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

Maja Popovic, Valentina Fiano, Francesca Fasanelli, Morena Trevisan, Chiara Grasso, Manuela Bianca Assumma, Anna Gillio-Tos, Silvia Polidoro, Laura De Marco, Franca Rusconi, Franco Merletti, Daniela Zugna, Lorenzo Richiardi

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

Epigenomics . 2017 Dec;9(12):1489-1502.
doi: 10.2217/epi-2017-0073. Epub 2017 Nov 6.

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

Abstract

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

Methods: Pairwise correlation coefficients between CpGs selected in two published EWAS, the top hits replication, Bonferroni p-values, Benjamini-Hochberg (BH) FDR and directional FDR r-values were calculated in the NINFEA cohort data. 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 0.12 at genome-wide level to 0.26 among the selected CpGs), affecting the empirical p-value distributions and the usual multiple testing control.

Conclusions: Bonferroni and BH-FDR are 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. 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. In replication studies based on epigenome-wide data, an appropriate null hypothesis must be considered as, for example, done by permutation procedures. Alternatively, directional false-discovery rate (FDR) control for the replicability null hypotheses - the so-called FDR r-value has been recently proposed [9, 10].

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, and then illustrate how this increased correlation may influence the p-value distribution under the null hypothesis and translate into biased interpretations of the results in replication analyses. Finally, we present one of the available methods appropriate for replication studies – r-value - that quantifies the strength of replication taking into account the underlying correlation structure [9].

Materials

Literature dataset

We used findings from two studies assessing DNA methylation in newborns in association with two 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 [11], and 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 [12]. Both studies involved analyses on DNA methylation from cord blood

samples measured using the Infinium HumanMethylation450K BeadChip array. CpG sites for replication were selected by using a fixed threshold of Benjamini and Hochberg FDR-corrected p-values of 0.05.

In addition, a publicly available data set (the Gene Expression Omnibus database accession number GSE77716) with whole blood DNA methylation data measured using the Infinium HumanMethylation450K BeadChip array for 573 participants of Mexican and Puerto Rican descent from the GALA II study [13] was used to determine the correlation between CpG sites selected in the PACE and CHAMACOS study. The complete GALA II data included pre-processed methylation data from 473,838 CpG sites [13].

NINFEA replication study

The selected CpG candidates from the two literature datasets described above were retested in epigenome-wide data coming from the NINFEA birth cohort [14]. 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 during the quality control checks, 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, while information on child's sex is obtained at the first follow-up questionnaire 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. DNA methylation at more than 485,000 CpG sites was measured by the Illumina Infinuim HumanMethylation450 BeadChip and expressed both as percentage (Beta values) and converted to M values by a logit transformation [15]. After quality control checks 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 two literature datasets (the PACE consortium and the CHAMACOS study) we retrieved the published selected altered CpG sites that were then used in the NINFEA and the GALA II datasets. Due to different probes filtering between the NINFEA study and the two literature datasets, there was an incomplete overlap of the top hits.

All the analyses were performed using R statistical computing software (version 3.4.0) and RStudio (version 0.99.491) [16].

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

Correlation analysis

For the two groups of selected CpG sites – derived from the literature examples – we estimated, in 138 subjects from the NINFEA dataset, the partial pairwise Spearman correlation coefficients between the CpG site M values controlling for batch. To ensure that an increased correlation was not influenced by the small sample size or different tissue type we performed sensitivity analyses by

calculating pairwise Spearman correlation coefficients between CpG sites M values measured from whole blood of 573 GALA II participants. The 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 [17]). 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 and GALA II datasets.

Replication analyses

Replication of CpG sites associated with maternal smoking and those associated with child's sex was then conducted in the NINFEA data. It should be noted that the replication analyses were performed only for demonstration purposes, as the NINFEA dataset was underpowered to replicate findings from the discovery studies. Our main aim was to permute and re-analyze the selected exposures in order to show the effect of an increased correlation on the empirical p-value distributions under the null hypothesis (see below).

For both replication analyses we specified models identical to the models of the discovery studies. Replication of the top hits associated with maternal smoking during pregnancy was performed using robust linear regression model adjusted for maternal age, maternal education (low, medium and high), parity and batch. Heteroscedasticity consistent standard errors were calculated using vcovHC function with the HC2 estimator, available in the package sandwich implemented in the R system for statistical computing [18].

Methylation levels at CpG sites selected in the CHAMACOS study were related to child's sex using linear regression models with heteroscedasticity consistent standard errors, adjusted for batch. To improve the models fit, the discovery study on child's sex adjusted the models also for the cell composition estimated directly from the samples [12]. We did not adjust for cell composition as, to the best of our knowledge, no widely-accepted reference data set for the saliva cell composition

exists. The most commonly used reference-free method [19] has been shown to have poor performance in scenarios with binary phenotypes [20], may diminish important phenotypic variation, and we are not aware of studies assessing its performance in saliva samples. Finally, the association between sex and cell composition is unlikely, and even if present the cell composition would likely be on the pathway between child's sex and DNA methylation levels.

Histograms and quantile-quantile (QQ) plots were used to graphically evaluate the observed versus the expected uniform null distribution of p-values. Deviations from the uniform distribution were also formally tested using the Kolmogorov-Smirnov test [21].

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 CpG sites associated with maternal smoking or CpG sites associated with child's sex were estimated in each replicate using the same models as for the replication analyses. 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 [21, 22] and Anderson-Darling tests [22–24]. 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.

Random permutations of the exposure variables within each batch were performed using permute function developed as a part of gtools package [25], skewness was calculated using moments package [26], while foreach package [27] was used for constructing permutation loops.

Multiple testing corrections and r-values

Multiple comparisons correction of the NINFEA results using Bonferroni or Benjamini-Hochberg FDR procedure would not be appropriate due to the underlying correlation structure. Under scenario of highly correlated tests, permutation-based methods are the methods of choice.

Alternatively, Heller et al [9] developed r-values to quantify the evidence for replication while controlling FWER or FDR in genome-wide association studies. This procedure uses multiple testing correction to control for proportion of false replicability claims among all those called replicated when both discovery and replication samples are available. FDR r-value is defined as the lowest FDR at which the finding can be called replicated, and with its modified version accounts for arbitrary dependence between the p-values within the primary study [9]. This method has been further extended [10] to incorporate the direction of observed associations, i.e. to replicate only associations with the same direction in both studies.

For each CpG site of the two datasets (maternal smoking and child's sex) we computed both directional FDR r-values and its modified version that accounts for the underlying correlation (modified r-values) using R function included in the script available in RunMyCode [28]. Default

199 settings were selected for all the parameters included in the r-value computation. A CpG site is
200 considered replicated if the r-value < 0.05 [9].

201 For demonstration purposes we also present p-values corrected using Bonferroni correction and
202 Benjamini-Hochberg FDR procedure [8]. More details on computation of Bonferroni correction
203 (Family-Wise Error Rate [FWER]), Benjamini-Hochberg FDR, FDR r-value and its modified
204 version are summarized in the **Technical note** of the **Supplemental Material**.

205 **Results**

206 ***CpG sites selection***

207 As a result of quality control exclusions and different probes filtering criteria there was an
208 incomplete CpG overlap between literature and the NINFEA EWAS datasets: 4794 CpG sites
209 (78.9% of the selected CpG sites) were included in the analyses on maternal smoking, and 2544
210 CpG sites (83.9% of the selected CpG sites) for the analyses on child's sex. There was a complete
211 overlap between CpG sites selected in the two literature datasets and the GALA II EWAS data.

212 A total of 6 children from the NINFEA data set (4.3%) were exposed to maternal sustained smoking
213 during pregnancy and matched to the unexposed children (N=30) by batch in which samples were
214 analyzed, keeping a constant 1:5 ratio between exposed and unexposed children. Therefore, a total
215 of 36 children were included in the analyses on maternal smoking.

216 The analyses on child's sex were performed in 80 children, by choosing the maximum number of
217 exposed children (females) available within each batch that could be matched with unexposed
218 children (males) from the same batch to keep a constant 1:3 ratio between "exposed" (N=20) and
219 "unexposed" (N=60) subjects.

220 ***Correlation analyses***

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

224 When being pre-selected in the discovery studies, such as in the examples presented here, the
225 average correlation between CpG sites tends to increase depending on the exposure under study.
226 For example, the mean correlation of 0.26 between several thousands of CpG sites associated with
227 maternal smoking during pregnancy was much higher than the original genome-wide mean
228 correlation of 0.12. The variance of correlations in the pre-selected CpG sites also increased
229 substantially compared with the genome-wide CpG sites (all p-values for F test $<2.2x10^{-16}$, visual
230 inspection of **Figure 2**).

231 The same analyses performed on the GALA II data, with DNA methylation levels measured from
232 whole blood in 573 children study, showed similar correlation patterns (**see Supplemental**
233 **Material; see Table S1**). The NINFEA and GALA II datasets had the same mean genome-wide
234 correlation coefficient of 0.12. Compared with the NINFEA study, the mean correlation coefficient
235 in the GALA II study was lower between CpG sites associated with maternal smoking and higher
236 between CpG sites associated with child's sex, (**Table 1**, **see Supplemental Material; see Table**
237 **S1**).

238 When CpG sites from the PACE and CHAMACOS study were selected on the basis of Bonferroni
239 correction, the pairwise correlation coefficients calculated in the NINFEA and GALA II datasets
240 were even higher than when the selection was based on the Benjamini-Hochberg FDR control (data
241 not shown).

242 ***Replication analyses***

243 **Figure 3** reports the p-value distributions and the QQ plots for the replication analyses of the top
244 CpG sites for maternal smoking and child's sex in the NINFEA data. For both exposures, there was
245 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 393 CpG sites (15.5%) with a p-value <0.05 and 1989 CpG sites (78.2%) with the same direction of the effect as in the CHAMACOS study. Maternal smoking during pregnancy was associated with 424 CpG sites (8.8%) at conventional 5% level of significance, and 2199 CpG sites (45.9%) had the same direction of the effect as in the PACE study.

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 variables 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 [24], 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) [29].

To further explore the impact of the correlation structure on the empirical p-value distributions we plotted the skewness of the underlying p-value distributions from the 10,000 replications for each of

271 the examples (**Figure 4**). Symmetric distributions, such as the uniform or normal distribution, have
272 the skewness value zero, while right- or left-skewed distribution have positive or negative values,
273 respectively. The average absolute skewness was 0.34 and 0.22 for 10,000 permutations of maternal
274 smoking and child's sex, respectively. On the contrary, the average absolute skewness was much
275 lower when both, exposures and CpG sites, were selected at random (0.15 for maternal smoking
276 and 0.17 for child's sex). From **Figure 4**, it can be noted that in the presence of a higher correlation
277 between CpG sites, such as in the examples presented here, the skewness of the p-value
278 distributions has a larger variation and is shifted towards positive values (right-skewed
279 distributions) compared to the distributions of genome-wide randomly selected CpG sites. A similar
280 pattern was also observed when all 138 subjects were analyzed with CpG sites associated with
281 maternal smoking during pregnancy (**see Supplemental Material; see Figure S2**), thus ruling out a
282 possible impact of the small sample size on the empirical p-value distributions in the example with
283 maternal smoking during pregnancy.

284 It is noteworthy that the biases that we have so far described are mainly due to the underlying
285 correlation structure. For demonstration purposes we have selected 256 out of 4794 CpG sites
286 related to maternal smoking during pregnancy and 129 out of 2544 CpG sites related to child's sex
287 that have all pairwise correlation coefficients below an arbitrary level of 0.40 in the NINFEA
288 dataset. Mean absolute correlation coefficient was 0.09 for both low-correlated data sets, and thus
289 lower than the underlying genome-wide mean correlation of 0.12.

290 P-value distributions of the 10,000 random permutations of the exposure variables were non-
291 uniform, i.e. associated with a Kolmogorov-Smirnova p-value < 0.05 in 17.0% permutations of
292 maternal smoking and 5.7% permutations of child's sex. The average absolute skewness was 0.09
293 for maternal smoking and 0.10 for child's sex, with standard deviations much smaller than that for
294 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 Bonferroni or Benjamini-Hochberg FDR multiple testing correction. After the Benjamini-Hochberg correction at the 0.05 FDR level methylation levels at fourteen CpG sites were associated with child's sex, while only one CpG site remained associated with maternal smoking during pregnancy, reflecting the small number of exposed subjects (N=6) in the NINFEA dataset (**Table 3**). The two top ranked CpG sites that passed Benjamini-Hochberg correction (both p-values=0.02) remained associated with child's sex also after more conservative, Bonferroni correction (both p-values=0.04), and the only CpG site associated with maternal smoking at 0.05 FDR level remained associated also after Bonferroni correction (p=0.04).

One of the approaches that would be correct for a replication study is the FDR-based replication p-value (r-value). For the analyses on sex differences in methylation levels, only one CpG site was replicated (cg03168896) with the directional FDR r-value=0.04, and it remained replicated in the NINFEA cohort also after considering the underlying correlation (modified r-value=0.04). It should be noted that the methylation level at the replicated cg03168896 was positively associated with female sex both in CHAMACOS and in the NINFEA study, and had a Benjamini-Hochberg FDR p-value=0.02. Other thirteen CpG sites that passed the Benjamini-Hochberg FDR correction, despite having the same direction of the effect in the CHAMACOS and the NINFEA study, 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 0.12 at the genome-wide level to 0.26 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 using only two datasets, one with child saliva DNA methylation,and one with cord blood DNA methylation. The underlying correlation between the pre-selected CpG sites was higher in both datasets compared to the genome-wide mean correlation coefficient of 0.12. 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.

As the examples presented here [11,12], most of the EWAS studies use Benjamini-Hochberg FDR method to adjust for multiple tests, both in the discovery and replication analysis [30, 31]. 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 [32]. The only assumption behind permutation procedures is that the observations are exchangeable under the null hypothesis [32], 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, for example methods dealing specifically with linkage-disequilibrium in GWAS such as p_{ACT} method [33], SNPSpD [34] and permutation-based method by Dudbrige and Koeleman [35], or more general resampling-based FDR for correlated tests [36] and Benjamini-Yekutieli modification of standard FDR [37]. 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.* [38] 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 correlated test statistics, and might be an effective alternative to standard methods also for the replication studies.

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 directional r-values as an FDR-based measure - a valuable method specifically developed for replication studies. The modified version of r-value guarantees false-discovery rate control under arbitrary dependence between tests. Moreover, directional FDR r-values quantify the evidence of replication that accounts for the consistency between the directions

of associations in the discovery and replication studies [10]. In the GWAS context Sofer et al. [10] showed that r-value approach provides better control of false discovery error rate compared to commonly used approaches, while retaining the same power, and a gain in power of the replication study the larger the discovery study is.

The r-value computation largely depends on the nature of the replicability problem and the design of the study. As pointed out in Heller et al. [9] the advantage of combining evidence from the discovery and replication study offers new perspectives for developing methods that take into account the relative importance given to the replication study, i.e. in the context of replication of EWAS findings, the use of unequal penalties to the errors of the discovery and replication studies.

As the directional FDR r-value approach addresses the issues of the consistency in the direction of the effects between the discovery and the replication studies and the underlying correlation between pre-selected CpG sites, we applied this method for demonstration purposes. However, our study was not designed to test the robustness of this method given particular scenarios, or to compare its performance with other available methods dealing with correlated tests in the context of replication studies. Further investigations are required to provide evidence on the gold-standard methods for EWAS replication studies, and best approaches for the determination of sample size in the discovery and replication studies.

One of the limitations of our study is the relatively small sample size used for the replication analyses (36 subjects for analyses on maternal smoking and 80 subjects for analyses on child's sex). In fact, since p-values depend on a combination of sample size and effect size, the NINFEA study was underpowered to replicate the findings, especially in the case of the PACE study that had a much larger sample size compared to the NINFEA study. Our study, however, had illustrative purposes and we showed that a false impression of replication might arise when correlation structure was not taken into account (even in presence of a small sample size for the replication study). Specifically, the main aim of this study was to illustrate how increased correlation in the

replication phase of EWAS influences the empirical p-value distribution, and consequently the usual Bonferroni and Benjamini-Hochberg FDR control. The permutation procedures that we performed were conducted under the null hypothesis, where the issue of small sample size is less relevant. We also conducted sensitivity analyses by considering scenarios of increasing sample size (from 36 to 138 subjects) and showed that the very small sample size did not affect the empirical p-value distribution under the null hypothesis. Moreover, the impact of sample size on the correlation structure has been further evaluated by using an external data set with a sample size of 573.

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. Another option is the computation of r-values, which focus specifically on the strength of replication in the presence of highly correlated tests, as in the context of epigenome-wide replication studies.

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 influences empirical p-value distributions, affecting also the usual control by Benjamini-Hochberg FDR procedure.
- Bonferroni correction and 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.

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.

441

442

443 **References**

444 1. Eckhardt F, Lewin J, Cortese R *et al*. DNA methylation profiling of human chromosomes 6,
445 20 and 22. *Nat Genet* 38(12), 1378–1385 (1987).

446 2. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol* 196(2),
447 261–282 (1987).

448 3. Ong ML, Holbrook JD. Novel region discovery method for Infinium 450K DNA
449 methylation data reveals changes associated with aging in muscle and neuronal pathways.
450 *Aging Cell* 13(1), 142–155 (2014).

451 4. Jaffe AE, Murakami P, Lee H *et al*. Bump hunting to identify differentially methylated
452 regions in epigenetic epidemiology studies. *Int J Epidemiol* 41(1), 200–209 (2012).

453 5. Sofer T, Schifano ED, Hoppin JA, Hou L, Baccarelli AA. A-clustering: a novel method for
454 the detection of co-regulated methylation regions, and regions associated with exposure.
455 *Bioinformatics* 29(22), 2884–2891 (2013).

456 6. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis.
457 *BMC Bioinformatics* 9, 559 (2008).

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

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

462 9. Heller R, Bogomolov M, Benjamini Y. Deciding whether follow-up studies have replicated
463 findings in a preliminary large-scale omics study. *Proc Natl Acad Sci U S A* 111(46),
464 16262–16267 (2014).

10. Sofer T, Heller R, Bogomolov M, et al. A powerful statistical framework for generalization testing in GWAS, with application to the HCHS/SOL. *Genet Epidemiol* 41, 251–258 (2017).
11. 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).
12. 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).
13. Rahmani E, Zaitlen N, Baran Y, et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat Methods* 13(5), 443–445 (2016).
14. 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).
15. 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).
16. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2017). https://www.R-project.org/.
17. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika* 10(4), 507–521 (1915).
18. Zeileis A. Econometric Computing with HC and HAC Covariance Matrix Estimators. *J Stat Softw* 11(10), 1–17 (2004).
19. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* 30(10), 1431–1439 (2014).
20. McGregor K, Bernatsky S, Colmegna I, et al. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biol.* 17, 84 (2016).
21. Massey FJJ. The Kolmogorov–Smirnov test for goodness of fit. *J Am Stat Assoc* 46, 68–78

Formattato: Inglese (Regno Unito)

Formattato: Italiano (Italia)

(1951).

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

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

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

25. Warnes GR, Bolker B, Lumley T. gtools: Various R Programming Tools. (2015). https://CRAN.R-project.org/package=gtools

26. Komsta L, Novomestky F. moments: Moments, cumulants, skewness, kurtosis and related tests. (2015). https://CRAN.R-project.org/package=moments

27. Revolution Analytics, Weston S. foreach: Provides Foreach Looping Construct for R. (2015). https://CRAN.R-project.org/package=foreach

28. Heller R, Bogomolov M, Benjamini Y. Deciding whether follow-up studies have replicated findings in a preliminary large-scale "omics' study" Available at: http://www.runmycode.org/companion/view/542 [Accessed July 17, 2017]

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

30. Morales E, Vilahur N, Salas LA, et al. Genome-wide DNA methylation study in human placenta identifies novel loci associated with maternal smoking during pregnancy. Int J Epidemiol 45(5), 1644-1655 (2016).

31. Gruzieva O, Xu CJ, Breton CV, et al. Epigenome-Wide Meta-Analysis of Methylation in Children Related to Prenatal NO2 Air Pollution Exposure. Environ Health Perspect 125(1), 104-110 (2017).

Formattato: Inglese (Regno Unito)

Formattato: Italiano (Italia)

Formattato: Italiano (Italia)

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

33. 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).

34. 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).

35. 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).

36. 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).

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

38. 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 partial 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	321,084	0.01	0.12	0.09	0.47
Child's sex	2544	0.01	0.18	0.13	0.64
Maternal smoking during pregnancy	4794	0.01	0.26	0.19	0.77

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	98.4
Random CpG sites	91.4
Child's sex	95.3
Random CpG sites	91.8

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

580

581

582

583

584

585

586

587

588

589 **Table 3.** Smoking- and sex-associated CpG sites that passed Benjamini-Hochberg (BH) FDR correction in the NINFEA replication study, and
590 corresponding discovery and replication p-values, FWER (Bonferroni-corrected p-values), BH FDR p-values, FDR r-values and modified r-values.

Smoking-associated CpG sites	**Discovery study two-sided p-value**	**Replication study two-sided p-value**	**FWER**	**BH FDR p-value**	**FDR r-value[a]**	**Modified r-value[b]**
cg12793610	4.42e-05	9.12e-06	0.04	0.04	0.91	1.00
Sex-associated CpG sites						
cg23092538	7.43e-05	1.69e-05	0.04	0.02	0.18	0.74
cg03168896	1.86e-08	1.73e-05	0.04	0.02	0.04	0.04
cg14022202	1.17e-05	2.55e-05	0.06	0.02	0.16	0.39
cg25438440	3.72e-18	6.76e-05	0.17	0.04	0.07	0.08
cg15089217	8.44e-06	9.52e-05	0.24	0.04	0.12	0.36
cg19544707	8.12e-12	9.98e-05	0.25	0.04	0.07	0.08
cg12763978	1.13e-06	1.17e-04	0.30	0.04	0.07	0.26
cg03298305	5.27e-04	1.38e-04	0.35	0.04	0.31	1.00
cg23332732	1.68e-05	1.38e-04	0.35	0.04	0.17	0.41
cg26955850	5.55e-04	1.44e-04	0.37	0.04	0.33	1.00
cg14546619	1.57e-04	1.67e-04	0.42	0.04	0.24	1.00
cg01063965	3.42e-06	1.67e-04	0.42	0.04	0.08	0.27
cg26213873	3.34e-18	2.15e-04	0.55	0.04	0.09	0.14
cg18305433	2.24e-05	2.17e-04	0.55	0.04	0.17	0.41

[a] Directional FDR r-value
[b] Conservative r-value modification that accounts for arbitrary dependence between tests

Formattato: Larghezza 20,99 cm, Altezza: 29,7 cm, Numerazione: continua

Figure 1. The main steps of the analysis

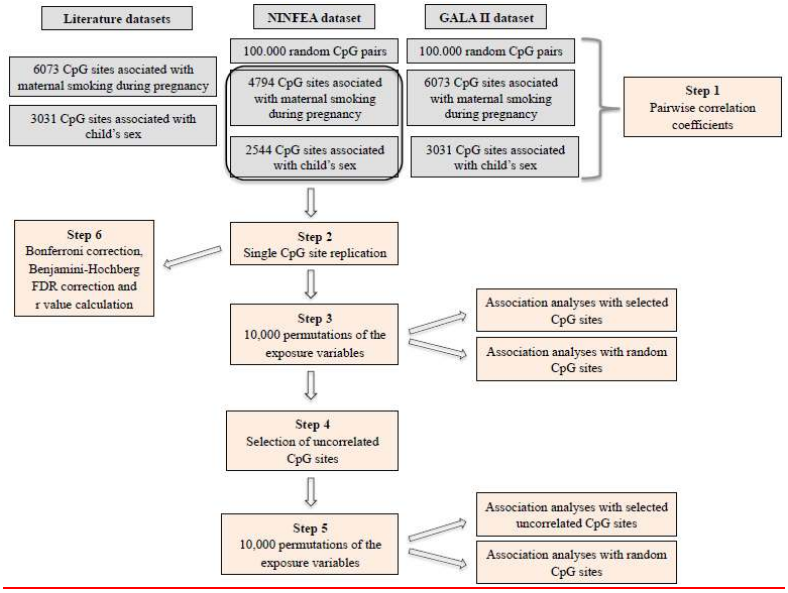
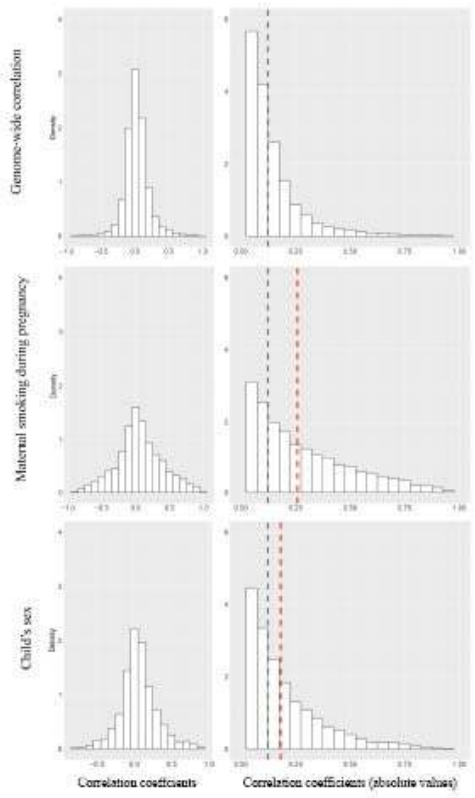


Figure 2. Distribution of correlation coefficients (left side) and their absolute values (right side) for genome-wide CpG sites, 4794 CpG sites associated with maternal smoking during pregnancy, and 2544 CpG sites associated with child's sex. Vertical gray line indicates genome-wide mean correlation coefficient (absolute values). Vertical red lines indicate mean correlations coefficients (absolute values) for each set of the selected CpG sites.



599

600 **Figure 3.** Replication in the NINFEA cohort: Distribution of replication p-values and Q-Q plots of
601 observed versus expected p-values for the associations between methylation levels at smoking-
602 related (N=4794) and sex-related (N=2544) CpG sites and maternal smoking during pregnancy, and
603 child's sex.

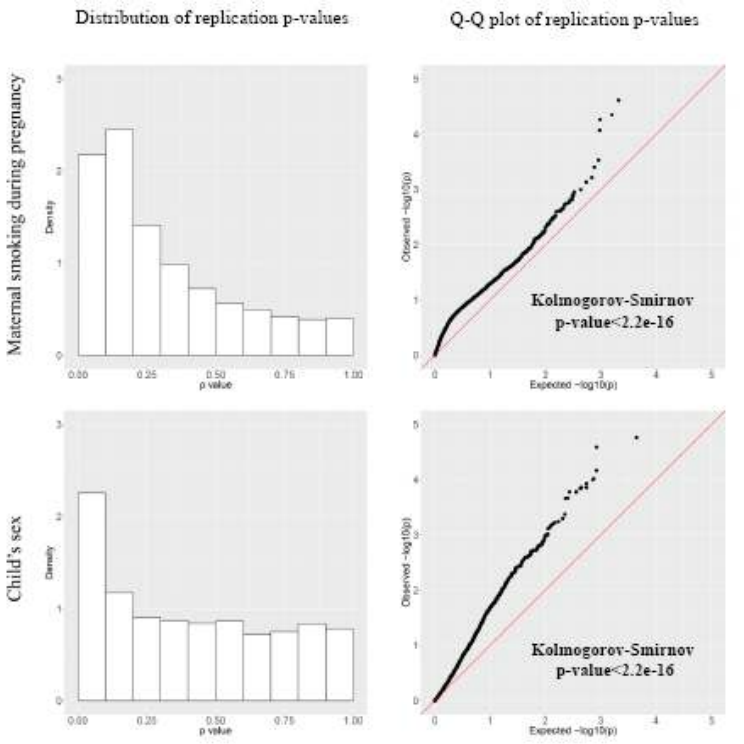


Figure 4. Skewness of p-value distributions from the analyses of the association between smoking-related (N=4794) and sex-related (N=2544) CpG sites and permutations of maternal smoking during pregnancy and child's sex from 10,000 replications. "*Random*" indicates random permutations of both CpG sites and exposure under study.

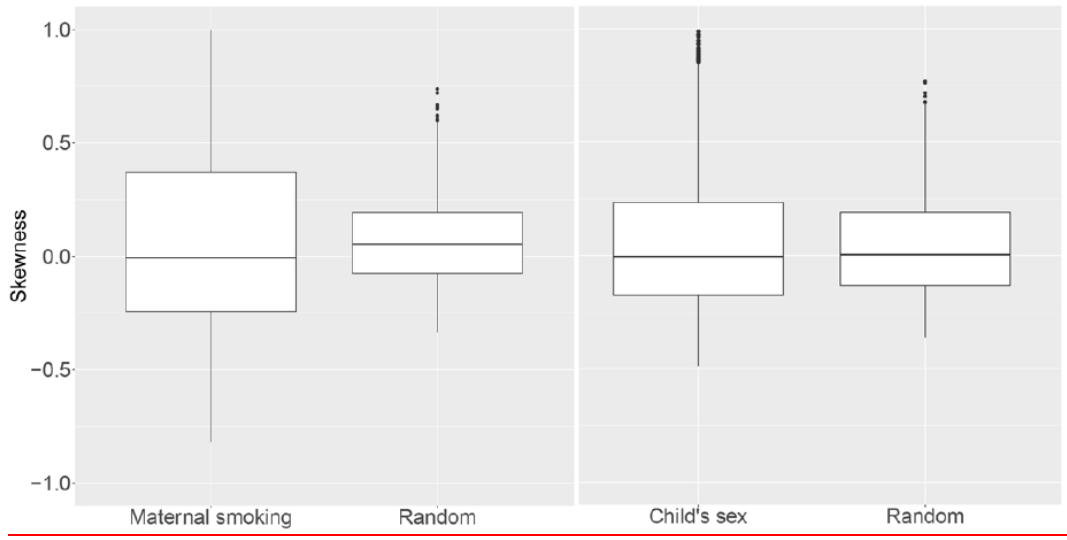
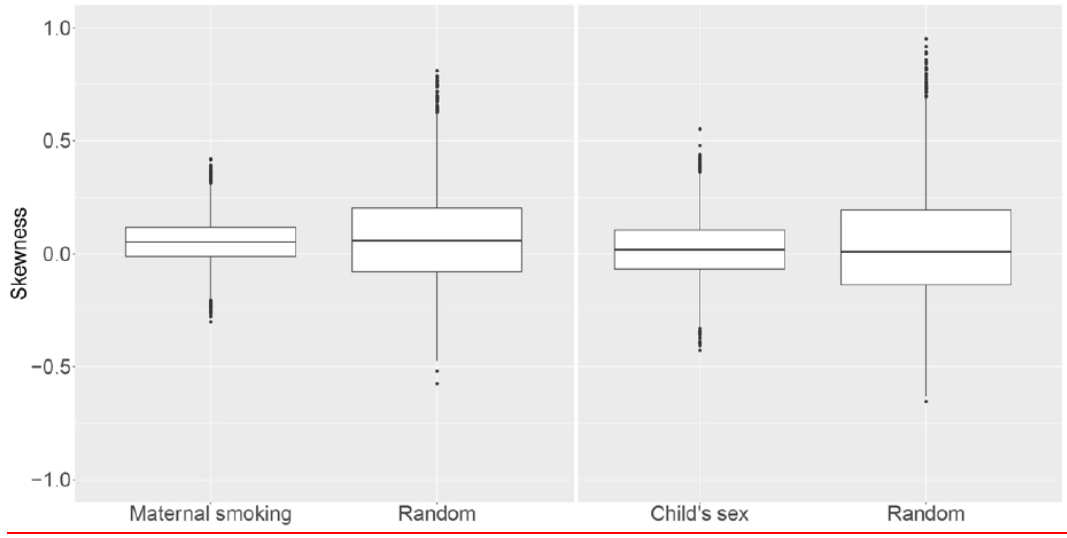


Figure 5. Skewness of p-value distributions from the analyses of the association between smoking-related "low-correlated" (N=256) and sex-related "low-correlated" (N=129) pre-selected CpG sites and permutations of maternal smoking/child's sex from 10,000 replications. "*Random*" indicates random permutations of both CpG sites and exposure under study.



Supplemental Material

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

Table of Contents

Methods

Technical note

We assume that we are testing m independent null hypotheses H_{01}, H_{02},...,H_{0m}. The possible outcomes when testing m hypotheses simultaneously are summarized as follows:

	Rejecting Ho	Accepting Ho	Total
True null hypothesis	V	U	m_0
False null hypothesis	S	T	m_1
Total	R	W	m

where

- V is the number of false rejections (or false discoveries),
- U is the number of true acceptances,
- S is the number of true rejections,
- T is the number of false acceptances.

The total number of true null hypotheses, m_0, is fixed but unknown. Random variables V, S, U, and T are not observable, while the random variables R=S+V and W=U+T, the number of rejected and accepted null hypotheses, respectively, are observable.

In a single study analysis, there are two different approaches to address the issue of multiple testing: the family wise error rate (FWER) and the false discovery rate (FDR).

FWER
It is the probability of falsely rejecting at least one null hypothesis. In formula:

$$FWER = P(V \geq 1)$$

FDR
It is the expected proportion of falsely rejected hypotheses among all rejected hypotheses. In formula:

$$FDR = E\left[\frac{V}{\max(R,1)}\right]$$

The maximum between "R" and 1 guarantees that FDR is equal to 0 when no hypothesis is rejected.

In Heller et al.[1] a generalization of FWER and FDR was developed in order to give a formal method to declare that findings from a discovery study have been replicated in a replication study.

Consider a family of null hypotheses H_j tested in each of two independent studies. Let h_{ij} be the indicator of whether H_j is false in study i:

$h_{ij} = 0$ if H_j is true in study i (i.e. $\beta = 0$)
$h_{ij} = 1$ if H_j is false in study i (i.e. $\beta \neq 0$)

where i=1,2 (1=discovery study; 2=replication study) and j is the index that refers to a specific test, hereafter referred as locus in the context of epigenome-wide association studies.
Let $\mathcal{H}_j$ be the set of the four possible results for the specific locus j:

$$\mathcal{H}_j = \{H_j = (h_{1j}, h_{2j}) : h_{ij} \in \{0,1\}\}$$
$$= \{(0,0), (0,1), (1,0), (1,1)\}.$$

R is the total number of replicability claims. Denote $S = R_{11}$ the number of true positives and $R - S = R_{00} + R_{01} + R_{10}$ the number of false positives. Note that in a single study V is the number of false positives, while in a discovery and replication analysis the number of false positives is the sum of the three terms ($R_{00} + R_{01} + R_{10}$).

The FWER and FDR for replicability analysis are defined as:

$$FWER_r = P(R - S \geq 1).$$

$$FDR_r = E\left[\frac{R - S}{\max(R, 1)}\right].$$

The $FWER_r/FDR_r$ r-value for a specific locus is defined as the lowest FWER/FDR level at which we can say that the finding has been significantly replicated.

These definitions of r-values do not account for the direction of the observed association. For this reason the r-values approach was then extended by Sofer et al.[2] to incorporate the direction of observed associations. Define the left-sided (right-sided) alternative as the scenario in which a given locus is negatively (positively) associated with an exposure/outcome in a given study. Let

$h_{ij} = 1$ if the right-sided alternative is true for locus j in study i (i.e. β >0)
$h_{ij} = 0$ if the right-sided alternative is true for locus j in study i (i.e. β =0)
$h_{ij} = -1$ if the left-sided alternative is true for locus j in study i (i.e. β <0)

where i=1,2 (1=discovery study; 2=replication study) and j is the index that refers to a specific locus.

Let $\mathcal{H}_j$ be the set of the nine possible results for the specific locus j:

$$\mathcal{H}_j = \{H_j = (h_{1j}, h_{2j}) : h_{ij} \in \{-1,0,1\}\}$$
$$= \{(-1,-1), (-1,0), (-1,1), (0,-1), (0,0), (0,1), (1,-1), (1,0), (1,1)\}.$$

Suppose that R is the total number of replicability claims, i.e. the number of rejected hypotheses in the replication analysis. Denote R_j^R and R_j^L the indicators of whether the null rejections are made in the right or left direction, respectively, for locus j. The number of erroneously rejected hypotheses is $R - S$, where :

$$S = \sum_{\{j: H_j=(1,1)\}} R_j^R + \sum_{\{j: H_j=(-1,-1)\}} R_j^L$$

The directional replication FWER and FDR are defined as:

$$FWER_{r_{dir}} = P(R - S \geq 1),$$

$$FDR_{r_{dir}} = E\left[\frac{R-S}{\max(R,1)}\right].$$

The $FWER_{r_{dir}}/FDR_{r_{dir}}$ r-value for a specific locus is defined as the lowest FWER/FDR level at which we can say that the locus association has been significantly replicated with the same direction.

The FWER/FDR controlling procedures for testing the family of no replicability null hypotheses in the replication studies are described in Heller et al.[1] and Sofer et al.[2] for r-values and directional r-values, respectively. These procedures require data and parameters as input for r-values computation:

1. *m*, the number of hypotheses examined in the discovery study;
2. R_1, the set of loci selected for replication based on the discovery study results;
3. the directional p-values for the followed-up loci $\{(p_{1j}, p_{2j}): j \in R_1\}$;
4. $l_{00} \in [0,1)$, the user-specified lower bound on the fraction of locus associations, out of the *m* loci examined in the discovery study, that are null in both studies (default value $l_{00} = 0.8$);
5. $c_2 \in (0,1)$, the emphasis given to the follow-up study (default value $c_2 = 0.5$).

These procedures declare as replicated all findings with FWER/FDR r-values ≤ q.

Heller et al.[1] gave a theorem that shows that:

- if the p-values in the discovery study are independent, and the p-values from the replication study are jointly independent or are positive regression dependent on the subset of null hypotheses, then the FWER/FDR on false replicability claims is controlled at level q;
- for arbitrary dependence among the p-values in the discovery study, replacing *m* by

  $$m^* = m \cdot \sum_{i=1}^{m} \frac{1}{i}$$

  in the r-value computation, the FWER/FDR on false replicability claims is controlled at level q.

The procedure with *m** instead of *m* computes the modified r-values that takes into account arbitrary dependencies among tests.

In this paper, we computed FWER, FDR, directional FDR r-values (r-values) and directional FDR r-values with m* modification for arbitrary dependence among p-values (modified r-values).

References

1. 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).

2. Sofer T, Heller R, Bogomolov M, et al. A powerful statistical framework for generalization testing in GWAS, with application to the HCHS/SOL. Genet Epidemiol 41, 251–258 (2017).

Results

Table S1. Summary statistics of the correlation coefficients' distributions, expressed as absolute values, in 573 children of the GALA II study

Set of CpG sites	N	3rd percentile	Mean	Median	97th percentile
Genome-wide	473,838	0.01	0.12	0.10	0.32
Child's sex	3031	0.01	0.20	0.18	0.48
Maternal smoking during pregnancy	6073	0.01	0.15	0.13	0.45

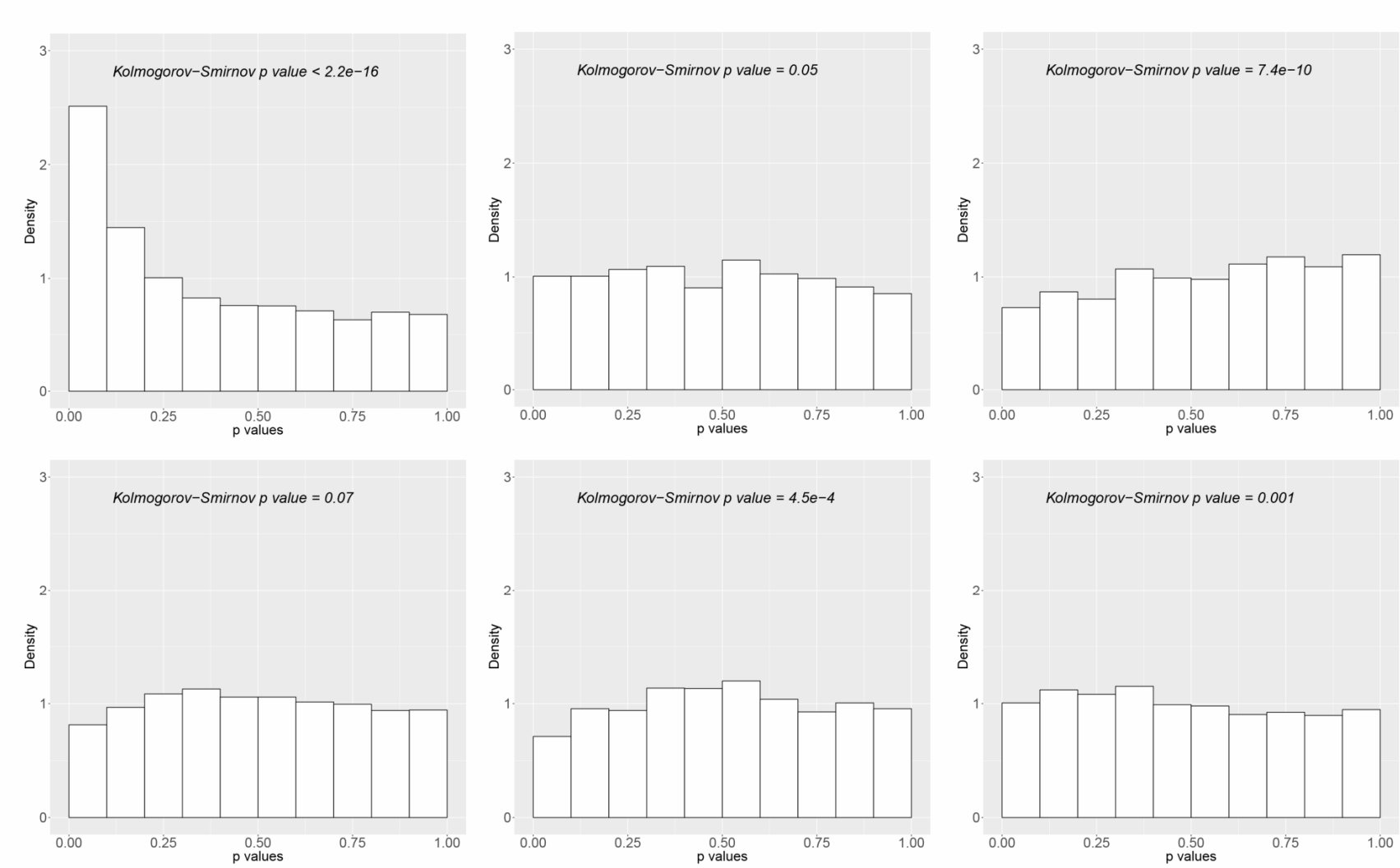


Figure S1. Histograms of p-value distributions from random permutations and Kolmogorov-Smirnov p value assessing whether the observed p-value distributions come from a hypothesized uniform distribution

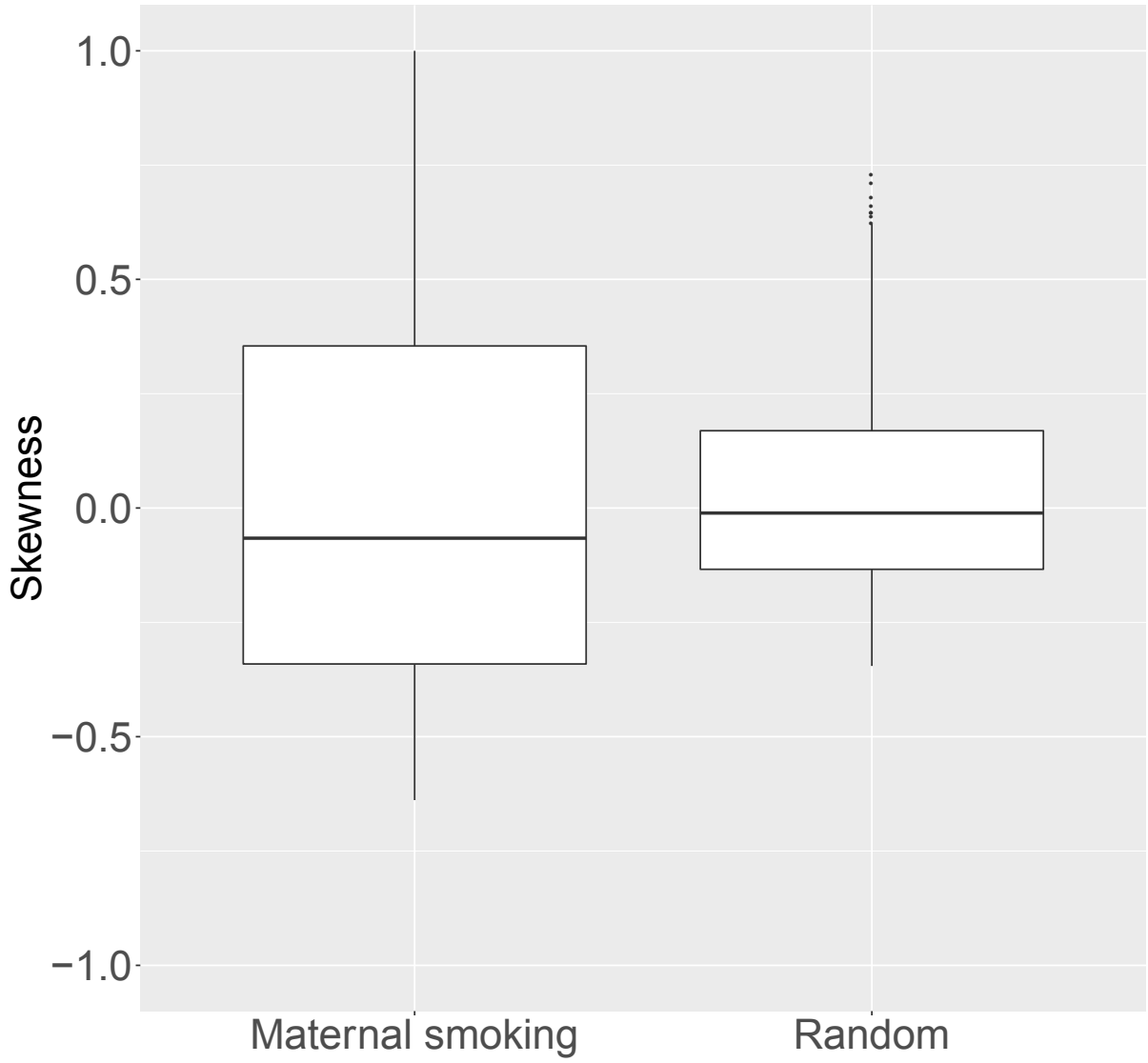


Figure S2. Skewness of p-value distributions from the analyses of the association between 4794 CpG sites associated with maternal smoking and 10,000 permutations of an imaginary exposure for 138 subjects from the NINFEA cohort. "*Random*" indicates random permutations of both CpG sites and exposure under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

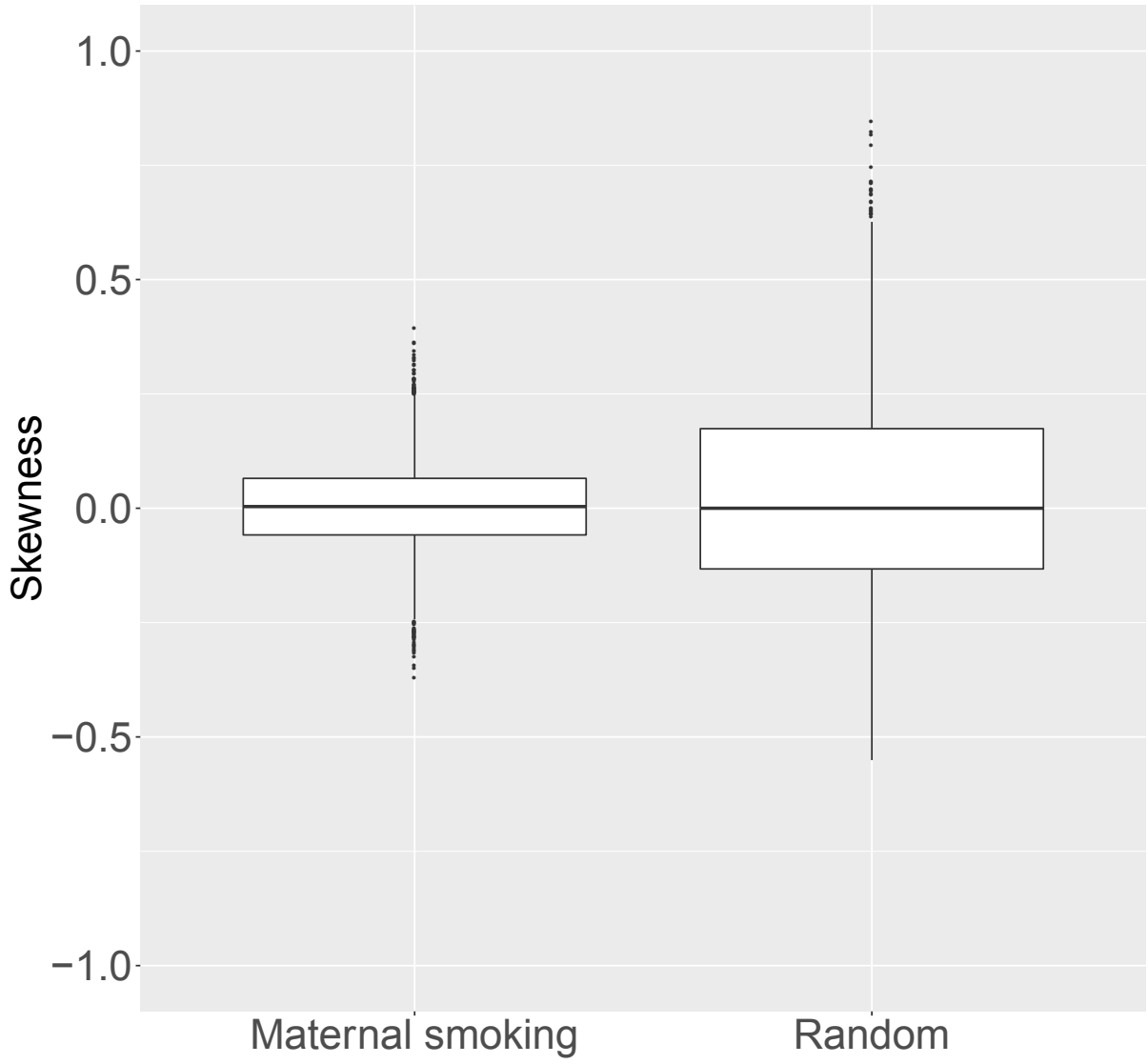


Figure S3. Skewness of p-value distributions from the analyses of the association between 256 low-correlated CpG sites associated with maternal smoking and 10,000 random permutations of an imaginary exposure for 138 subjects from the NINFEA cohort. "*Random*" indicates random permutations of both CpG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