Landes Bioscience www.landesbioscience.com

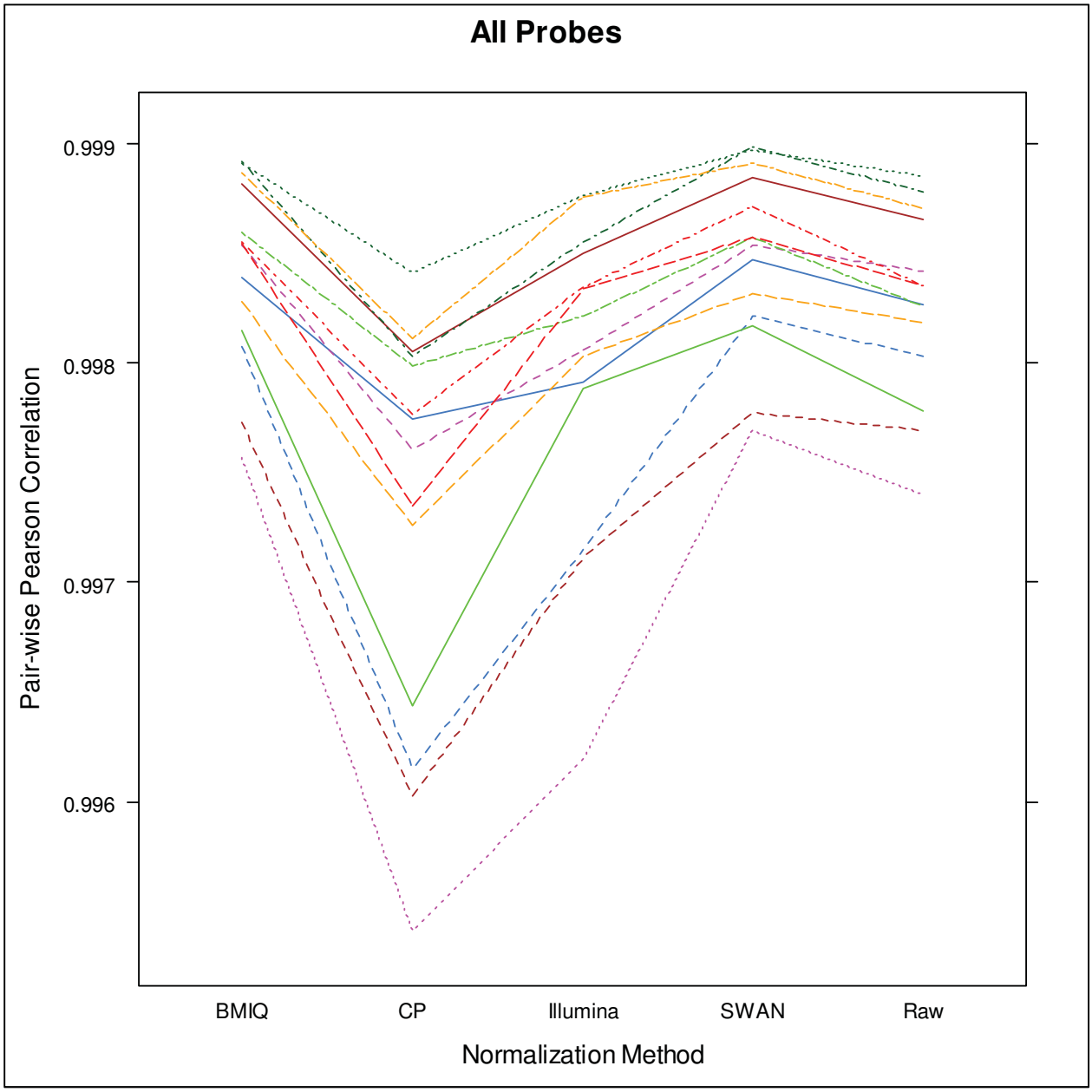
# Supplemental Material to:

Michael C Wu, Bonnie R Joubert, Pei-fen Kuan, Siri E Håberg, Wenche Nystad, Shyamal D Peddada, and Stephanie J London

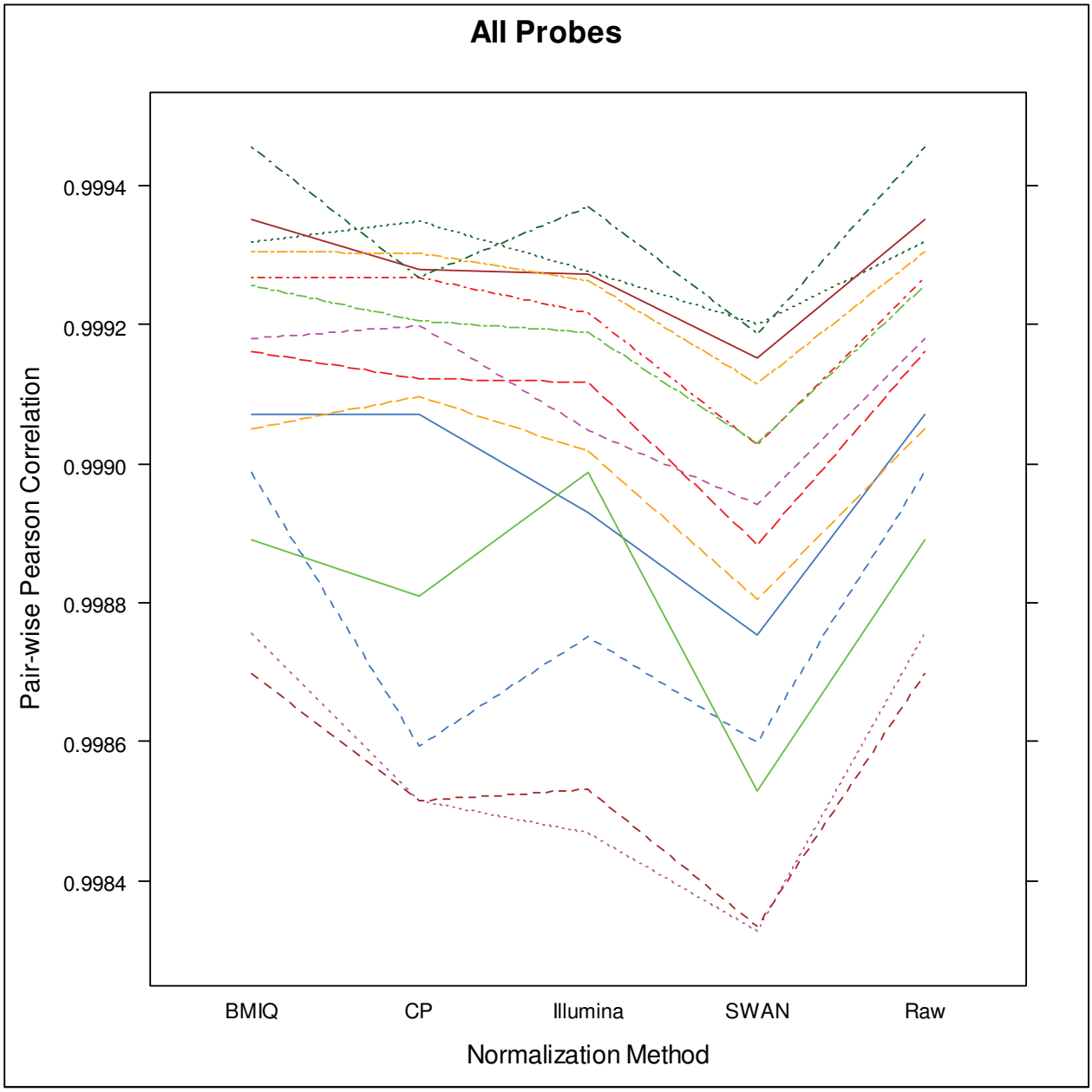
## A systematic assessment of normalization approaches for the Infinium 450k methylation platform

Epigenetics 2013; 9(2)
http://dx.doi.org/10.4161/epi.27119
http://www.landesbioscience.com/journals/epigenetics/article/27119/

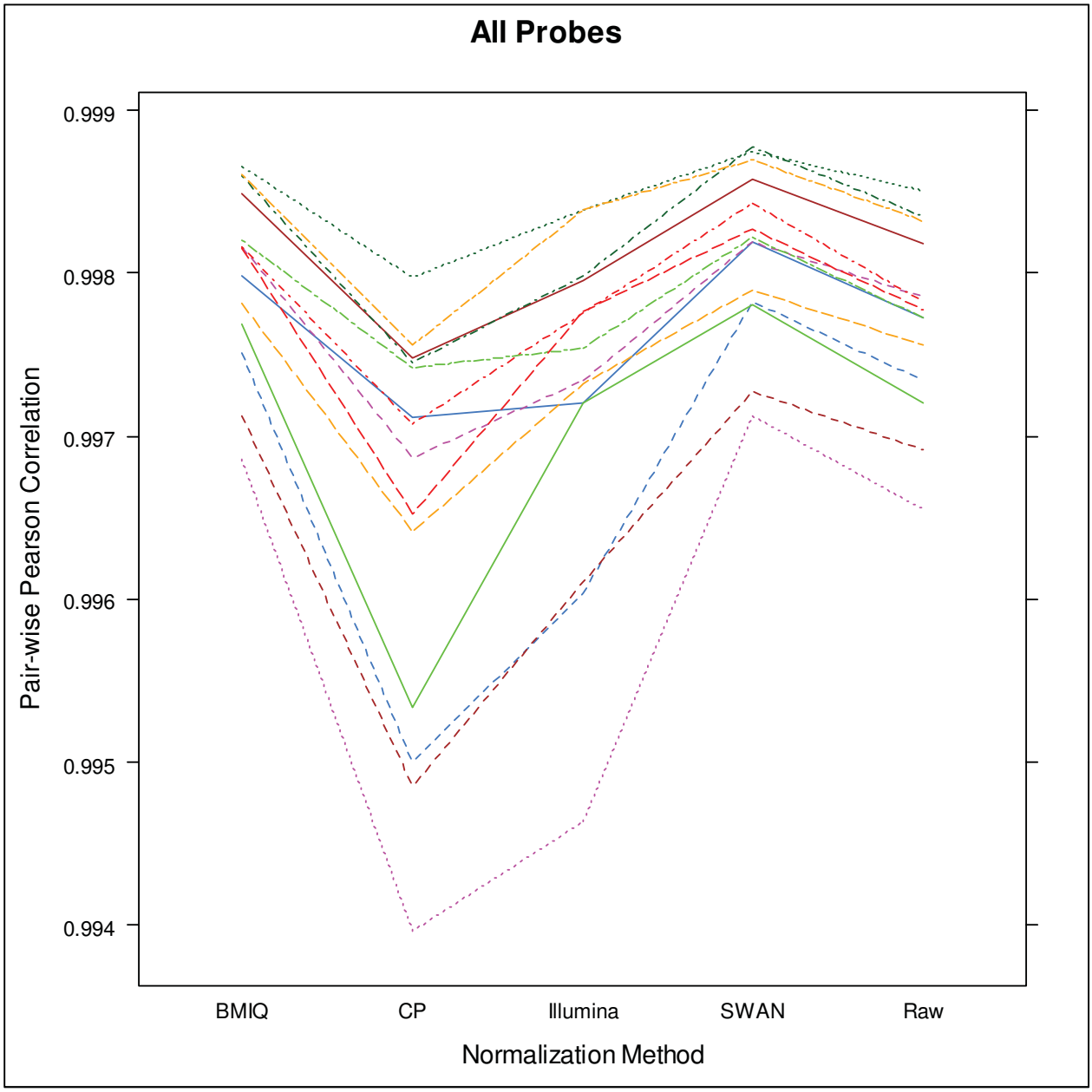
**Supplemental Figure 1.** Pairwise Pearson correlation between duplicate pairs following application of each normalization using all probes. Each line corresponds to a different duplicate pair.



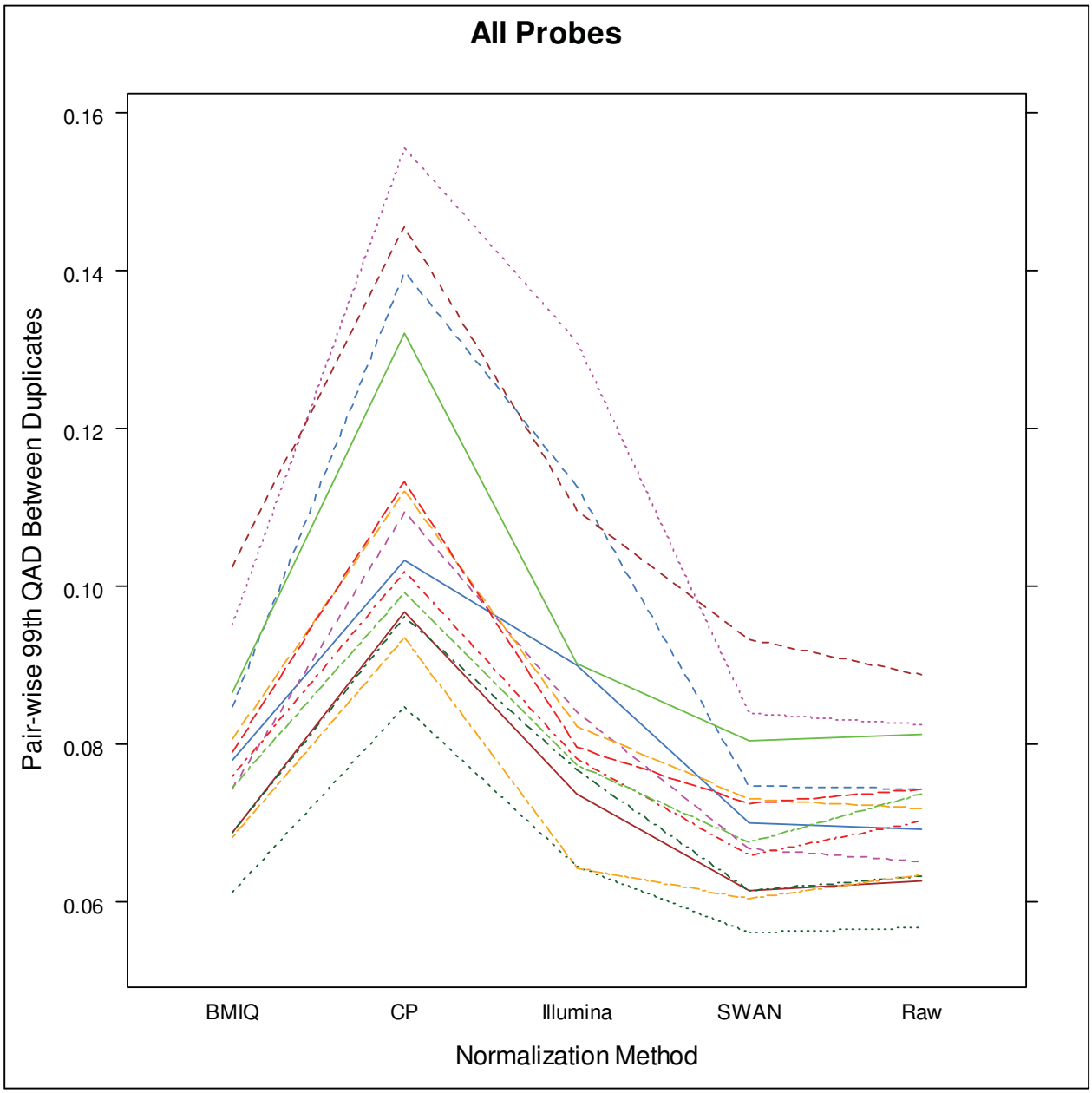
**Supplemental Figure 2.** Pairwise Pearson correlation between duplicate pairs following application of each normalization using just Type I probes. Each line corresponds to a different duplicate pair.



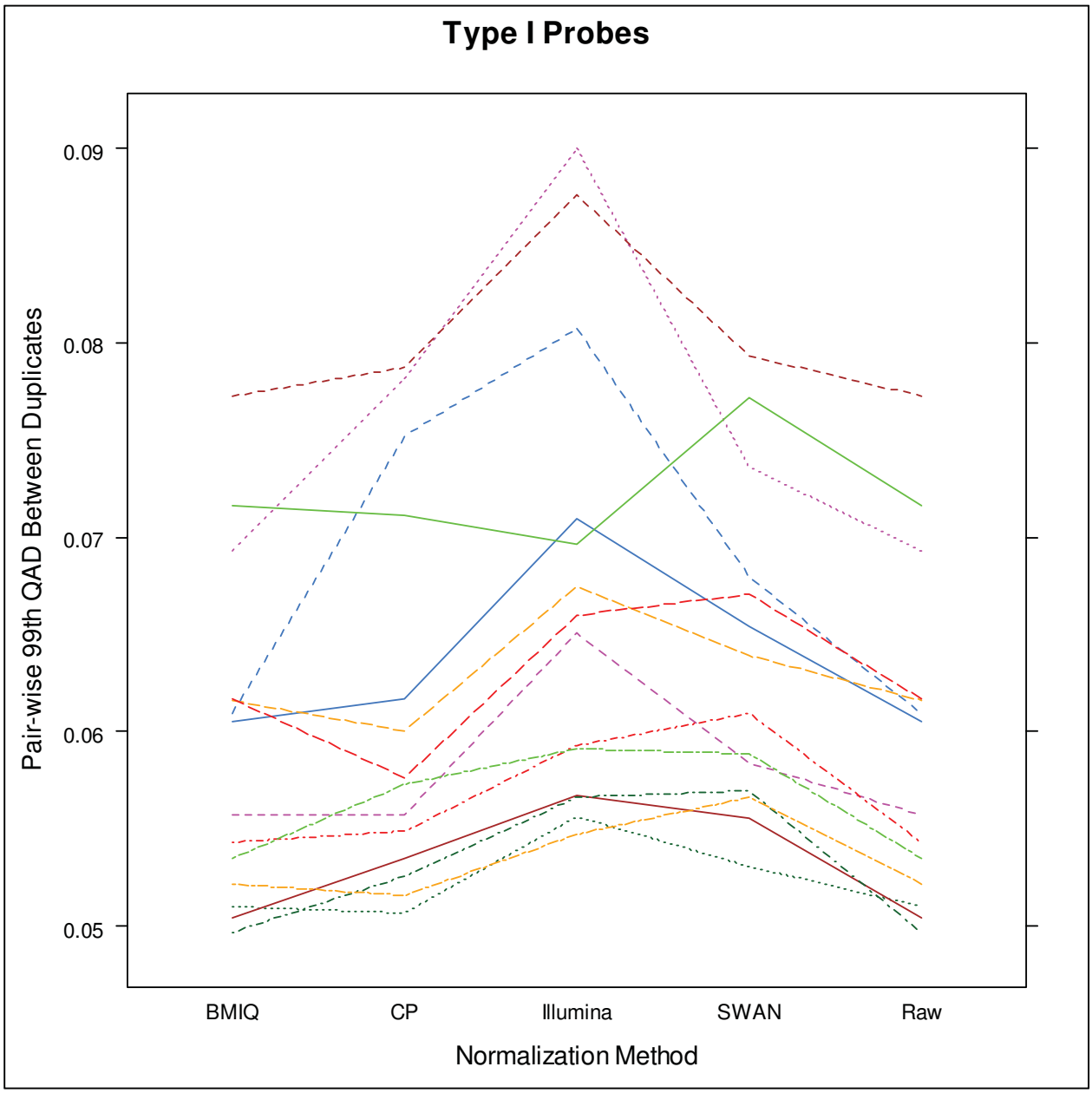
**Supplemental Figure 3.** Pairwise Pearson correlation between duplicate pairs following application of each normalization using just Type II. Each line corresponds to a different duplicate pair.



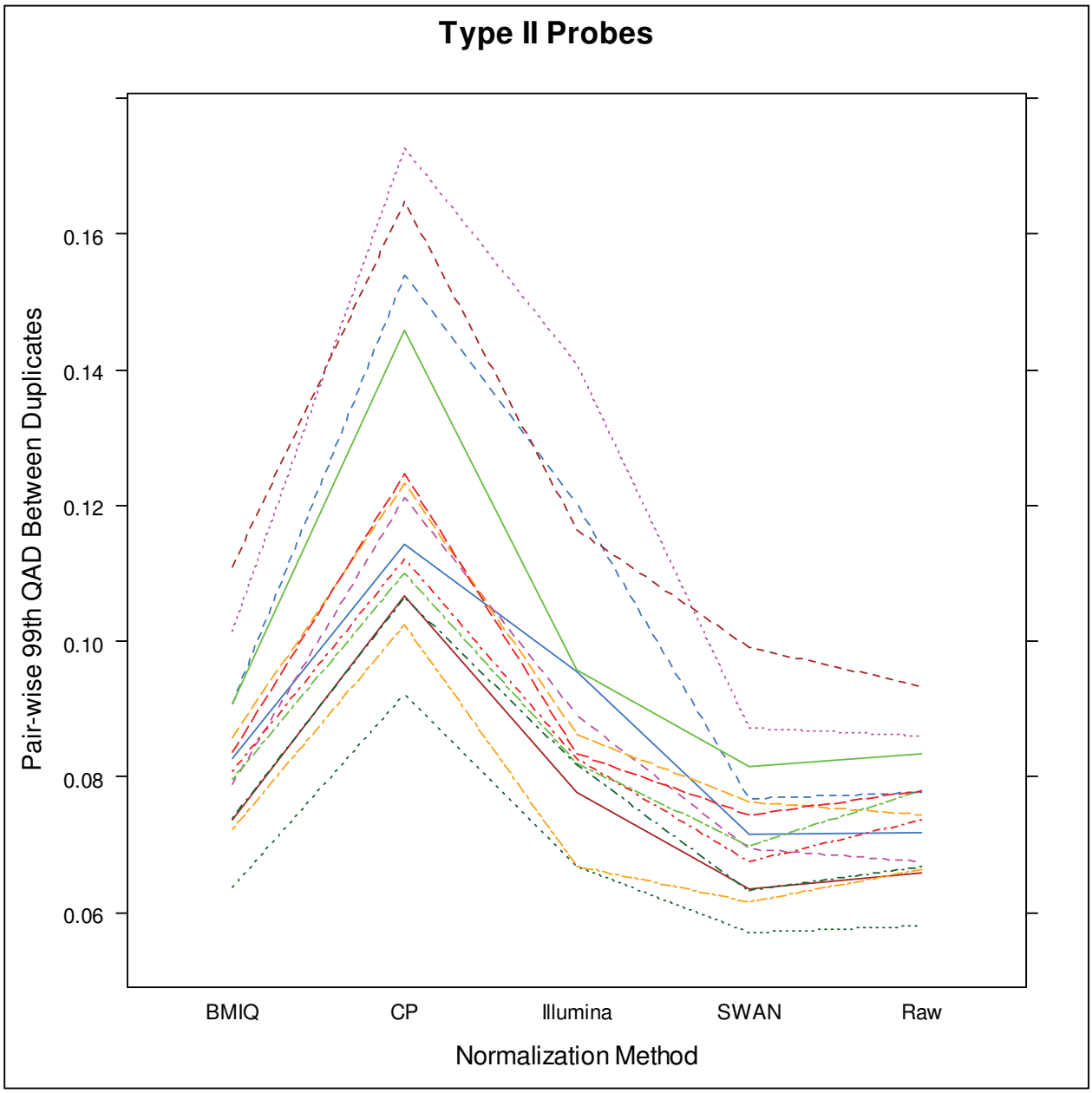
**Supplemental Figure 4.** Pairwise 99$^{th}$-QAD between duplicate pairs following application of each normalization using all probes. Each line corresponds to a different duplicate pair.



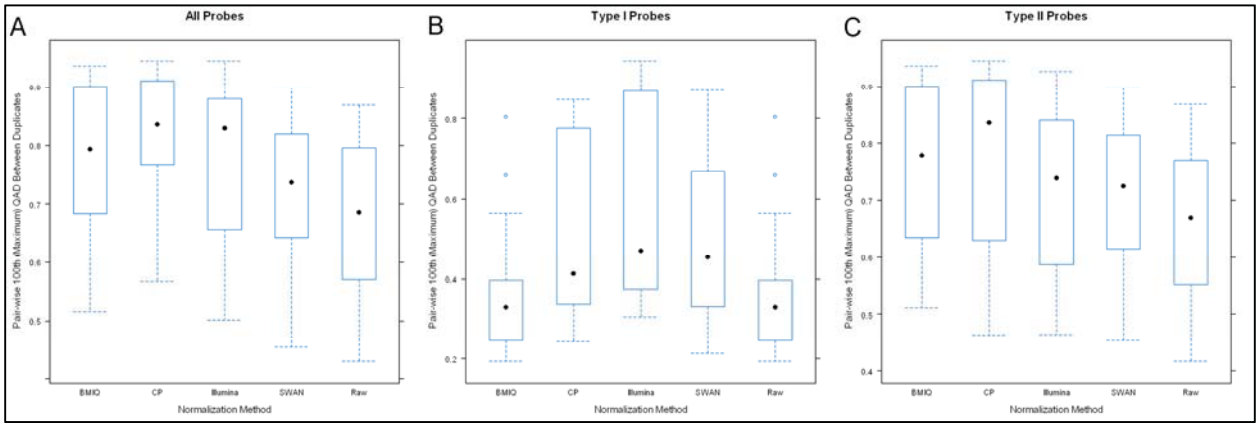
**Supplemental Figure 5.** Pairwise 99$^{th}$-QAD between duplicate pairs following application of each normalization using just Type I probes. Each line corresponds to a different duplicate pair.



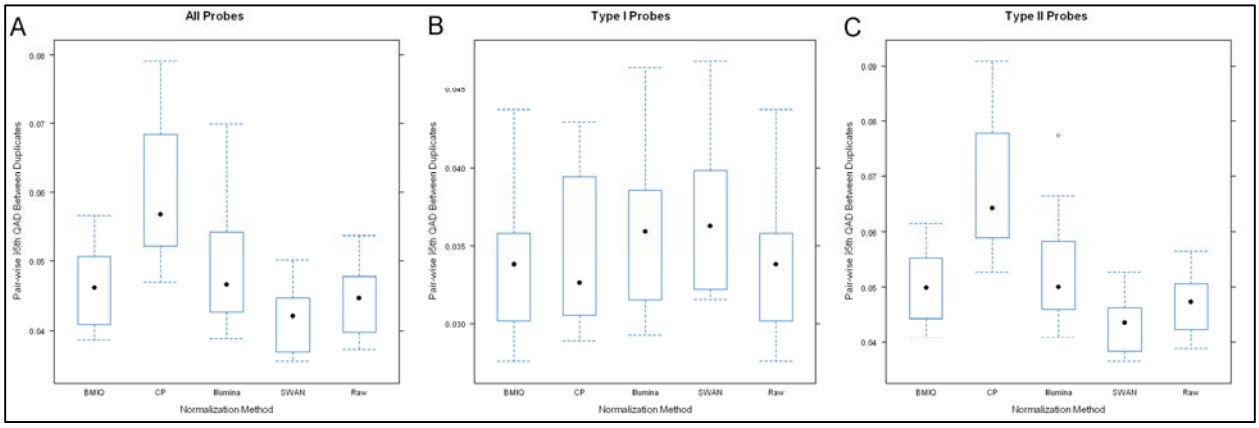
**Supplemental Figure 6.** Pairwise 99$^{th}$-QAD between duplicate pairs following application of each normalization using just Type II probes. Each line corresponds to a different duplicate pair.



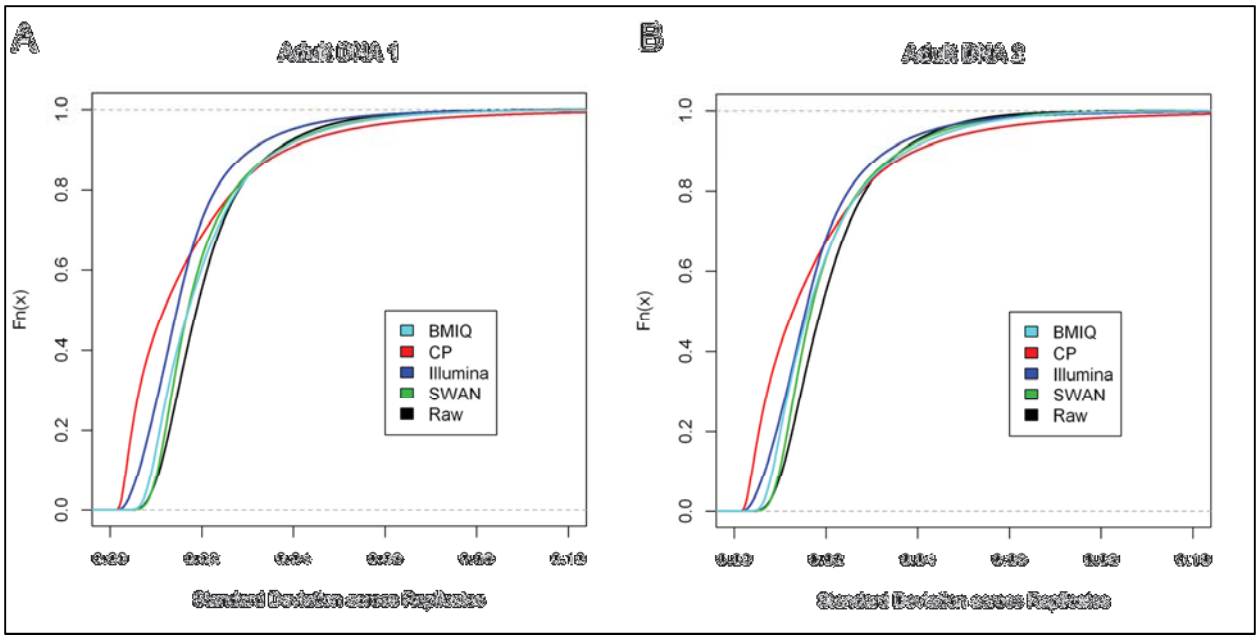
**Supplemental Figure 7.** Pairwise 100$^{th}$-QAD between duplicate pairs computed using (A) all probes, (B) just Type I probes, or (C) just Type II probes.



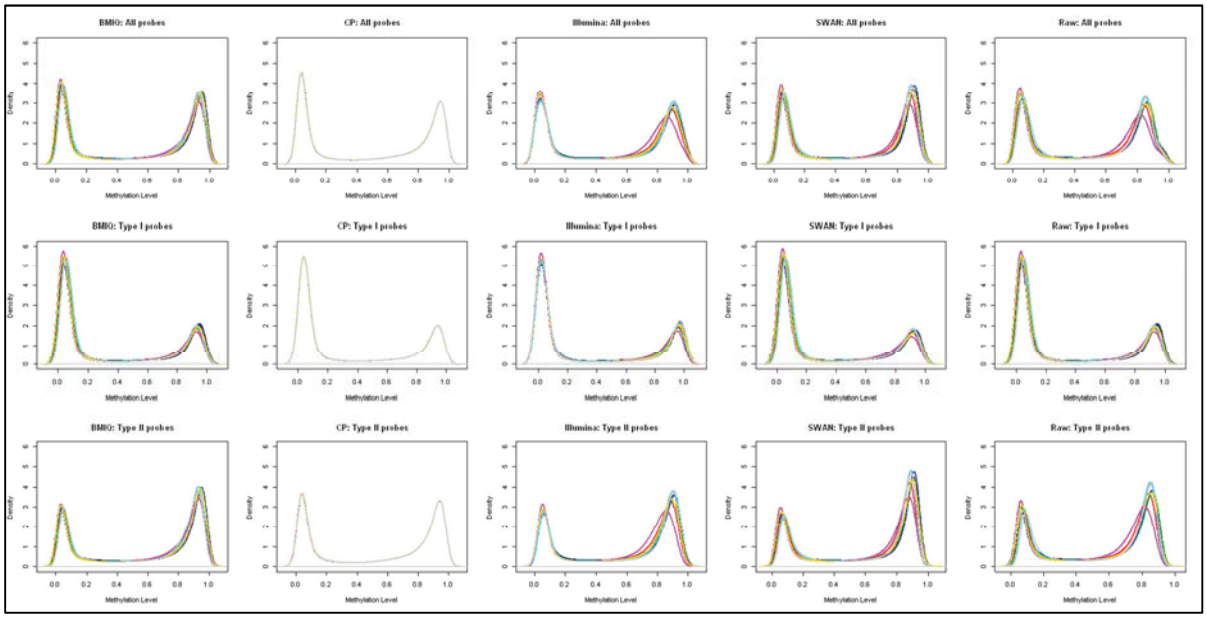
**Supplemental Figure 8.** Pairwise 95$^{th}$-QAD between duplicate pairs computed using (A) all probes, (B) just Type I probes, or (C) just Type II probes.



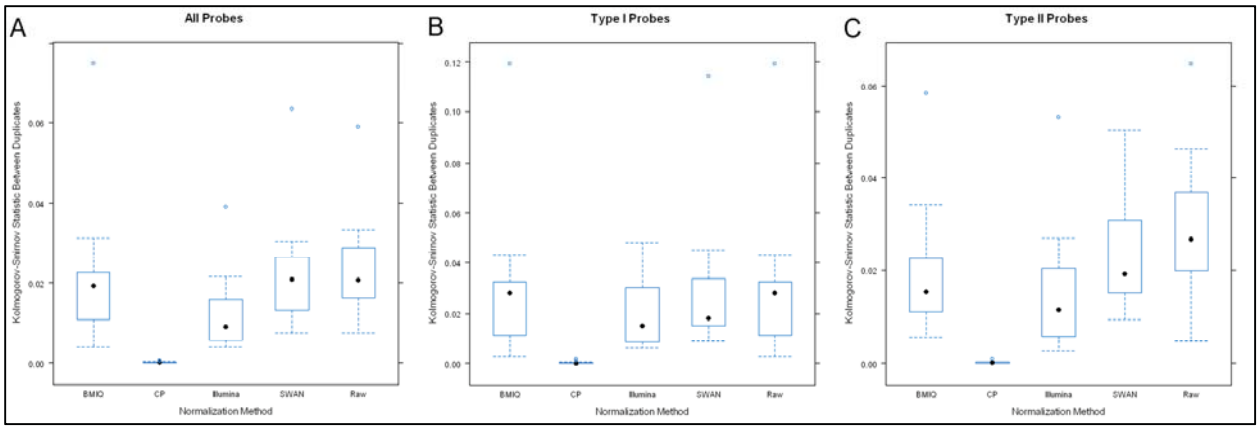
**Supplemental Figure 9.** Cumulative distribution functions (CDF) of the probe standard deviations across the replicates for (A) Adult DNA sample 1 and (B) Adult DNA Sample 2. Plots show that the upper tails of the standard deviation distribution for some methods are much heavier, indicating reduced reproducibility for a subset of the probes.



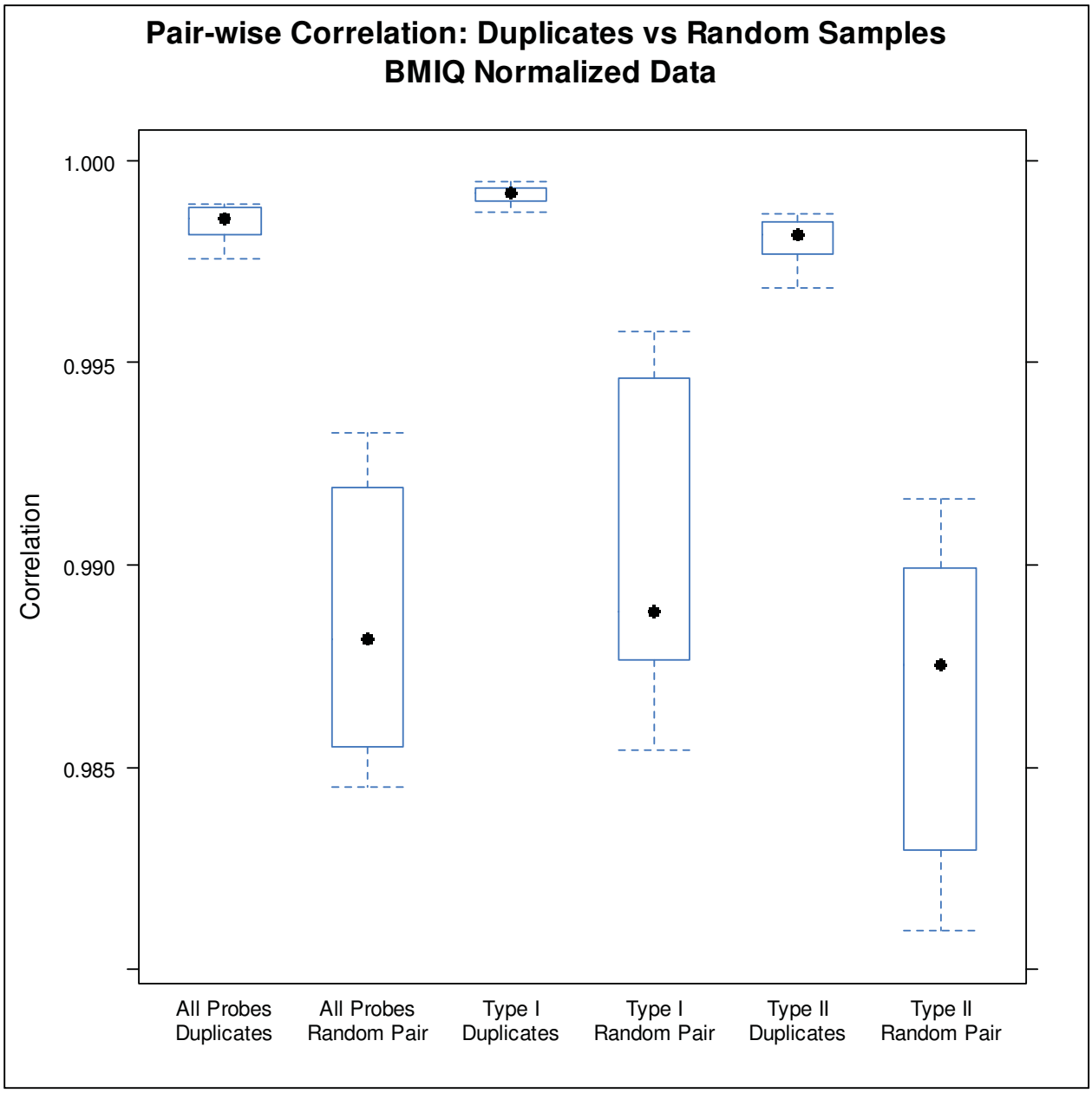
**Supplemental Figure 10.** Density plots of the methylation distributions following normalization (using all probes, just Type I probes, or just Type II probes) for eight different individual samples. Each line corresponds to a separate sample.



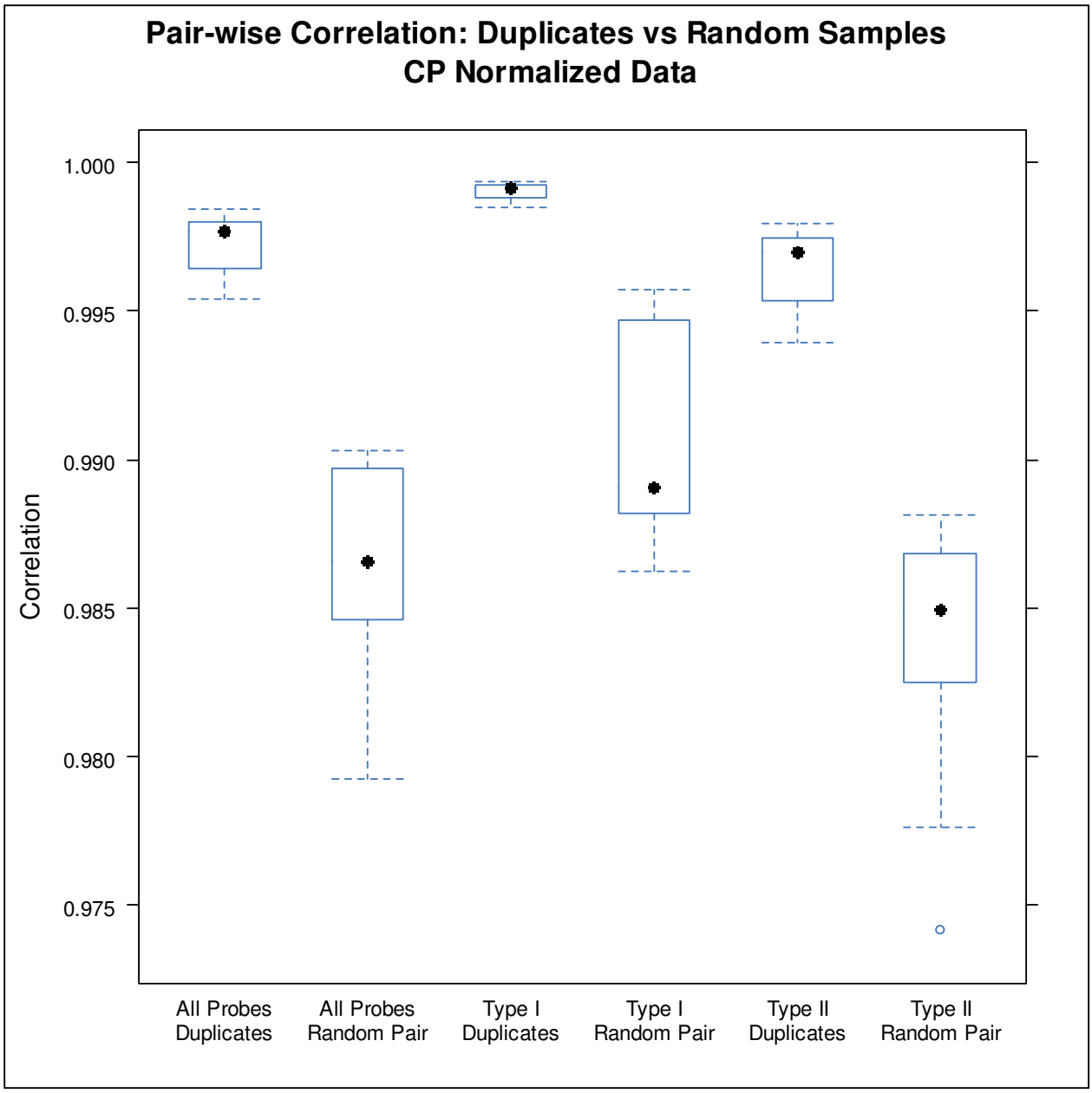
**Supplemental Figure 11.** Pair-wise Kolmogorov-Smirnov statistics of the methylation distributions computed between duplicate pairs using (A) all probes, (B) just Type I probes and (C) just Type II probes. Overall, methods behave similarly except for the CP approach which enforces similarity in overall distributions.



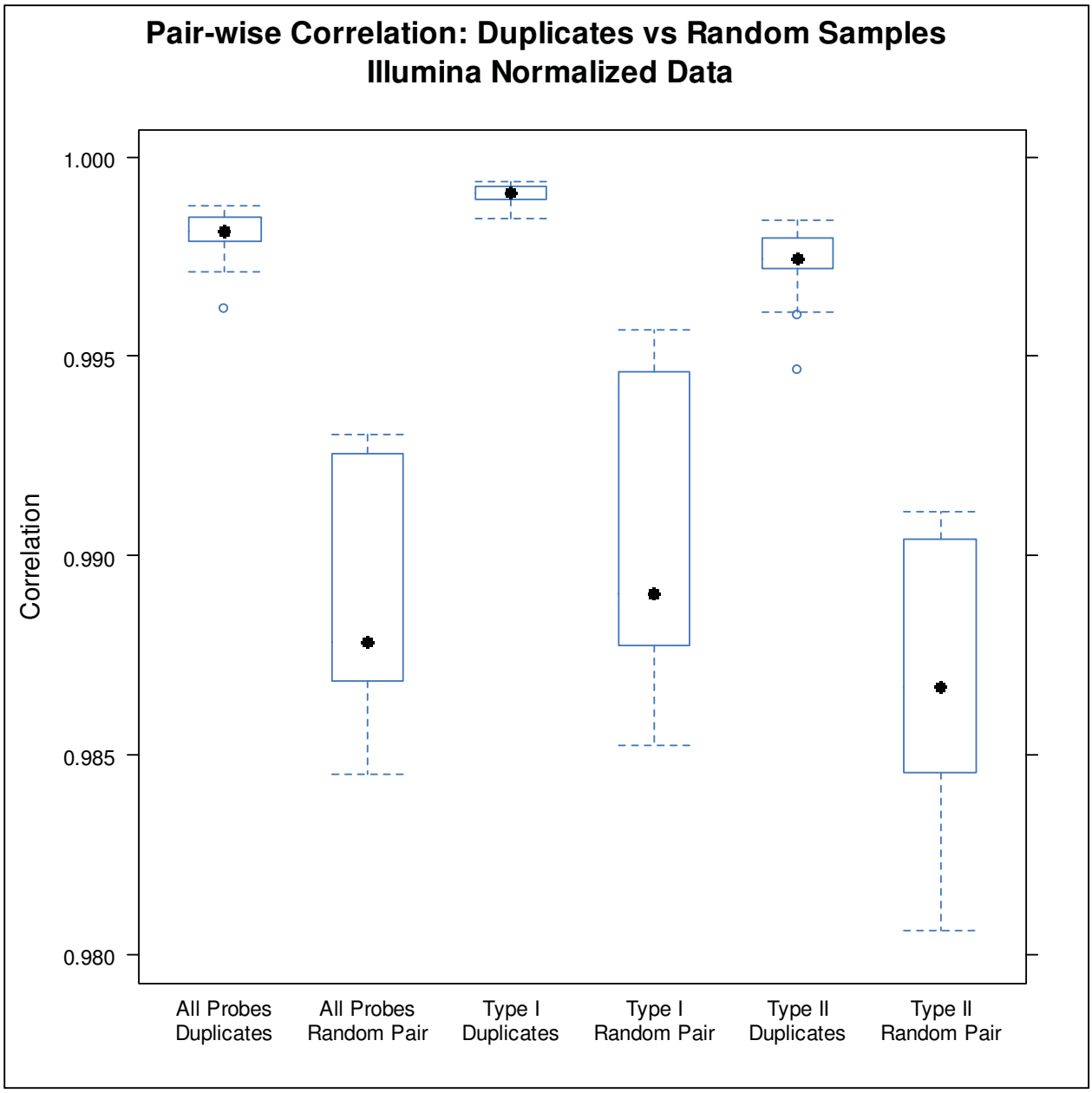
**Supplemental Figure 12.** Pairwise Pearson correlation between duplicate pairs and between non-duplicate pairs following BMIQ normalization using all probes, just Type I probes and just Type II probes.



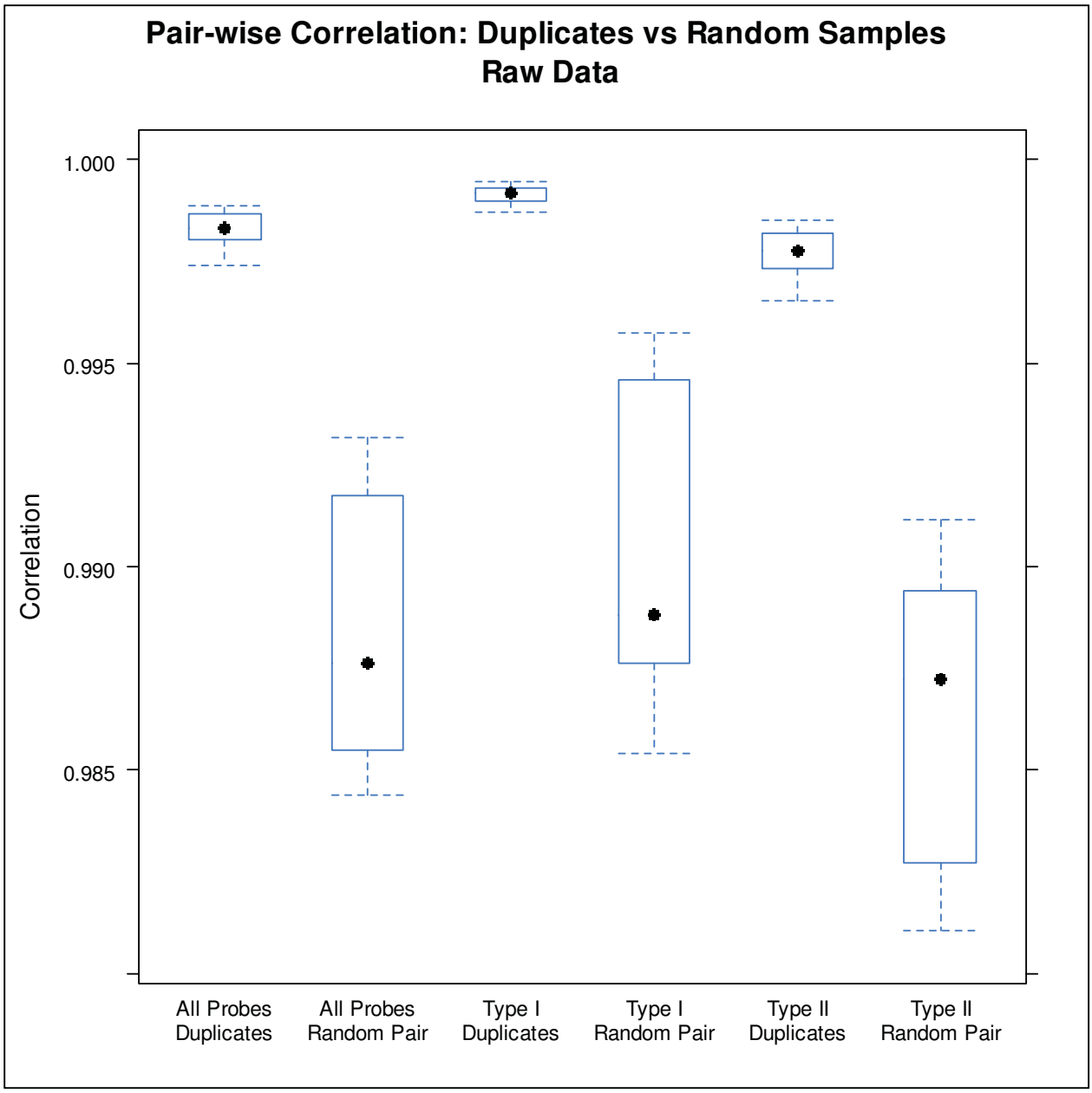
**Supplemental Figure 13.** Pairwise Pearson correlation between duplicate pairs and between non-duplicate pairs following CP normalization using all probes, just Type I probes and just Type II probes.



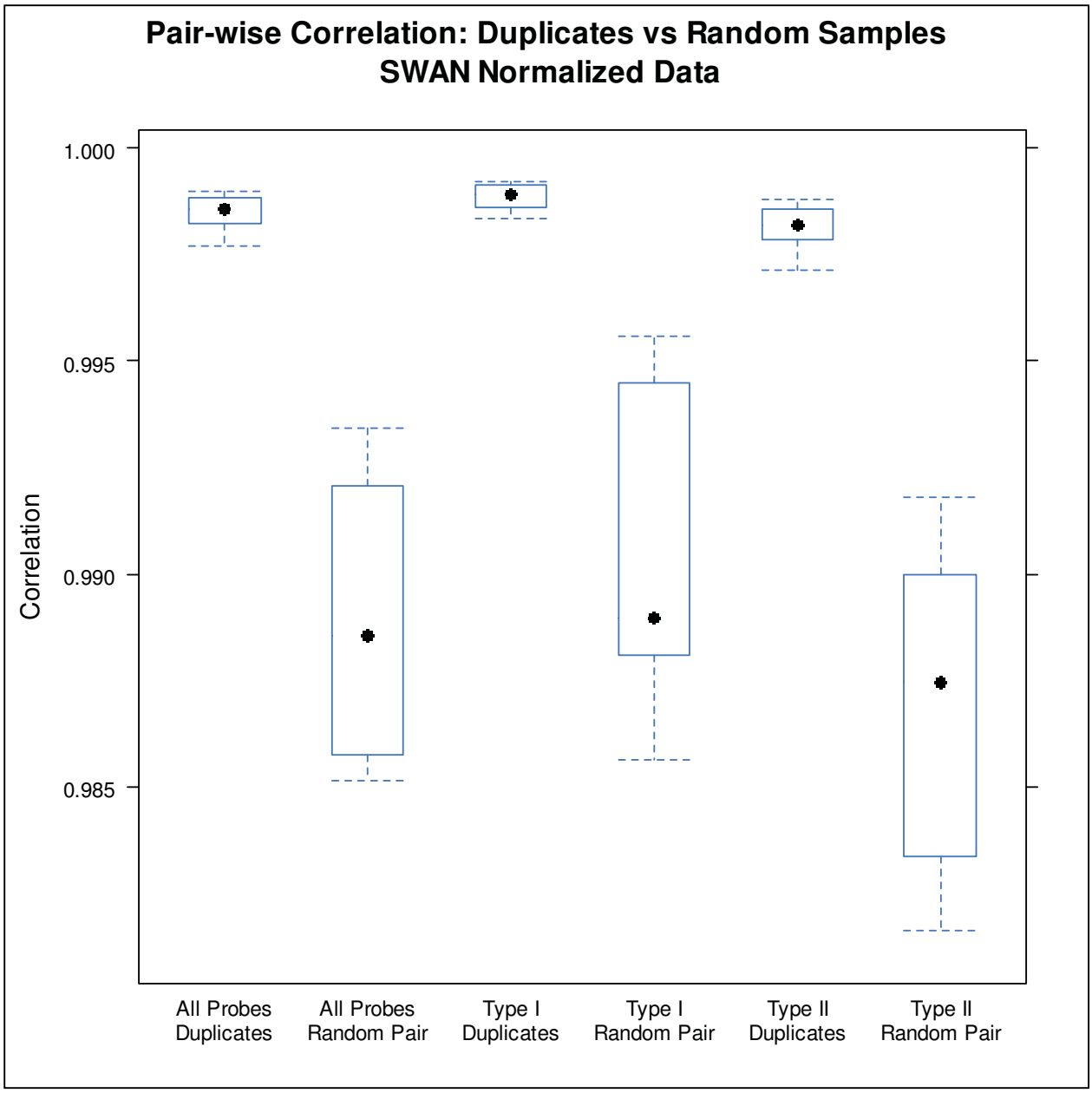
**Supplemental Figure 14.** Pairwise Pearson correlation between duplicate pairs and between non-duplicate pairs following Illumina normalization using all probes, just Type I probes and just Type II probes.



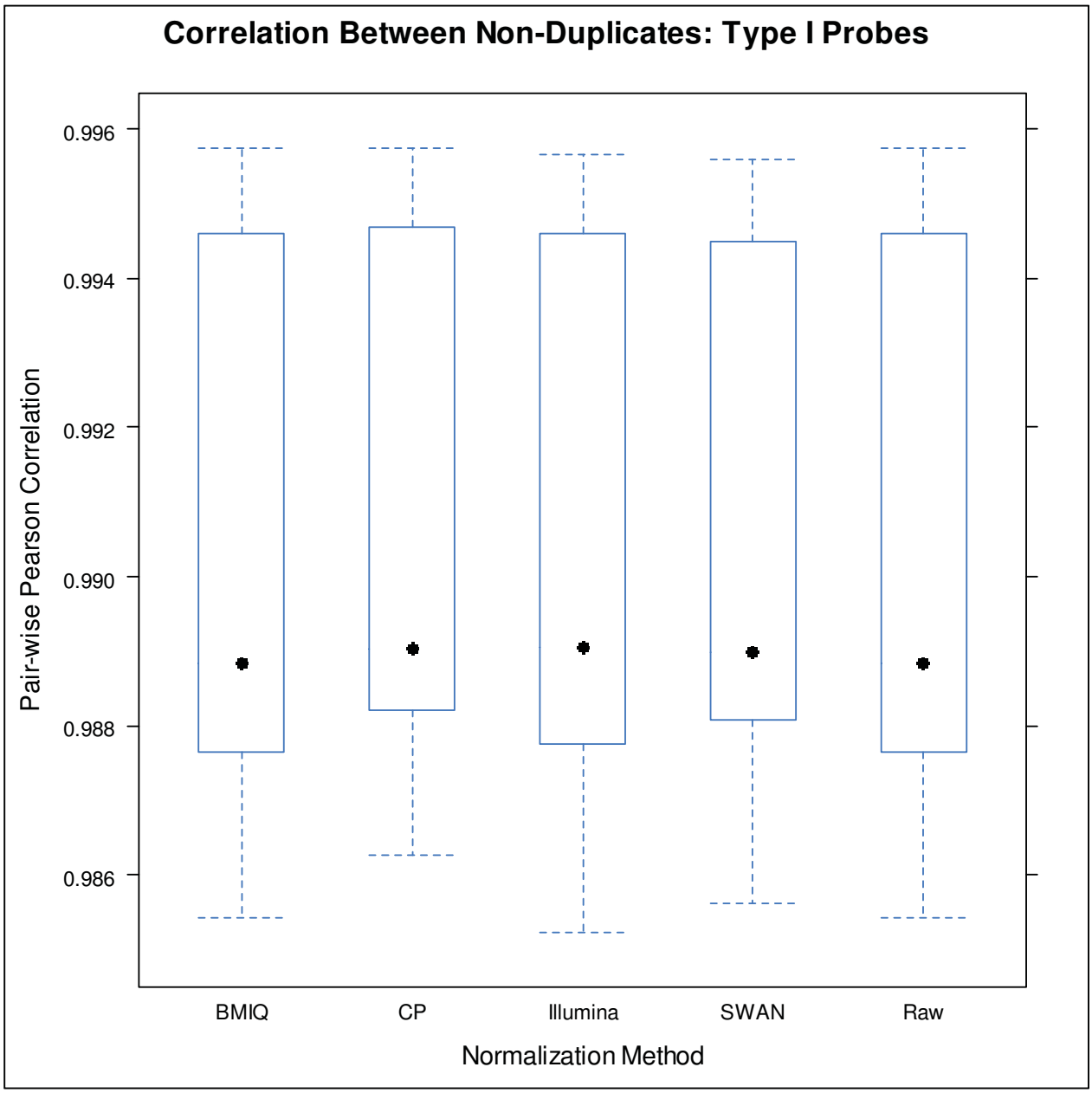
**Supplemental Figure 15.** Pairwise Pearson correlation between duplicate pairs and between non-duplicate pairs using the raw data without any normalization using all probes, just Type I probes and just Type II probes.



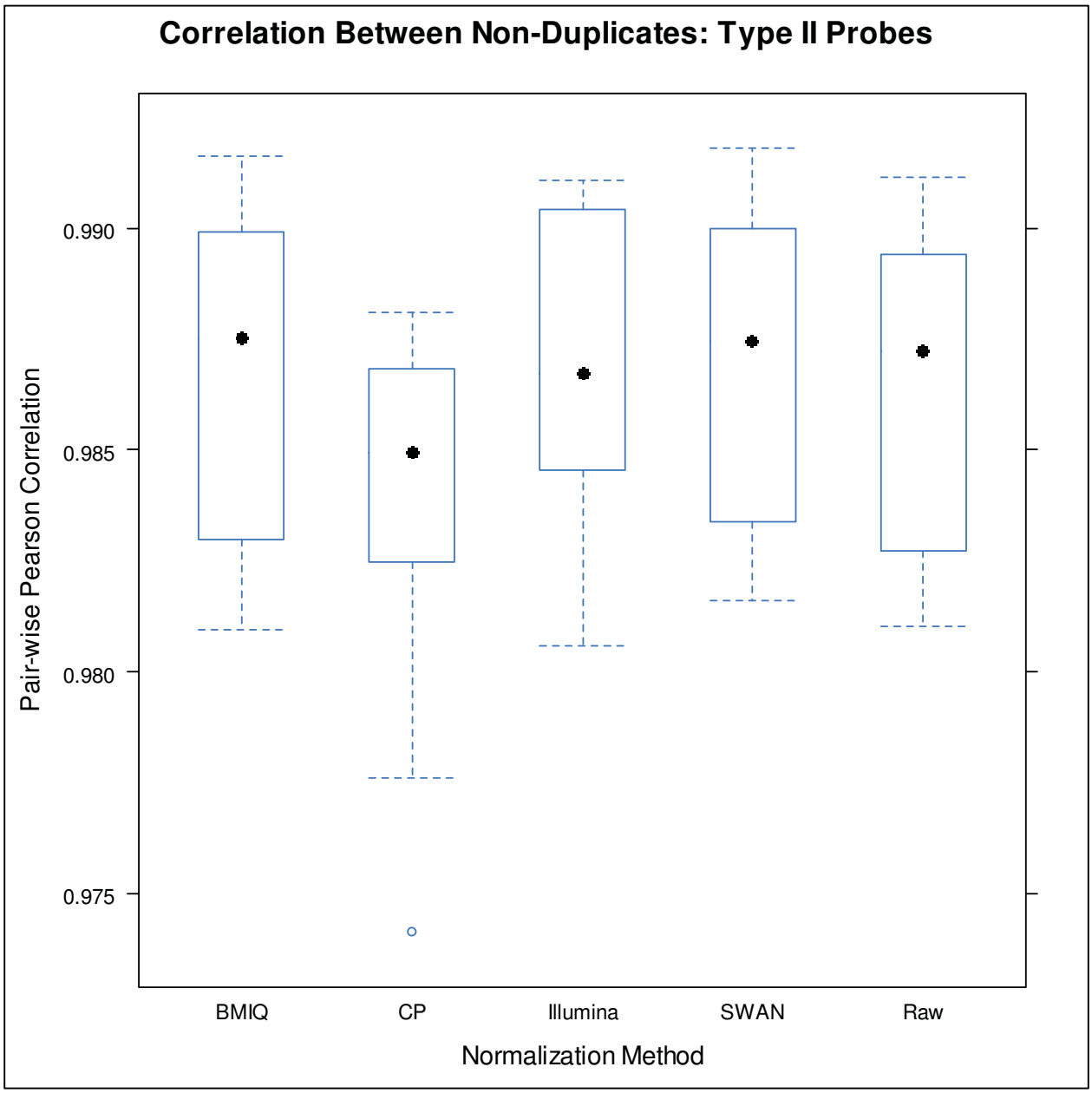
**Supplemental Figure 16.** Pairwise Pearson correlation between duplicate pairs and between non-duplicate pairs following SWAN normalization using all probes, just Type I probes and just Type II probes.



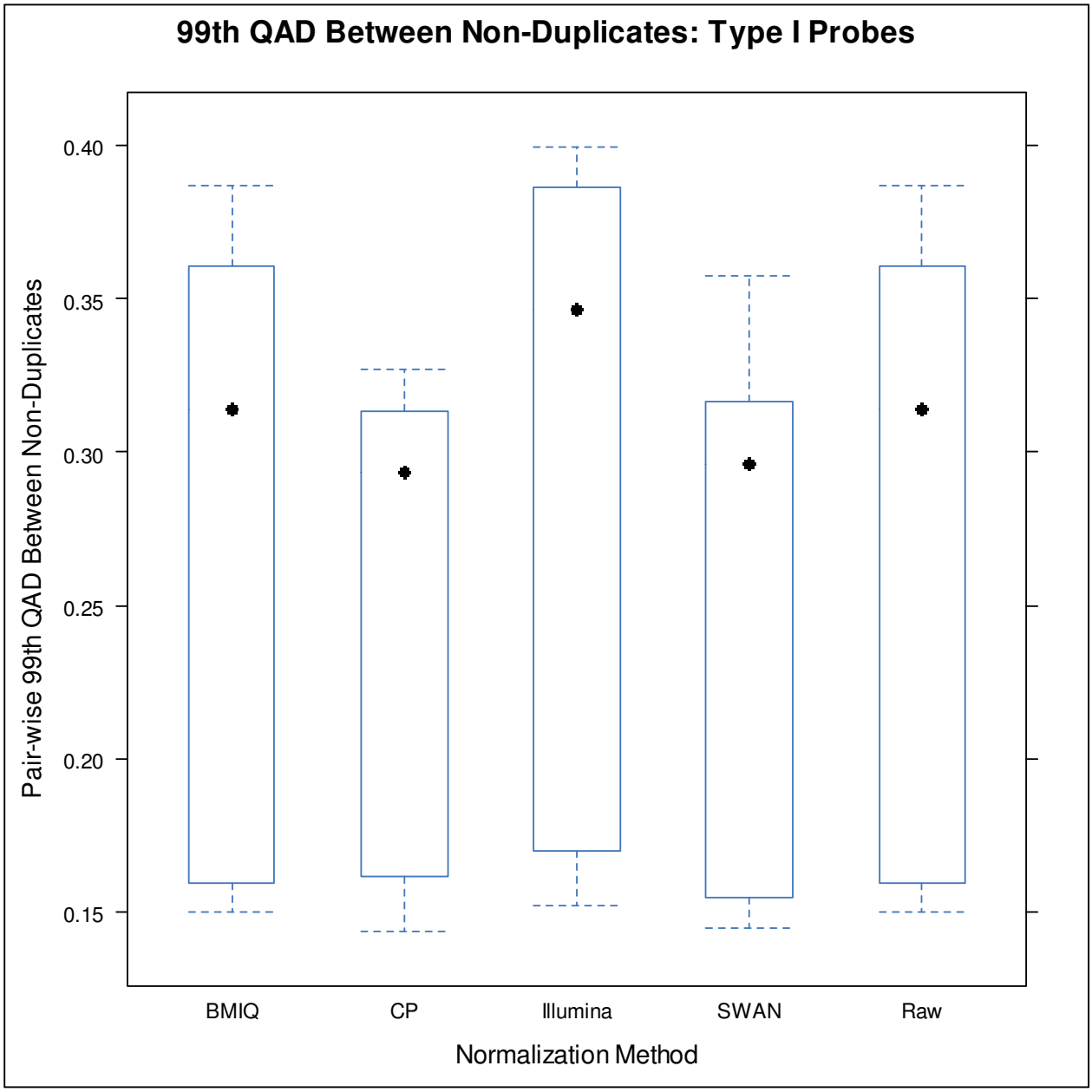
**Supplemental Figure 17.** Pairwise Pearson correlation between non-duplicates based on just Type I probes.



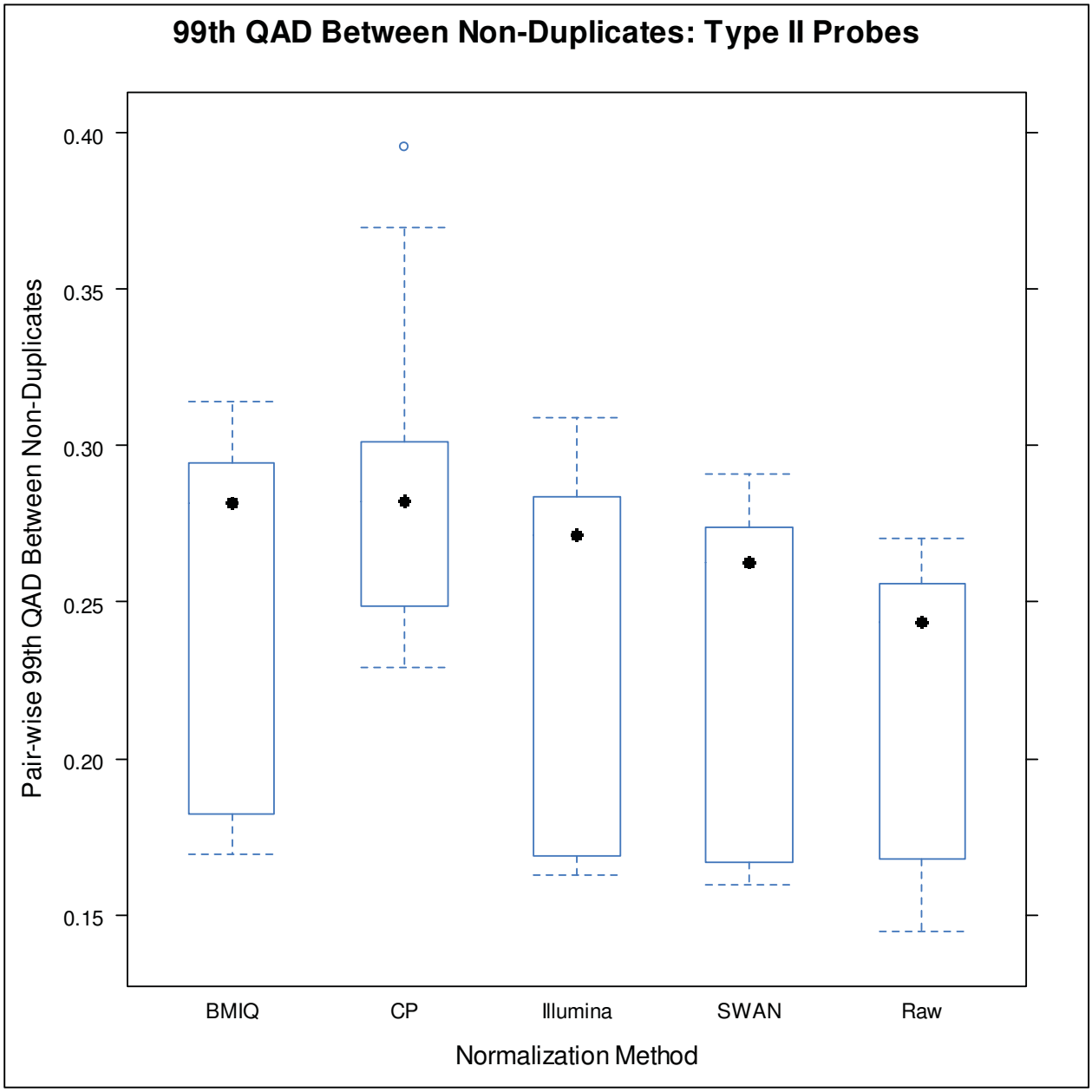
**Supplemental Figure 18.** Pairwise Pearson correlation between non-duplicates based on just Type II probes.



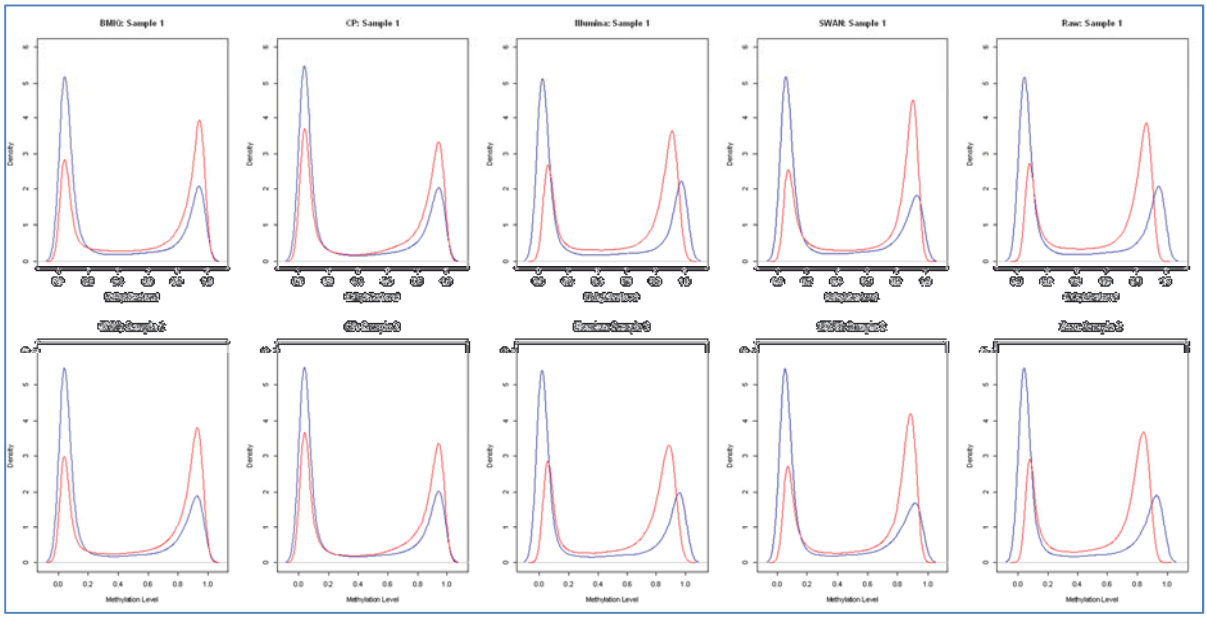
**Supplemental Figure 19.** Pairwise 99$^{th}$-QAD between non-duplicates based on just Type I probes.



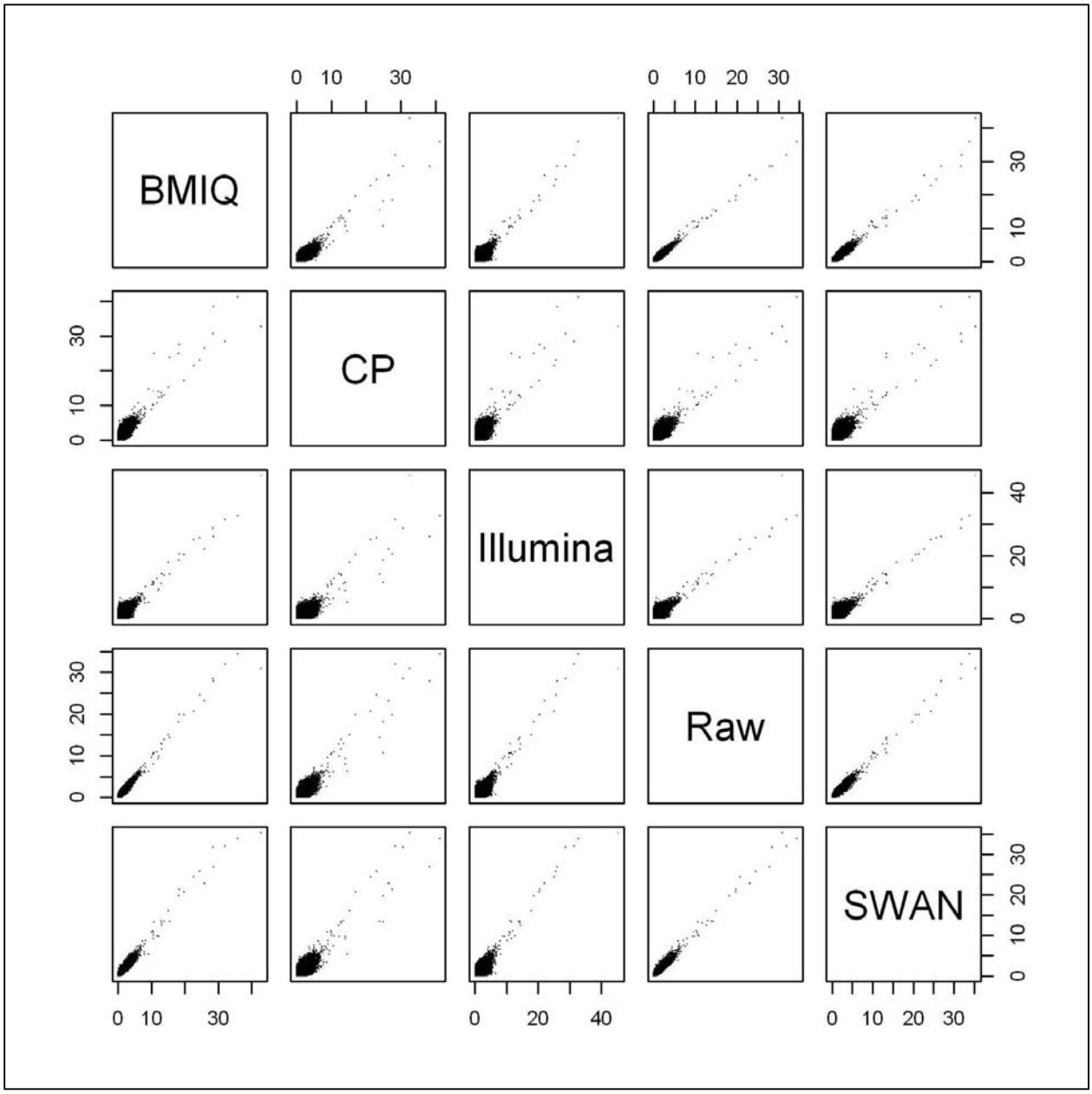
**Supplemental Figure 20.** Pairwise 99$^{th}$-QAD between non-duplicates based on just Type II probes.



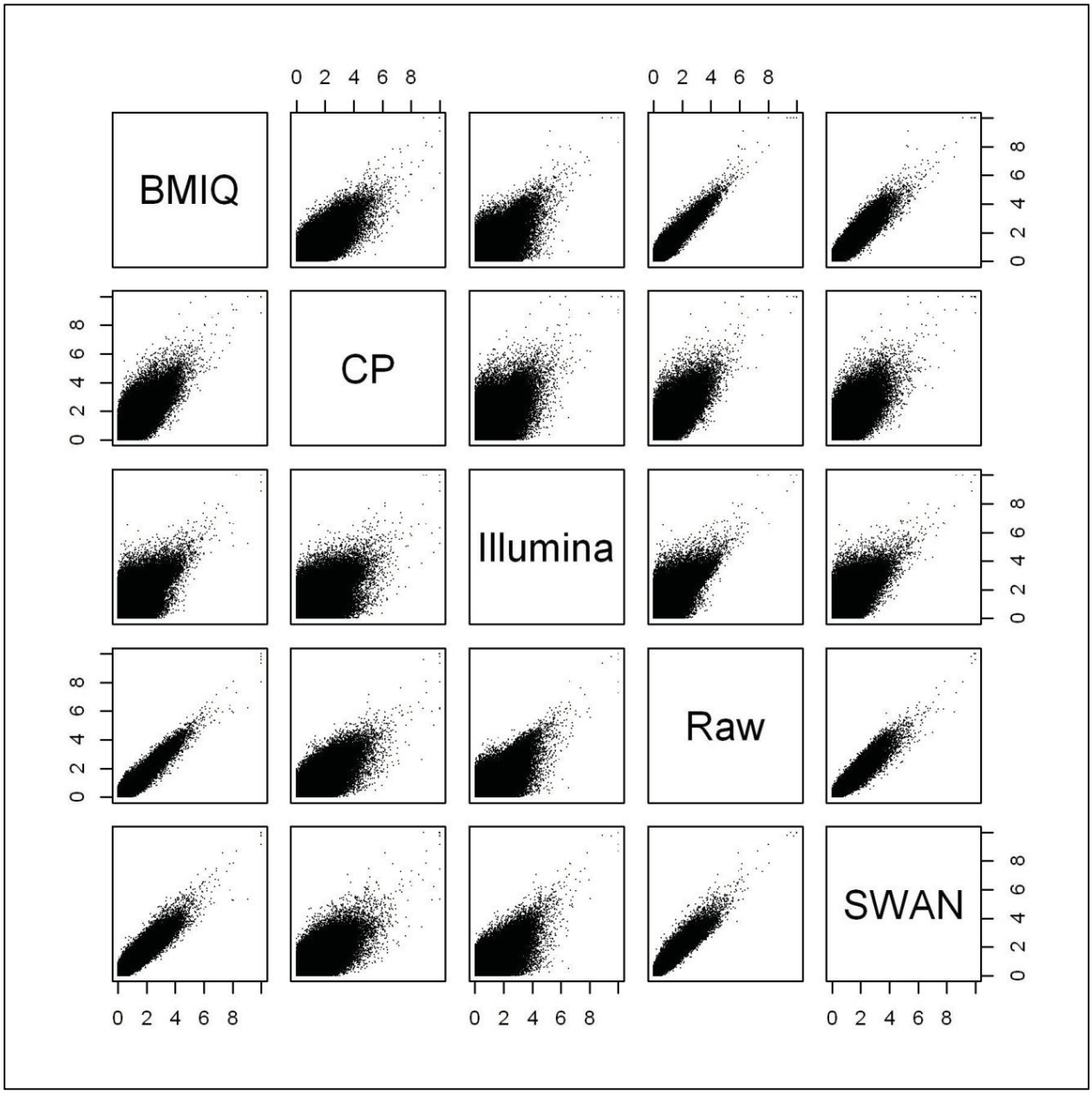
**Supplemental Figure 21.** Comparison of the density plots for overall distribution of Type I (blue lines) and Type II (red lines) probes in two different samples following application each normalization approach.



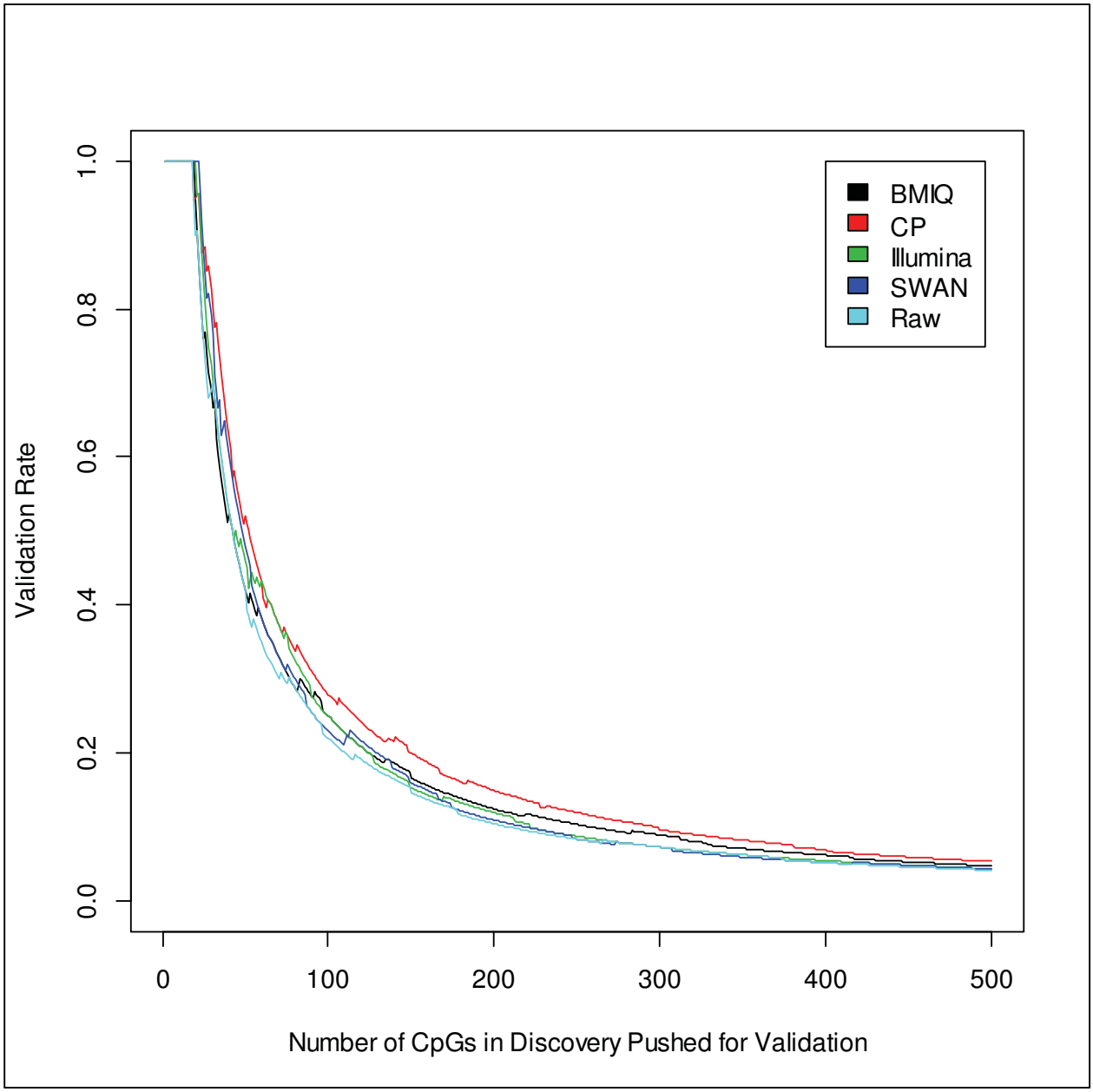
**Supplemental Figure 22.** Scatterplots of the $-\log_{10}$ p-values from the analysis of the MoBa epigenetic study.



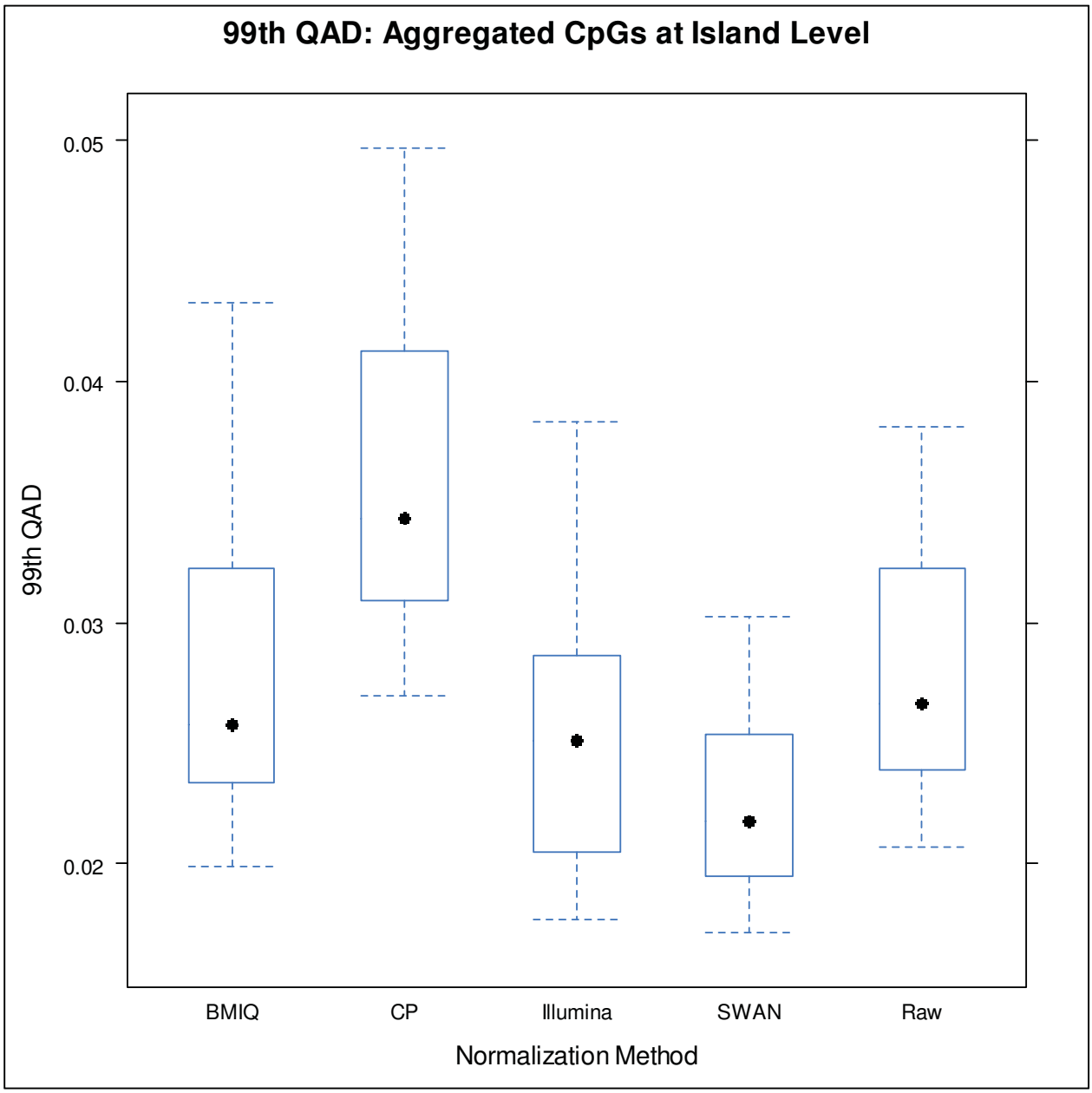
**Supplemental Figure 23.** Scatterplots of the $-\log_{10}$ p-values from the analysis of the MoBa epigenetic study with small p-values truncated at $10^{-9}$.



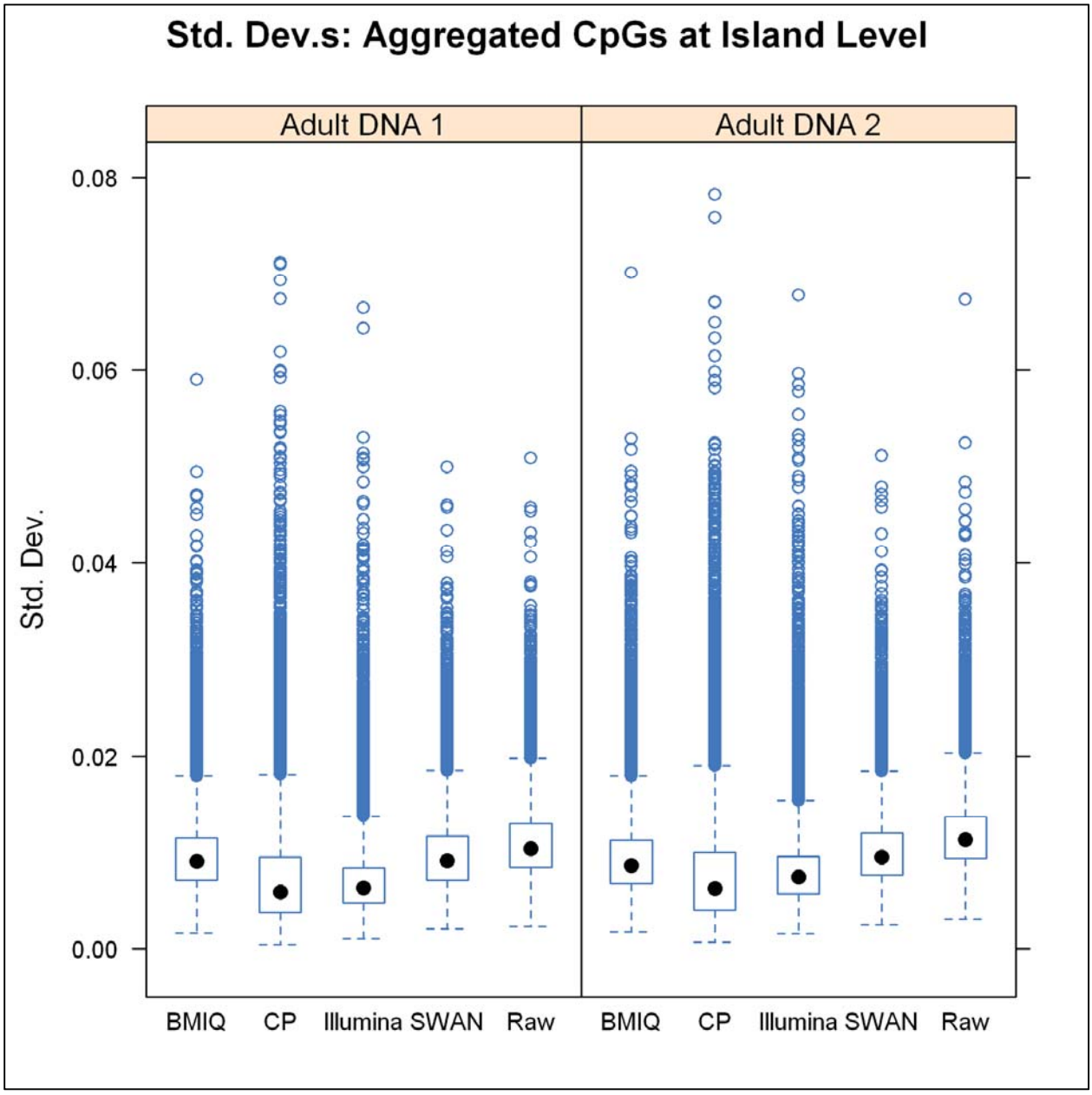
**Supplemental Figure 24.** Plot of the validation rate as a function of the number of CpGs pushed for validation from the discovery set based on the split sample analysis with 531 subjects used for discovery and 531 subjects used for validation. The x-axis is the number of CpGs that are pushed for validation and the y-axis is calculated as the proportion of the CpGs pushed for significance that would reach a Bonferroni corrected significance level of 0.05. This mimics an analysis in which ranking, rather than significance, is used to determine the CpGs to be validated. Note that these validation rates are considered upper bounds (best case) for methods that apply cross sample normalization since both discovery and validation samples were normalized together as in a true split sample analysis.



**Supplemental Figure 25.** Pairwise 99$^{th}$-QAD between duplicates with data aggregated at the CpG island level.



**Supplemental Figure 26.** Standard deviation of the data aggregated at the CpG island level across technical replicates of two different adult DNA samples.



**Supplemental Table 1.** Overlapping numbers of validated CpGs in the full split sample analysis (half of the entire sample used for each of discovery and validation). Each number corresponds to the number of CpGs called validated by both the method at the top of the table and the method at the left side of the table. For the discovery stage, significance was called at the Benjamini-Hochberg FDR = 5% level while significance at the validation stage was determined using Bonferroni correction. Note that the precise significance threshold for the validation stage depends on the number of CpGs called significant in the discovery stage.

| | BMIQ | CP | Illumina | SWAN | Raw |
|---|---|---|---|---|---|
| BMIQ | 21 | 4 | 13 | 13 | 14 |
| CP | | 30 | 3 | 5 | 4 |
| Illumina | | | 21 | 13 | 13 |
| SWAN | | | | 22 | 14 |
| Raw | | | | | 19 |

**Supplemental Table 2.** Comparison of validation rates when using data sets of different sizes. Here, *n* subjects were randomly selected to be a discovery set and a separate *n* subjects were randomly selected to be the validation set (with no overlapping subjects). Association between cotinine levels and each CpG was tested using robust regression adjusted for possible confounders within just the Discovery set with significance determined at the Benjamini-Hochberg FDR = 5% level. For each method, significant CpGs were validated in the Validation set with significance determined using Bonferroni adjustment. The validation rate is the proportion of validated discoveries for each method. Note that these validation rates are considered upper bounds (best case) for methods that apply cross sample normalization since both discovery and validation samples were normalized together as in a true split sample analysis.

| *n* | BMIQ | CP | Illumina | SWAN | Raw |
|---|---|---|---|---|---|
| 531[a] | 62% | 10% | 54% | 67% | 86% |
| 250 | 25% | 2% | 23% | 20% | NA[b] |
| 100 | 8% | 1% | 2% | 6% | NA[b] |

[a] This is the same as using the full data set on which complete covariate information was available.

[b] No CpGs were determined to be significant within the Discovery set at the FDR = 5% level.

**Supplemental Table 3.** The number of significant CpGs in the full discovery set, the full validation set, and the percentage of CpGs that are validated for each normalization procedure. Significance for Discovery was determined at the Bonferroni corrected $\alpha=10^{-7}$ level. These CpGs were validated in the Validation set. CpGs significant in the Validation set following Bonferroni adjustment for the number of significant Discovery CpGs are considered to be validated.

| | BMIQ | CP | Illumina | SWAN | Raw |
|---|---|---|---|---|---|
| Discovery | 19 | 29 | 20 | 17 | 17 |
| Validation | 19 | 24 | 20 | 17 | 17 |
| Proportion Validated | 100% | 83% | 100% | 100% | 100% |