvement in power even at six replicate pairs in our simulation and six replicate pairs were sufficient to outperform variance-based filtering in our Normative Aging Study application as well. However, for FDR control, the improvement was marginal for six technical replicate pairs, compared to the power of no filtering in our simulation. Therefore, we recommend that CpG

filtering be performed when at least six replicate pairs are available.

We also performed the the variance-based filtering. However, the method performed much worse than ICC-based methods. In many cases, it did not achieve any improvement in power at any quantile cutoff, compared to no filtering. This is due to the fact that larger variance is not necessarily correlated with the test statistic under the alternative hypothesis, and a large variance may be due to a large measurement error (small ICC) instead of the group difference. Therefore, the traditional variance-based filtering appears not to be suitable for CpG association studies, at least in the current simulation setting, and it has the potential problem of enriching for noisy CpGs (with low ICCs). Further investigation is needed to justify its use.

```
np ← 10000       # Number of CpG sites
ns ← 1000        # Number of samples
nr  ← 10         # Number of replicates
ntp ← 1000       # Number of differential CpGs
ntn ← 9000       # Number of non-differential CpGs
eff.m ← 0.2      # Mean of effect sizes
eff.sd ← 0.05    # SD of effect sizes
# Biological variability (vb)
vb ← rep(1.0, np)
# Techincial variability (ve)
ve ← sample(c(exp(rnorm(np * 0.5, −1.5, 0.5)),
    exp(rnorm(np * 0.5, 1.5, 0.5))))
Y1 ←  matrix(NA, np, ns)    # Non-replicated samples
Y21 ← matrix(NA, np, nr)    # Replicate sample 1
Y22 ← matrix(NA, np, nr)    # Replicate sample 2

for (j in 1:np) {
   if (j ≤ ntp) {
      # Differential CpGs
      eff ← rnorm(1, eff.m, eff.sd)
      eff ← c(rep(eff, ns/2), rep(0, ns/2))
      y1 ← rnorm(ns, eff, sqrt(vb[j])) +
      rnorm(ns,0,sqrt(ve[j]))
   } else {
       # Non-Differential CpGs
       y1 ← rnorm(ns,0,sqrt(vb[j])) +
       rnorm(ns,0,sqrt(ve[j]))
   }

```

```
30      y20 ← rnorm(nr, 0, sqrt(vb[j]))
31      y21 ← y20 + rnorm(nr, 0, sqrt(ve[j]))
32      y22 ← y20 + rnorm(nr, 0, sqrt(ve[j]))
33      Y1[j, ] ← y1
34      Y21[j, ] ← y21
35      Y22[j, ] ← y22
36  }
```

References

Meng, H., et al (2010) A statistical method for excluding non-variable CpG sites in high-throughput DNA methylation profiling, *BMC Bioinformatics*, **11**, 227.

Bourgon R., et al (2010). Independent filtering increases detection power for high-throughput experiments, *PNAS*, **107**, 21.

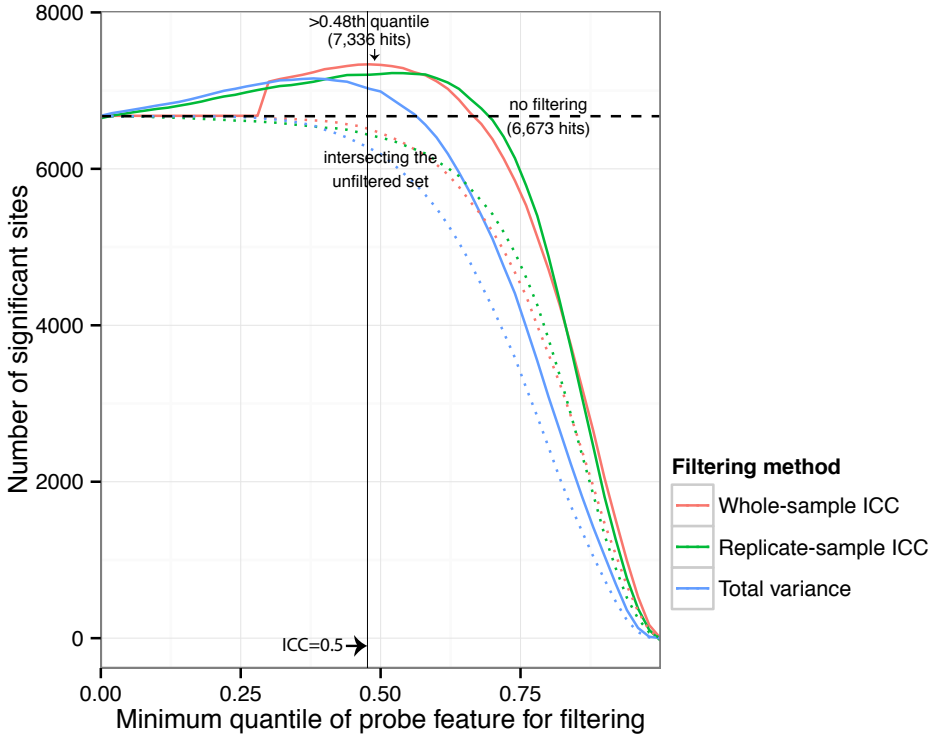

Figure S1: Effects of ICC cutoff values on the number of significant sites identified in the Normative Aging Study data set. CpG filtering was performed at different cutoffs using the three methods being compared. An epigenome-wide analysis of age was then performed on the subsets of filtered CpG sites, followed by Bonferroni correction. CpG sites with Bonferroni-adjusted p values less than 0.05 were declared significant hits. The x-axis is the stringency of the filtering cutoff, and the y-axis is the number of significant sites. The abrupt change of the whole-sample ICC method is due to the large number of probes (30%) with an estimated ICC of 0. The dashed horizontal line shows the 6,673 hits when using no filter. The dotted lines are the intersection of those 6,673 results with the hits from a given level of the filtering methods. The maximum number of significant sites achieved is from the whole-sample ICC method with 7,336 significant sites using a filter above the 0.48th quantile (analyzing only the 251,152 sites with an ICC above 0.52). This cutoff leads to a boost of an additional 10% of sites, compared to the analysis using all sites. At the 0.48th quantile cutoff, 6,508 sites overlapped with those discovered without CpG filtering. In comparison, the maximum number of significant sites achieved for replicate-sample ICC and total variance filter are 7,255 and 7,156 significant sites using a filter above the 0.52th and 0.38th quantile, respectively.

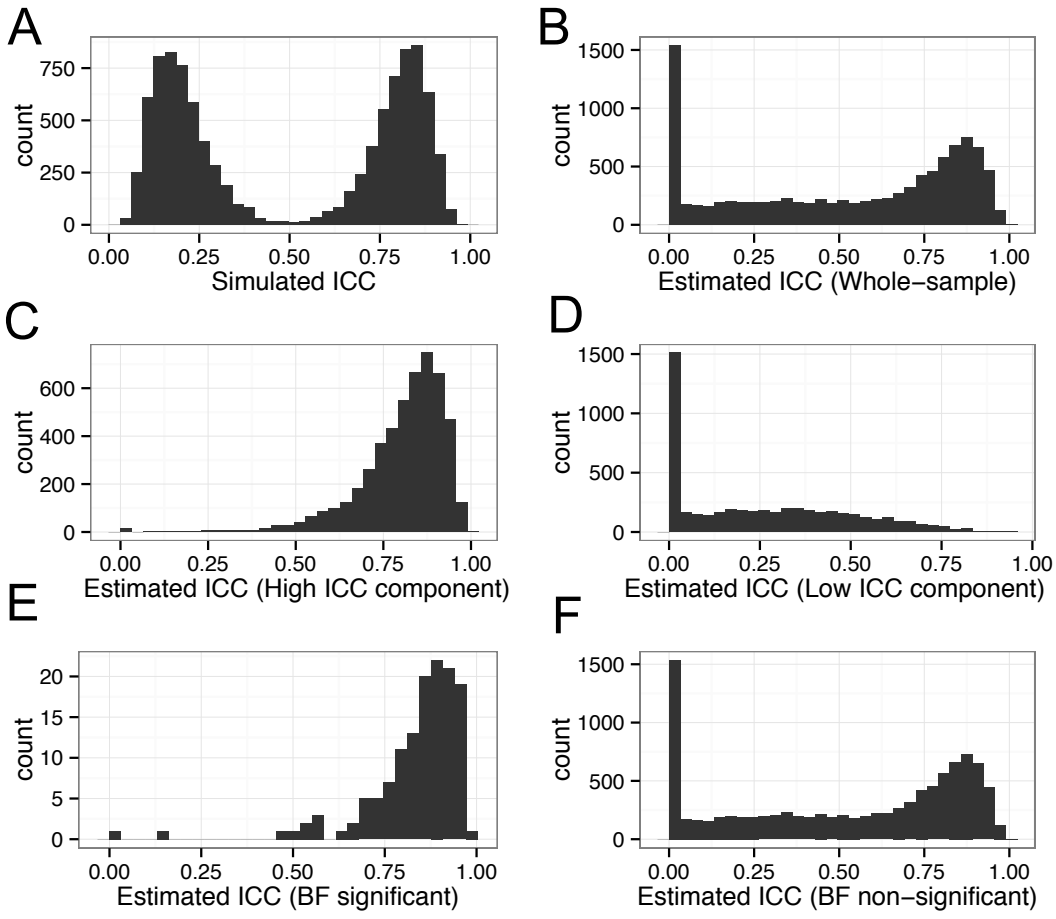

Figure S2: Distribution of simulated and estimated ICCs. Whole-sample ICC method was used with 10 replicate pairs. The performance of ICC estimation was compared on different subsets of ICCs.

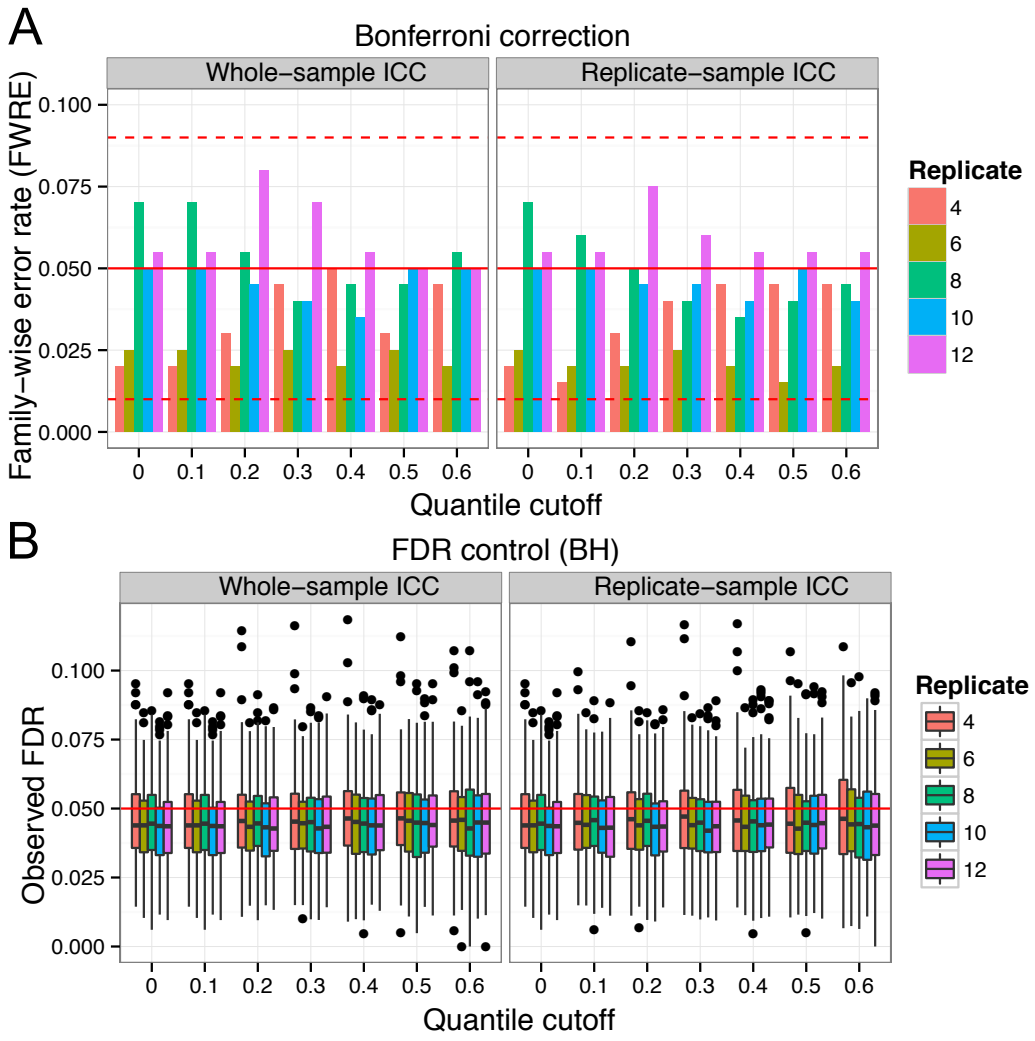

Figure S3: Effects of ICC-based CpG filtering on the type I error of association tests based on simulation. Both Bonferroni Correction (BF) and FDR control (Benjamini-Hochberg procedure (BH)) were investigated. FWER is defined as the proportion of the simulations that makes at least one false claim. Observed FDR is defined as the proportion of false positives in the claimed positives. The solid red line indicates the desired level, and dashed lines indicate the 95% confidence interval.

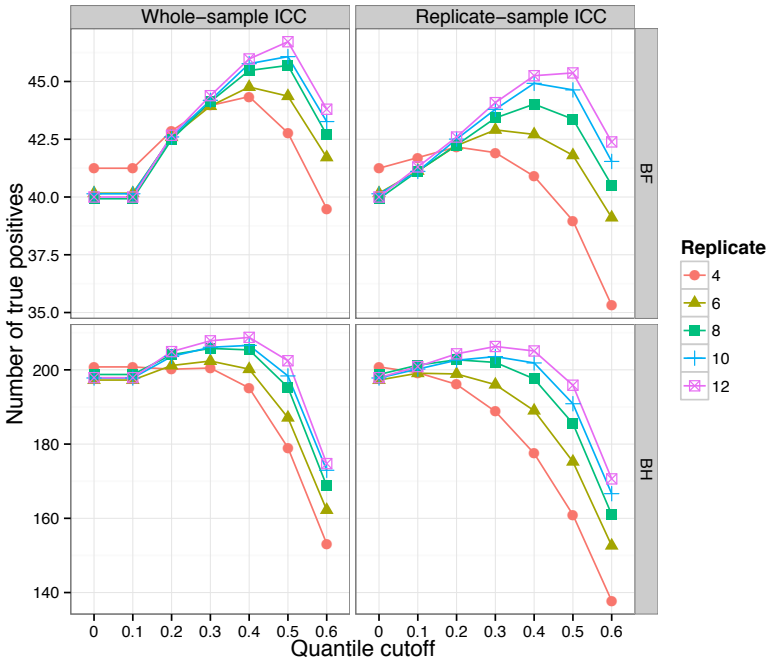

Figure S4: Comparison of the statistical power of whole-sample ICC and replicate-sample ICC method for CpG filtering based on simulation. Power (number of true positives identified) was evaluated at different numbers of replicates, and at different quantile cutoff values. Both Bonferroni Correction (BF) and FDR control (Benjamini-Hochberg procedure (BH)) were investigated.