Increased correlation between methylation sites in epigenome-wide replication studies: impact on analysis and results

Popovic Maja[1], Fasanelli Francesca[1], Fiano Valentina[1], Biggeri Annibale[2], Richiardi Lorenzo[1]

[1] Department of Medical Sciences, University of Turin and CPO Piemonte, Turin, Italy
[2] Department of Statistics, Computer Science, Applications «G. Parenti», University of Florence, Florence, Italy

Corresponding author:

Maja Popovic
Department of Medical Sciences
University of Turin
Via Santena 7,
10126 Turin, Italy
E-mail: maja_popovic@hotmail.com
Phone: +39 (0) 116334628

Abstract

Aims: To show that an increased correlation between CpGs after selection through an EWAS might translate into biased replication results.

Methods: The NINFEA cohort data were used to calculate pairwise correlation coefficients between CpGs selected in three published EWAS. We specified the appropriate replication null hypothesis, calculated r-values and Benjamini-Hochberg (BH) FDR p-values. Exposures' random permutations were performed to show the empirical p-value distributions.

Results: The average pairwise correlation coefficients between CpGs were enhanced after selection for the replication (e.g. from 0.14 at genome-wide level to 0.29 among the selected CpGs), affecting the empirical p-value distributions and the usual BH-FDR control.

Conclusions: BH-FDR method might be inappropriate for the EWAS replication phase, and methods that account for the underlying correlation need to be used.

Key words: epigenetics, replication study, correlation, bias, discovery study, EWAS

Introduction

Recent technological developments have enabled the widespread use of epigenome-wide association studies (EWAS) focused on identification of DNA methylation markers of disease state and progression and markers of a variety of exposures. Many large projects and some consortia have been established to reach a large sample size and allow comprehensive epigenetic mapping.

Although methylation occurs throughout the genome, it is often clustered along a chromosome with CpG sites likely being in the same methylation state when they are spatially close together [1]. CpG-rich areas, known as CpG islands [2], contain correlated sites with similar methylation state. The issue of correlation between nearby loci has been tackled to some extent in the EWAS by analyzing together areas with analogous functions. Region discovery [3], bump hunting [4], different clustering methods [5,6], or grouping by genomic annotations are only some of the strategies proposed in the literature that cope with correlated CpG sites. These methods offer biologically interpretable results but replication after the discovery phase is not straightforward [7].

As well recognized in the context of genome-wide association studies, replication and validation of epigenome-wide findings is essential and may be challenging. This task traditionally implies testing of few candidate CpG loci identified as top hits in the discovery sample, by applying *gold-standard* experimental methods, such as pyrosequencing, in an independent sample. Recently, high-throughput epigenome-wide studies focusing on exposures that have extensive impact on DNA methylation identify hundreds or thousands of potentially relevant single methylation sites. Replication/validation of these candidates with pyrosequencing is not possible in practice. Therefore, we often rely on replication in an independent sample with available epigenome-wide data, such as those from large epigenome consortia.

Under such scenario, it is intuitive that the average pairwise correlation between single sites in the large discovery EWAS will be lower than the average pairwise correlation between the few hundreds of single sites selected for the replication study. This fact is rarely taken into consideration in EWAS replication studies and the analyses in the replication sample may, thus, be biased. Benjamini-Hochberg False-discovery rate (FDR) correction [8], which is typically used both in the discovery and replication phase of the epigenome-wide studies is robust, yet does not take into account the underlying correlation structure. In replication studies based on epigenome-wide data, the false-discovery controlling procedure must consider an appropriate replicability null hypothesis as, for example, done by be the so-called r-value [9]. For what we have said insofar, the robustness of the procedure to the lack of independence is much more important for the replication than for the discovery study.

This article has an illustrative intent. We first show with real data examples that the average pairwise correlation between CpG sites increases after selection through an epigenome-wide discovery analysis. We then illustrate how this increased correlation may translate into biased interpretations of the results in replication analyses, and show the appropriate method for replication studies that quantifies the strength of replication taking into account the underlying correlation structure [9].

Materials

Literature dataset

We used findings from three studies assessing DNA methylation in newborns in association with three different exposures: i) a study on 6685 children from the Pregnancy and Childhood Epigenetics (PACE) consortium that identified 6073 over 464,628 CpG sites whose methylation levels were associated with maternal sustained smoking during pregnancy [10]; ii) a study on sex differences in DNA methylation in 111 Mexican-American newborns,

members of the CHAMACOS study, that identified 3031 over 410,072 CpG site candidates located on the autosomal chromosomes [11]; iii) and a study on 1988 newborns from two European cohorts that identified 443 over 419,905 CpG site candidates associated with maternal plasma folate levels during pregnancy (hereafter referred to as the "maternal plasma folate study") [12]. All these studies involved analyses on DNA methylation from newborn blood samples, and they selected CpG sites by using a fixed threshold of Benjamini and Hochberg FDR-corrected p-values of 0.05.

NINFEA replication study

We retested the selected CpG candidates from the three literature datasets described above, in epigenome-wide data coming from the NINFEA birth cohort [13]. The study design was a nested case-control study on 72 cases with at least one reported episode of wheezing between 6 and 18 months of age and 72 controls matched to cases by sex, age at sampling and seasonality/calendar year of sampling. In the NINFEA birth cohort saliva samples are routinely collected from infants at approximately 6 months of age using a mailed Oragene self-collection kit, and in the nested case control study we focused on saliva DNA methylation markers of childhood wheezing (data not published). DNA extracted from the saliva samples of cases and matched controls was assessed for epigenome-wide methylation using the Illumina Infinuim HumanMethylation450 BeadChip. Three cases and three matched controls were excluded due to high percentage of missing DNA methylation values, leading to a total of 138 subjects available for the analyses. The baseline NINFEA questionnaire is completed by mothers during pregnancy and includes questions on sustained smoking in pregnancy and intake of folic acid during pregnancy – no information is instead available on plasma folate levels. Information on child sex is obtained at the first follow-up questionnaire that is completed 6 months after delivery.

The Ethical Committee of the San Giovanni Battista Hospital and CTO/CRF/Maria Adelaide Hospital of Turin approved the NINFEA study (approval N. 0048362, and subsequent amendments), and all the participating mothers gave their informed consent before taking part in the study.

Methods

Statistical analyses

NINFEA cases and controls were pooled together. For each selected CpG site – see details below – we obtained a methylation percentage (Beta values) from the Illumina Infinuim HumanMethylation450 BeadChip and they were converted to M values by a logit transformation [14]. After quality control and probes filtering (probes corresponding the SNPs inside the probe body and SNPs at CpG sites, cross hybridizing and probes on the sex chromosomes) a total of 321,084 probes were available in the NINFEA dataset.

For each of the three literature datasets (the PACE consortium, the CHAMACOS study and the maternal plasma folate study) we retrieved the published selected altered CpG sites that were then used in the NINFEA dataset. Due to different probe filtering between the NINFEA study and the three literature datasets, there was an incomplete overlap of the top hits.

All the analyses were performed using RStudio (version 0.99.903).

The analytical flow is summarized in **Figure 1** and described below in details.

Correlation analysis

For each of the three groups of selected CpG sites – derived from the three literature examples – we estimated the pairwise Spearman correlation coefficients between the CpG site M values in the NINFEA replication study. The three distributions of correlation coefficients were compared with the distribution of genome-wide pairwise correlation coefficients between CpG sites (histograms, summary statistics with the 3rd, 50th and 97th percentiles, F test on

homogeneity of variance on Fisher's zeta transformation [15]). To obtain the genome-wide correlation distribution, we calculated the pairwise correlation coefficients between 100,000 randomly selected CpG pairs among all available CpG sites in the NINFEA data.

Replication analyses

Replication of CpG sites associated with maternal smoking and those associated with child's sex was then conducted in the NINFEA data. As data on maternal folate levels during pregnancy were not available in the NINFEA study, we did not identify children exposed and unexposed to folate to carry out the replication analyses.

For two groups of selected CpG sites we evaluated the associations of exposure to maternal smoking during pregnancy and child's sex with offspring saliva DNA methylation levels. A robust linear regression model was specified, and adjusted as in the discovery studies: child's sex was analyzed in univariate models, while the models for maternal smoking during pregnancy were adjusted for maternal age, maternal education (low, medium and high) and parity. Batch effect was controlled by within batch matching between exposed and unexposed subjects. We did not adjust for cell type composition as there is no reference data set for the saliva cell composition and we are not aware of studies assessing performance of the reference free method [16] in saliva samples. We used p-values of the test of the association between maternal smoking or child's sex and DNA methylation in infants. Histograms and quantile-quantile (QQ) plots were used to graphically evaluate the observed versus the expected null distribution of p-values. Deviations from the theoretical uniform distribution were also formally tested using the Kolmogorov-Smirnov test [17].

Assessment of the empirical p-value distributions

To evaluate the impact of the increased correlation among the selected CpG sites, we assessed the p-value distributions under the null-hypothesis of no effects of the exposures on the

methylation levels in the selected CpG sites. For this purpose, we generated 10,000 random shuffling of the exposed-unexposed status for each individual in the two datasets (maternal smoking during pregnancy and child's sex) while maintaining the same ratio between exposed and unexposed subjects within each batch as in the original data. The associations between the randomly attributed exposure and methylation in the 4794 CpG sites for maternal smoking or 2544 CpG sites for sex were estimated in each replicate using the same linear regression models. P-value distributions of the 10,000 replicates were described in terms of symmetry by estimating the skewness and in terms of deviation from a uniform distribution by performing Kolmogorov-Smirnov [17,18] and Anderson-Darling tests [18–20]. To compare empirical distributions, we generated additional 10,000 replicates for both examples (maternal smoking and child's sex) with random assignment of the exposure variables and random CpG sites selection.

To ensure that the low exposure frequency in the analyses on maternal smoking did not affect the underlying distribution under the null hypothesis, we analyzed all NINFEA subjects with available EWAS data by shuffling the imaginary exposure with 69 "cases" and 69 "controls" and relating it to methylation levels in 4794 smoking-related CpG sites.

Finally, to decrease the underlying correlation from both sets of CpG sites (maternal smoking and child's sex) we selected only sites that have all pairwise correlation coefficients below 0.40 in the NINFEA dataset. On these two subsets of low-correlated CpG sites associated with maternal smoking and child's sex we conducted the same analyses with 10,000 randomly assigned exposures and for comparison randomly assigned CpG sites.

Multiple testing correction and r-values

Multiple comparisons correction of the NINFEA results using Benjamini-Hochberg FDR procedure is not appropriate for a replication study and, therefore, we used r-values [9]. R-

value is defined as the lowest FDR at which the finding can be called replicated, and with its modified version accounts also for the underlying dependence between the p-values within the primary study [9]. For each CpG site of the two datasets (maternal smoking and child's sex) we computed both standard r-values for independent tests and its modified version that accounts for the underlying correlation. A CpG site is considered replicated if the r-value < 0.05. For demonstration purposes we also present p-values corrected using Benjamini-Hochberg FDR procedure [8].

Results

CpG sites selection

Of the 6073 top hits from the PACE study, 4794 (78.9%) CpG sites overlapped with those from the NINFEA data set. NINFEA children exposed to maternal sustained smoking during pregnancy (N=6, 4.3%) were matched to the unexposed children (N=30) by batch in which samples were analyzed, keeping a constant 1:5 ratio between exposed and unexposed children.

A total of 2544 CpG sites (83.9%) were available for the replication of the results on sex differences and DNA methylation in newborns. The analyses were performed in 80 children, by choosing the maximum number of exposed children available within each batch that could be matched with unexposed children from the same batch to keep a constant 1:3 ratio between "exposed" (females, N=20) and "unexposed" (males, N=60) subjects.

Out of the candidate CpG sites identified in the maternal plasma folate study, 344 (77.7%) were available in the NINFEA dataset.

Correlation analysis

Table 1 reports the summary statistics for the Spearman correlation coefficients calculated in the NINFEA data between the top CpG sites from each of the three literature datasets and for unselected genome-wide CpG pairs. The corresponding distributions are reported in **Figure 2**.

When being pre-selected in the discovery studies, such as in the examples presented here, the average correlation between CpG sites tends to increase depending on the exposure under study. For example, the mean correlation of 0.29 between several thousands of CpG sites associated with maternal smoking during pregnancy was much higher than the original genome-wide mean correlation of 0.14. The variance of correlations in the pre-selected CpG sites also increased substantially compared with the genome-wide CpG sites (all p-values for F test $<2.2x10^{-16}$, visual inspection of **Figure 2**).

Replication analyses

Figure 3 reports the p-value distributions and the QQ plots for the replication analyses of the top CpG sites for maternal smoking and child's sex in the NINFEA data. For both exposures, there was a clear deviation of the p-value distributions and QQ plots from what would be expected by chance (Kolmogorov-Smirnov p-value $<2.2 \times 10^{-16}$ in both analyses). The analysis on child's sex revealed 383 CpG sites (15.1%) with a p-value <0.05, while maternal smoking during pregnancy was associated with 7.8% of the analyzed CpG sites at conventional 5% level of significance. Based on these results, one could suggest that both the results for maternal smoking and the results for child's sex are globally replicated using the NINFEA saliva samples.

Assessment of the empirical p-value distributions

In the absence of correlation, by randomly permuting and re-analyzing the data we would expect the p-value distribution to be approximately uniform in most of the replications. Distributions as those observed in **Figure 3** - skewed versus lower p-values - are expected to

be seen in a small proportion of the replications. After visual inspection of the p-value distribution histograms from the 10,000 random permutations of the exposure variable we noticed that the percentage of replications not following the uniform p-value distribution was much higher than the expected 5%, both in the case of pre-selected CpG sites and in the case of genome-wide randomly selected CpG sites.

In fact, Kolmogorov-Smirnov p-values were low even when the p-value distribution histograms visually showed quite uniform patterns (**see Supplemental Material; see Figure S1**). Accordingly, as reported in **Table 2**, more than 90% of the replications were associated with a Kolmogorov-Smirnov p-value < 0.05. This proportion was higher in the case of pre-selected than randomly selected CpG sites. The Anderson-Darling test, considered more sensitive to the tails of a distribution than the Kolmogorov-Smirnov test [20], gave similar results (data not shown). However, it should be considered that, with large sample sizes, these test are likely to give strong evidence against the null hypothesis (i.e. they are able to detect even small departures from the theoretical distribution) [21].

To further explore the impact of the correlation structure on the p-value distributions we plotted the skewness of the underlying p-value distributions from the 10,000 replications for each of the examples (**Figure 4**). Symmetric distributions, such as the uniform or normal distribution, have the skewness value zero, while left- or right-skewed distribution have positive or negative values, respectively. From **Figure 4**, it can be noted that in the presence of a higher correlation between CpG sites, such as in the examples presented here, the skewness of the p-value distributions has a larger variation and is shifted towards positive values (left-skewed distributions) compared to the distributions of genome-wide randomly selected CpG sites. A similar pattern was also observed when all 138 subjects were analyzed with CpG sites associated with maternal smoking during pregnancy (**see Supplemental**

Material; see Figure S2), thus ruling out a possible impact of the small sample size in the example with maternal smoking during pregnancy.

It is noteworthy that the biases that we have so far described are mainly due to the underlying correlation structure. For demonstration purposes we have selected 256 out of 4794 CpG sites related to maternal smoking during pregnancy and 129 out of 2544 CpG sites related to child's sex that have all pairwise correlation coefficients below an arbitrary level of 0.40 in the NINFEA dataset. Mean correlation coefficient was 0.09 for both low-correlated data sets, and thus lower than the underlying genome-wide mean correlation of 0.14.

P-value distributions of the 10,000 random permutations of the exposure variables were "uniform" in 15.8% permutations of maternal smoking and 5.3% permutations of child's sex according to Kolmogorov-Smirnov test. The average skewness was 0.04 for maternal smoking and 0.004 for child's sex, with standard deviations much smaller than that for genome-wide randomly selected CpG sites (**Figure 5**). The results were similar when analyses on 256 CpG sites associated with maternal smoking were performed in all 138 subjects from the NINFEA data (**see Supplemental Material; see Figure S3**).

Multiple testing correction and r-values for replicability

After the initial replication performed in Step 2 (Figure 1, Figure 3) a standard naïve and incorrect practice would then be to consider the results of the single CpG sites, after implementing some of the procedures that take into account multiple testing and reduce the number of false positives, such as Benjamini-Hochberg FDR multiple testing correction. After the Benjamini-Hochberg correction at the 0.05 FDR level only five CpG sites were selected for sex differences in methylation levels, while no CpG site remained associated with maternal smoking during pregnancy, reflecting the small number of exposed subjects (N=6) in the NINFEA dataset.

A correct approach for a replication study would be to apply FDR-based replication p-values (r-values). For the analyses on sex differences in methylation levels, unmodified version of r-value revealed 4 replicated CpG sites (all r-values=0.02), while after considering the underlying correlation only three sex-associated CpG sites remained replicated in the NINFEA cohort (all r-values=0.03). It should be noted that all CpG sites considered replicated according to the r-value estimates had a Benjamini-Hochberg FDR p-value <0.05, while two CpG sites that passed the FDR naïve correction were not replicated (**Table 3**). No CpG site was replicated for maternal smoking during pregnancy.

Discussion

The large number of tests performed in epigenome-wide association studies requires statistical and computational methods to control for multiple testing both in the exploratory and in the replication phase. The most commonly used methods dealing with this issue, such as Bonferroni and Benjamini-Hochberg FDR corrections, rely on the assumption of independence of the tests. This assumption is often violated in EWAS, as spatially related CpG sites are very often in similar methylation state.

As shown in this paper, a certain degree of correlation already affects the discovery phase of EWAS, when analyses are carried out at the genome-wide level. This underlying correlation structure is enhanced in large sample size studies of exposures/outcomes that broadly affect DNA methylation, in which thousands of candidate CpG sites are selected for replication. The increase in correlation can be substantial: in one of the examples that we evaluated in this paper the mean pair-wise correlation coefficient increased from 0.14 at the genome-wide level to 0.29 among the selected CpG sites. Thus, the independency assumption of standard multiple testing procedures can be seriously violated, resulting in spurious replication findings. It should be noted that we analyzed the correlation structure and its impact on the results using only one dataset with saliva DNA methylation measured in children at

approximately 6 months of age. Average correlation at genome-wide level and that of pre-selected CpG sites might be different in other data sets, populations, age groups or tissues/biofluids.

We argue that in situations of high correlation it is important to explore its magnitude by conducting permutations in which the exposure/outcome status is randomly shuffled. The so-called permutation procedure that empirically generates a model-free p-value is based on this approach, and it is robust to the data correlation – a Family-wise Error Rate (FWER) control procedure (i.e. a procedure to control for type I errors in the context of multiple testing) based on permutations was proposed in the literature [22]. The only assumption behind permutation procedures is that the observations are exchangeable under the null hypothesis [22], while the most important limitation is the long computational time, especially in large EWAS. Several alternatives that account for the underlying correlation structure have been proposed and are shown to be as efficient as the permutation procedure [23–27]. The implementation of these approaches requires much less time, but to our knowledge, they are seldom used in the analysis of EWAS. Although not in the context of an increased correlation in replication studies, a recent study by van Iterson *at al.* [28] sheds light on the inflation and bias of test statistics in EWAS and transcriptome-wide association studies. They proposed a Bayesian method for the estimation of the empirical null distribution and bias and inflation correction in the presence of the correlated test statistics [28].

Several recently published EWAS used Benjamini-Hochberg FDR method to adjust for multiple tests, both in the discovery and replication analysis. Apart from using alternative methods to account for the underlying correlation, an option for the replication phase would be to select a subgroup of CpG sites using ad-hoc algorithms to decrease the correlation, including, for example, approaches based on the genomic location or the introduction of a maximum threshold for pairwise correlation coefficients. To our knowledge, the performance

and validity of possible selection criteria remains to be systematically investigated in methodological studies.

In this study we applied and recommended r-values as an FDR-based measure and an appropriate method for replication studies. The modified version of r-value guarantees false-discovery rate control under any type of dependency between tests, and is an appropriate approach for the replication phase of EWAS.

Finally, we have also shown that the Kolmogorov-Smirnov and Anderson-Darling tests, often used to assess departures from a uniform distribution of p-values, become extremely sensitive in presence of large sample sizes. Thus, if hundreds or thousands correlated CpG sites are selected for replication, these tests will almost invariably generate low p-values, and a spurious result of a global replication of the exploratory phase is very likely.

Conclusions

We caution against using FWER control procedures (e.g. the simple Bonferroni correction) or Benjamini-Hochberg FDR control in epigenome-wide replication studies, where the correlation between CpG sites can be substantial and the null hypothesis different than the null hypothesis of a discovery study. Permutation procedures are proposed as the method of choice to control FWER in the circumstances of highly correlated tests, but they are time-consuming when applied to large-scale studies, and are seldom used in EWAS. In replication studies, CpG sites for replication could also be selected *a priori*, based on different criteria or their combinations, such as significance in the discovery sample, correlation with other CpG sites, genomic location or biological significance. An option that is highly recommended for replication studies is the computation of r-values, which focus specifically on the strength of replication. In the context of epigenome-wide replication studies with highly correlated tests, we suggest to rely on robust FDR r-value.

Executive summary

- The most commonly used approaches dealing with multiple testing in the replication phase of epigenome-wide association studies are type I error rate and false-discovery rate controls that, although claimed to be robust, assume independence between tests.
- The correlation between CpGs is enhanced after selection during the discovery phase.
- In the replication phase of EWAS an increased correlation between CpGs might translate into biased empirical p-value distributions, affecting also the usual control by Benjamini-Hochberg FDR procedure.
- Benjamini-Hochberg FDR method might not be adequate for the replication phase of EWAS.
- Replication studies should consider methods that take into account the underlying correlation structure, including permutation procedures and r-values to detect replicated associations.

Acknowledgements

We are grateful to all the participants of the NINFEA birth cohort. We thank Silvia Polidoro for her contributions to the EWAS study nested in the NINFEA cohort and Giovanni Fiorito for his comments on an earlier version of the manuscript.

Financial and competing interests disclosure

The NINFEA study was partially funded by the Compagnia San Paolo Foundation. MP was supported by the Erasmus Mundus for Western Balkans (ERAWEB II) programme (reference number: E2.D2.14.270) and the present work is a part of her PhD thesis. AB was partially

supported by MIUR ex 60% funds. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval. Informed consent has been obtained from the participants involved.

References

1. Eckhardt F, Lewin J, Cortese R *et al*. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38(12), 1378–1385 (1987).
2. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol* 196(2), 261–282 (1987).
3. Ong ML, Holbrook JD. Novel region discovery method for Infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways. *Aging Cell* 13(1), 142–155 (2014).
4. Jaffe AE, Murakami P, Lee H *et al*. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* 41(1), 200–209 (2012).
5. Sofer T, Schifano ED, Hoppin JA, Hou L, Baccarelli AA. A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics* 29(22), 2884–2891 (2013).
6. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559 (2008).

7. Lin X, Barton S, Holbrook JD. How to make DNA methylome wide association studies more powerful. *Epigenomics* 8(8), 1117–1129 (2016).

8. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57(1), 289–300 (1995).

9. Heller R, Bogomolov M, Benjamini Y. Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proc Natl Acad Sci U S A* 111(46), 16262–16267 (2014).

10. Joubert BR, Felix JF, Yousefi P *et al*. DNA Methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. *Am J Hum Genet* 98(4), 680–696 (2016).

11. Yousefi P, Huen K, Davé V, Barcellos L, Eskenazi B, Holland N. Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. *BMC Genomics* 16, 911 (2015).

12. Joubert BR, den Dekker HT, Felix JF *et al*. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nat Commun* 7, 10577 (2016).

13. Richiardi L, Baussano I, Vizzini L *et al*. Feasibility of recruiting a birth cohort through the Internet: the experience of the NINFEA cohort. *Eur J Epidemiol* 22, 831–837 (2007).

14. Du P, Zhang X, Huang C-C *et al*. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11, 587 (2010).

15. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika* 10(4), 507–521 (1915).

16. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* 30(10), 1431–1439 (2014).

17. Massey FJJ. The Kolmogorov–Smirnov test for goodness of fit. *J Am Stat Assoc* 46, 68–78 (1951).

18. Razali NM, Wah YB. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *J. Stat. Modeling Anal* 2, 21–33 (2011).

19. Anderson TW, Darling DA. A test of goodness of fit. *J Am Stat Assoc* 49(268), 765–769 (1954).

20. Stephens MA. EDF statistics for goodness of fit and some comparisons. *J Am Stat Assoc* 69, 730–737 (1974).

21. Lin MF, Lucas HC, Shmueli G. Too big to fail: large samples and the p-value problem. *Inform Syst Res* 24, 906-917 (2013).

22. Good P. Permutation tests: A practical guide to resampling methods for testing hypotheses, 2nd edition. Springer-Verlag, New York, NY (1994).

23. Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet* 81(6), 1158–1168 (2007).

24. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Stat Plan Infer* 82, 171–196 (1999).

25. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29(4), 1165–1188 (2001).

26. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 74(4), 765–769 (2004).

27. Dudbridge F, Koeleman BPC. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet* 75(3), 424–435 (2004).

28. van Iterson M, van Zwet EW, BIOS Consortium, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol* 18(1), 19 (2017).

Table 1. Summary statistics of the correlation coefficients' distributions, expressed as absolute values, in 138 children of the NINFEA cohort.

Set of CpG sites	**N**	**3rd percentile**	**Mean**	**Median**	**97th percentile**
Genome-wide	321084	0.01	0.14	0.10	0.49
Maternal plasma folate study	344	0.01	0.15	0.11	0.46
Child's sex	2544	0.01	0.19	0.14	0.59
Maternal smoking during pregnancy	4794	0.01	0.29	0.24	0.76

Table 2. Kolmogorov-Smirnov test assessing the uniformity of the p-value distributions from 10,000 permutations

Permutations (N=10,000)	**Percentage of permutations associated with a Kolmogorov-Smirnov[a] p-value < 0.05 (%)**
Maternal smoking during pregnancy	96.4
Random CpG sites	93.0
Child's sex	95.9
Random CpG sites	93.8

[a] Kolmogorov-Smirnov test to determine if the distribution of p-values from each replication is equal to the expected uniform distribution.

Table 3. Sex-associated CpG sites that passed Benjamini-Hochberg (BH) FDR correction and corresponding uncorrected p-values, BH FDR p-values and FDR r-values

CpG sites	Discovery study p-value	Replication study p-value	BH FDR p-value	FDR r-value[a]	Modified r-value[b]
cg25438440	1.86e-18	1.81e-05	0.02	0.02	0.03
cg19544707	4.06e-12	9.77e-06	0.02	0.02	0.03
cg03168896	9.30e-09	1.53e-05	0.02	0.02	0.03
cg12763978	5.65e-07	1.40e-05	0.02	0.02	0.23
cg14022202	5.85e-06	2.89e-05	0.02	0.14	0.40

[a] R-value for the independent tests
[b] Conservative modification of r-value that accounts for any type of the dependency between tests

Figure 1. The main steps of the analysis

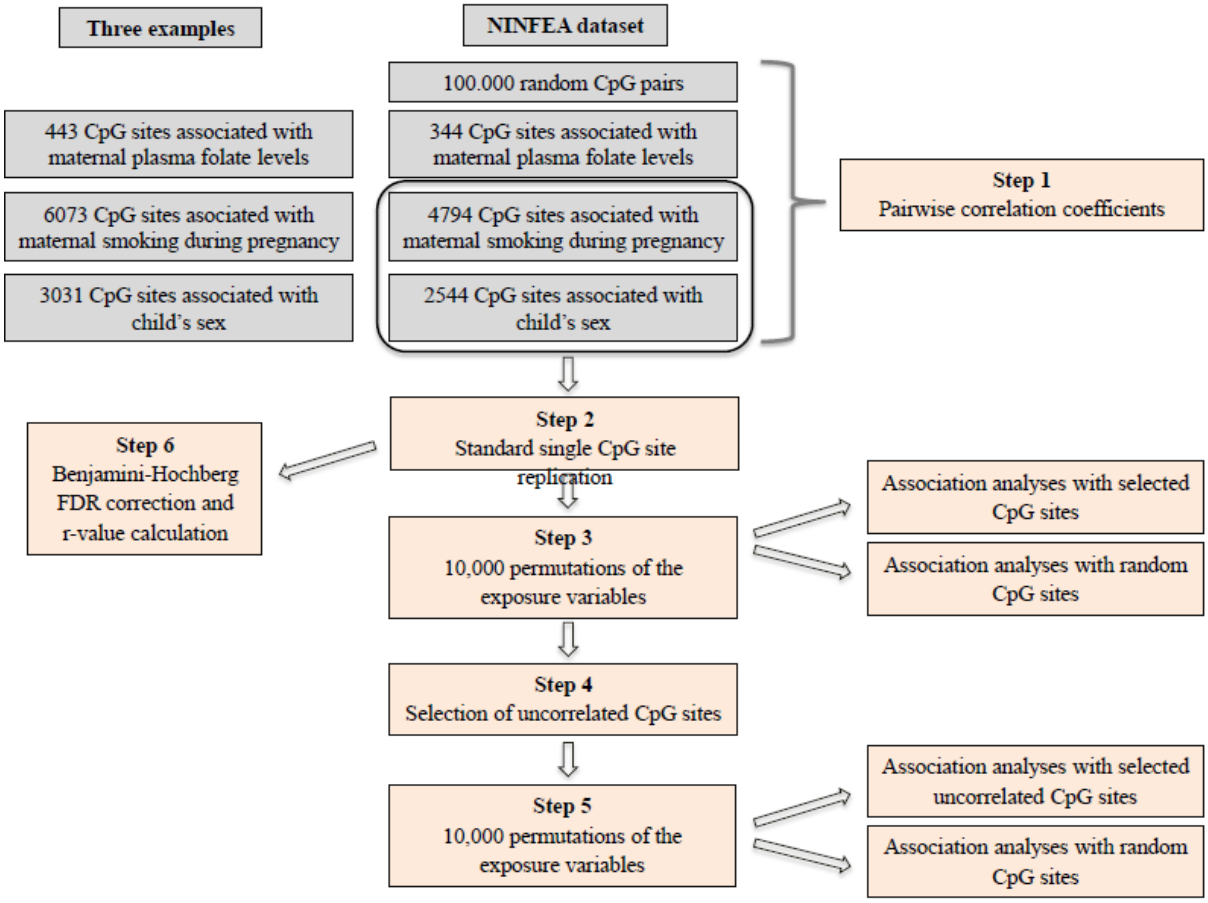


Figure 2. Distribution of correlation coefficients (left side) and their absolute values with the corresponding absolute mean correlation coefficients (right side) for genome-wide CpG sites,

4794 CpG sites associated with maternal smoking during pregnancy, 344 CpG sites associated with maternal folate levels and 2544 CpG sites associated with child's sex.

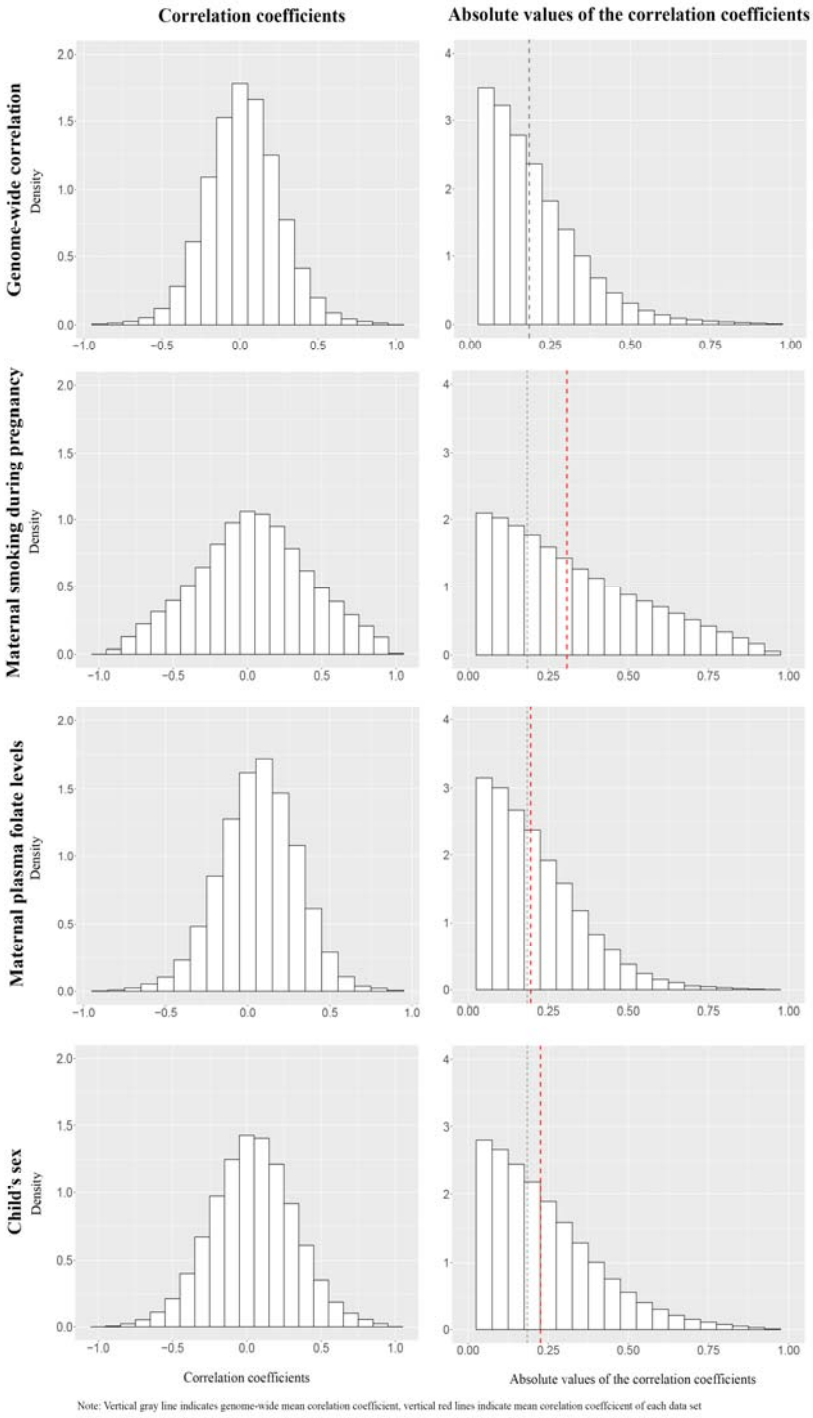


Figure 3. Distribution of replication p-values and QQ plot of observed versus expected p-values testing the null hypotheses of no association between methylation levels of pre-selected CpG sites and maternal smoking during pregnancy, and child's sex.

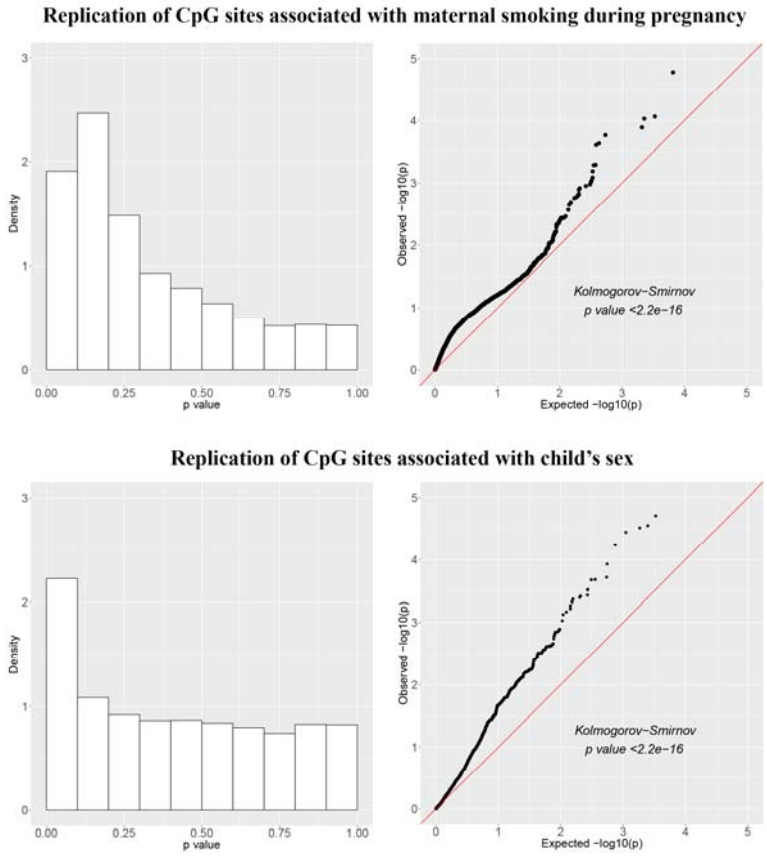


Figure 4. Skewness of p-value distributions from the analyses of the association between smoking-related and sex-related CpG sites and permutations of maternal smoking during pregnancy and child's sex from 10,000 replications. "*Random*" indicates random permutations of both CpG sites and exposure under study.

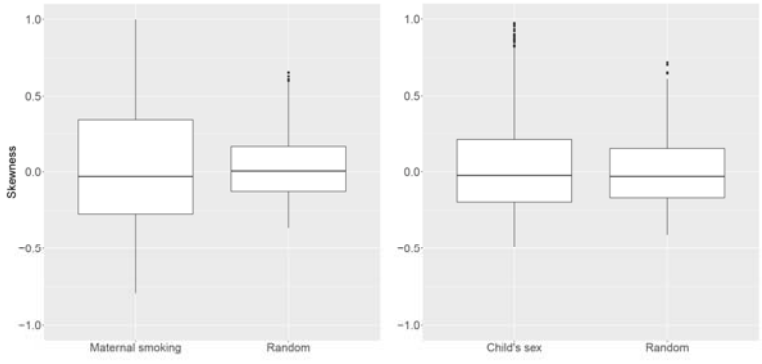
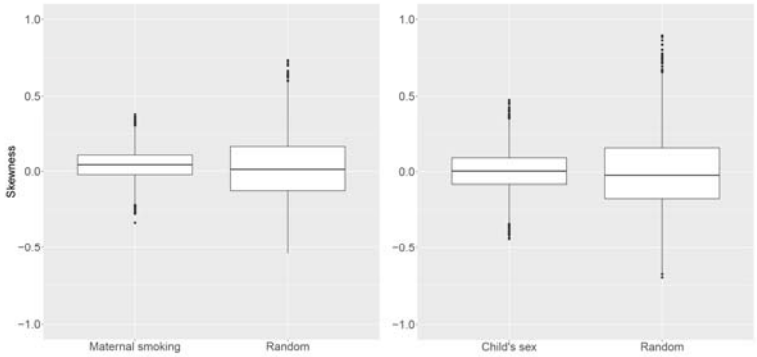


Figure 5. Skewness of p-value distributions from the analyses of the association between pre-selected low-correlated CpG sites and permutations of maternal smoking/child's sex from

10,000 replications. "*Random*" indicates random permutations of both CpG sites and exposure under study.



Supplemental Material

Increased correlation between methylation sites in epigenome-wide replication studies and its impact on analysis and results

Popovic Maja, Fasanelli Francesca, Fiano Valentina, Biggeri Annibale, Richiardi Lorenzo

Table of Contents

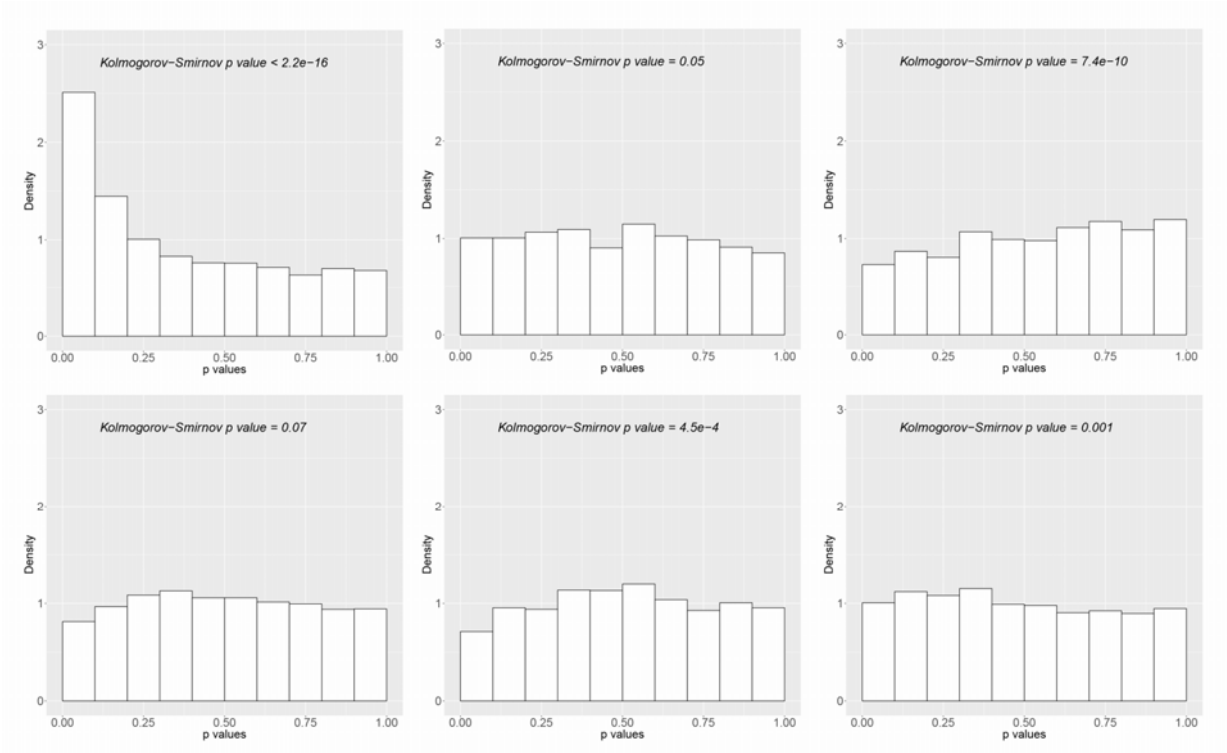


Figure S1. Histograms of p-value distributions from random permutations and Kolmogorov-Smirnov p value assessing whether the observed p-value distributions come from a hypothesized uniform distribution

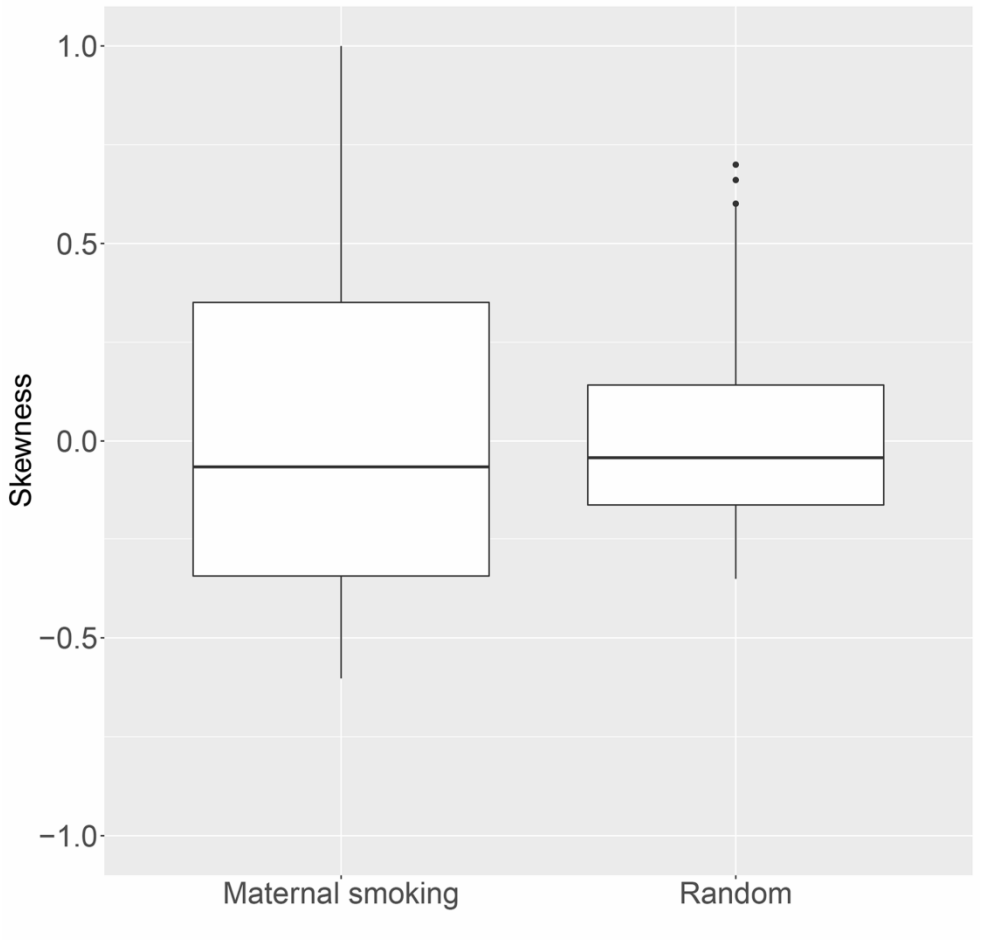


Figure S2. Skewness of p-value distributions from the analyses of the association between 4794 CpG sites associated with maternal smoking and 10,000 permutations of an imaginary exposure for 138 subjects from the NINFEA cohort. "*Random*" indicates random permutations of both CpG sites and exposure under study.

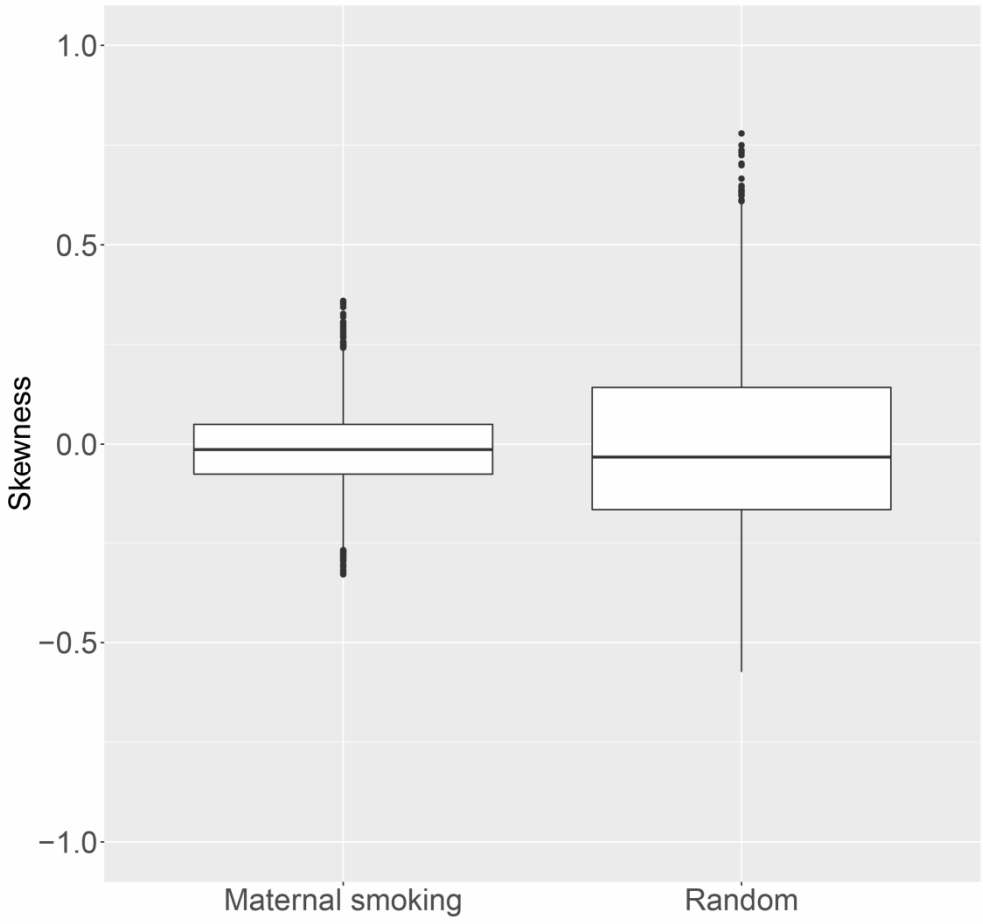


Figure S3. Skewness of p-value distributions from the analyses of the association between 256 low-correlated CpG sites associated with maternal smoking and 10,000 random permutations of an imaginary exposure for 138 subjects from the NINFEA cohort. "*Random*" indicates random permutations of both CpG sites and exposure under study.