



AperTO - Archivio Istituzionale Open Access dell'Università di Torino

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

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/1652237> since 2021-09-02T14:05:25Z

Published version:

DOI:10.2217/epi-2017-0073

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)



This is the author's final version of the contribution published as:

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.

The publisher's version is available at:

<http://hdl.handle.net/2318/1652237>

When citing, please refer to the published version.

Link to this full text:

<http://hdl.handle.net/2318/1652237>

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

3
4
5 **Abstract**

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

8 **Methods:** Pairwise correlation coefficients between CpGs selected in two published EWAS, the top
9 hits replication, Bonferroni p-values, Benjamini-Hochberg (BH) FDR and directional FDR r-values
10 were calculated in the NINFEA cohort data. Exposures' random permutations were performed to
11 show the empirical p-value distributions.

12 **Results:** The average pairwise correlation coefficients between CpGs were enhanced after selection
13 for the replication (e.g. from 0.12 at genome-wide level to 0.26 among the selected CpGs), affecting
14 the empirical p-value distributions and the usual multiple testing control.

15 **Conclusions:** Bonferroni and BH-FDR are inappropriate for the EWAS replication phase, and
16 methods that account for the underlying correlation need to be used.

17

18

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

20

21

22

23

24

25

26 **Introduction**

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

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

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

48 Under such scenario, it is intuitive that the average pairwise correlation between single sites in the
49 large discovery EWAS will be lower than the average pairwise correlation between the few
50 hundreds of single sites selected for the replication study. This fact is rarely taken into consideration

51 in EWAS replication studies and the analyses in the replication sample may, thus, be biased.
52 Benjamini-Hochberg False-discovery rate (FDR) correction [8], which is typically used both in the
53 discovery and replication phase of the epigenome-wide studies is robust, yet does not take into
54 account the underlying correlation structure. For what we have said insofar, the robustness of the
55 procedure to the lack of independence is much more important for the replication than for the
56 discovery study. In replication studies based on epigenome-wide data, an appropriate null
57 hypothesis must be considered as, for example, done by permutation procedures. Alternatively,
58 directional false-discovery rate (FDR) control for the replicability null hypotheses - the so-called
59 FDR r-value has been recently proposed [9, 10].

60 This article has an illustrative intent. We first show with real data examples that the average
61 pairwise correlation between CpG sites increases after selection through an epigenome-wide
62 discovery analysis, and then illustrate how this increased correlation may influence the p-value
63 distribution under the null hypothesis and translate into biased interpretations of the results in
64 replication analyses. Finally, we present one of the available methods appropriate for replication
65 studies – r-value – that quantifies the strength of replication taking into account the underlying
66 correlation structure [9].

67 **Materials**

68 *Literature dataset*

69 We used findings from two studies assessing DNA methylation in newborns in association with two
70 different exposures: i) a study on 6685 children from the Pregnancy and Childhood Epigenetics
71 (PACE) consortium that identified 6073 over 464,628 CpG sites whose methylation levels were
72 associated with maternal sustained smoking during pregnancy [11], and ii) a study on sex
73 differences in DNA methylation in 111 Mexican-American newborns, members of the
74 CHAMACOS study, that identified 3031 over 410,072 CpG site candidates located on the
75 autosomal chromosomes [12]. Both studies involved analyses on DNA methylation from cord blood

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

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

85 ***NINFEA replication study***

86 The selected CpG candidates from the two literature datasets described above were retested in
87 epigenome-wide data coming from the NINFEA birth cohort [14]. The study design was a nested
88 case-control study on 72 cases with at least one reported episode of wheezing between 6 and 18
89 months of age and 72 controls matched to cases by sex, age at sampling and seasonality/calendar
90 year of sampling. In the NINFEA birth cohort saliva samples are routinely collected from infants at
91 approximately 6 months of age using a mailed Oragene self-collection kit, and in the nested case
92 control study we focused on saliva DNA methylation markers of childhood wheezing (data not
93 published). DNA extracted from the saliva samples of cases and matched controls was assessed for
94 epigenome-wide methylation using the Illumina Infiniuim HumanMethylation450 BeadChip. Three
95 cases and three matched controls were excluded during the quality control checks, leading to a total
96 of 138 subjects available for the analyses. The baseline NINFEA questionnaire is completed by
97 mothers during pregnancy and includes questions on sustained smoking in pregnancy, while
98 information on child's sex is obtained at the first follow-up questionnaire completed 6 months after
99 delivery.

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

104 **Methods**

105 *Statistical analyses*

106 NINFEA cases and controls were pooled together. DNA methylation at more than 485,000 CpG
107 sites was measured by the Illumina Infinium HumanMethylation450 BeadChip and expressed both
108 as percentage (Beta values) and converted to M values by a logit transformation [15]. After quality
109 control checks and probes filtering (probes corresponding the SNPs inside the probe body and SNPs
110 at CpG sites, cross hybridizing and probes on the sex chromosomes) a total of 321,084 probes were
111 available in the NINFEA dataset.

112 For two literature datasets (the PACE consortium and the CHAMACOS study) we retrieved the
113 published selected altered CpG sites that were then used in the NINFEA and the GALA II datasets.
114 Due to different probes filtering between the NINFEA study and the two literature datasets, there
115 was an incomplete overlap of the top hits.

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

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

119 *Correlation analysis*

120 For the two groups of selected CpG sites – derived from the literature examples – we estimated, in
121 138 subjects from the NINFEA dataset, the partial pairwise Spearman correlation coefficients
122 between the CpG site M values controlling for batch. To ensure that an increased correlation was
123 not influenced by the small sample size or different tissue type we performed sensitivity analyses by

124 calculating pairwise Spearman correlation coefficients between CpG sites M values measured from
125 whole blood of 573 GALA II participants. The distributions of correlation coefficients were
126 compared with the distribution of genome-wide pairwise correlation coefficients between CpG sites
127 (histograms, summary statistics with the 3rd, 50th and 97th percentiles, F test on homogeneity of
128 variance on Fisher's zeta transformation [17]). To obtain the genome-wide correlation distribution,
129 we calculated the pairwise correlation coefficients between 100,000 randomly selected CpG pairs
130 among all available CpG sites in the NINFEA and GALA II datasets.

131 *Replication analyses*

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

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

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

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

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

157 *Assessment of the empirical p-value distributions*

158 To evaluate the impact of the increased correlation among the selected CpG sites, we assessed the
159 p-value distributions under the null-hypothesis of no effects of the exposures on the methylation
160 levels in the selected CpG sites. For this purpose, we generated 10,000 random shuffling of the
161 exposed-unexposed status for each individual in the two datasets (maternal smoking during
162 pregnancy and child's sex) while maintaining the same ratio between exposed and unexposed
163 subjects within each batch as in the original data. The associations between the randomly attributed
164 exposure and methylation in the CpG sites associated with maternal smoking or CpG sites
165 associated with child's sex were estimated in each replicate using the same models as for the
166 replication analyses. P-value distributions of the 10,000 replicates were described in terms of
167 symmetry by estimating the skewness and in terms of deviation from a uniform distribution by
168 performing Kolmogorov-Smirnov [21, 22] and Anderson-Darling tests [22–24]. To compare
169 empirical distributions, we generated additional 10,000 replicates for both examples (maternal
170 smoking and child's sex) with random assignment of the exposure variables and random CpG sites
171 selection.

172 To ensure that the low exposure frequency in the analyses on maternal smoking did not affect the
173 underlying distribution under the null hypothesis, we analyzed all NINFEA subjects with available

174 EWAS data by shuffling the imaginary exposure with 69 “cases” and 69 “controls” and relating it
175 to methylation levels in 4794 smoking-related CpG sites.

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

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

184 *Multiple testing corrections and r-values*

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

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

196 For each CpG site of the two datasets (maternal smoking and child’s sex) we computed both
197 directional FDR r-values and its modified version that accounts for the underlying correlation
198 (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 $<2.2 \times 10^{-16}$, 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

246 (Kolmogorov-Smirnov p-value <2.2 x 10⁻¹⁶ in both analyses). The analysis on child's sex revealed
247 393 CpG sites (15.5%) with a p-value <0.05 and 1989 CpG sites (78.2%) with the same direction of
248 the effect as in the CHAMACOS study. Maternal smoking during pregnancy was associated with
249 424 CpG sites (8.8%) at conventional 5% level of significance, and 2199 CpG sites (45.9%) had the
250 same direction of the effect as in the PACE study.

251 ***Assessment of the empirical p-value distributions***

252 In the absence of correlation, by randomly permuting and re-analyzing the data we would expect
253 the p-value distribution to be approximately uniform in most of the replications. Distributions as
254 those observed in **Figure 3** - skewed versus lower p-values - are expected to be seen in a small
255 proportion of the replications. After visual inspection of the p-value distribution histograms from
256 the 10,000 random permutations of the exposure variables we noticed that the percentage of
257 replications not following the uniform p-value distribution was much higher than the expected 5%,
258 both in the case of pre-selected CpG sites and in the case of genome-wide randomly selected CpG
259 sites.

260 In fact, Kolmogorov-Smirnov p-values were low even when the p-value distribution histograms
261 visually showed quite uniform patterns (**see Supplemental Material; see Figure S1**). Accordingly,
262 as reported in **Table 2**, more than 90% of the replications were associated with a Kolmogorov-
263 Smirnov p-value < 0.05. This proportion was higher in the case of pre-selected than randomly
264 selected CpG sites. The Anderson-Darling test, considered more sensitive to the tails of a
265 distribution than the Kolmogorov-Smirnov test [24], gave similar results (data not shown).
266 However, it should be considered that, with large sample sizes, these test are likely to give strong
267 evidence against the null hypothesis (i.e. they are able to detect even small departures from the
268 theoretical distribution) [29].

269 To further explore the impact of the correlation structure on the empirical p-value distributions we
270 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-Smirnov 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

295 256 CpG sites associated with maternal smoking were performed in all 138 subjects from the
296 NINFEA data (see **Supplemental Material**; see **Figure S3**).

297 ***Multiple testing correction and r-values for replicability***

298 After the initial replication performed in Step 2 (**Figure 1**, **Figure 3**) a standard naïve and incorrect
299 practice would then be to consider the results of the single CpG sites, after implementing some of
300 the procedures that take into account multiple testing and reduce the number of false positives, such
301 as Bonferroni or Benjamini-Hochberg FDR multiple testing correction. After the Benjamini-
302 Hochberg correction at the 0.05 FDR level methylation levels at fourteen CpG sites were associated
303 with child's sex, while only one CpG site remained associated with maternal smoking during
304 pregnancy, reflecting the small number of exposed subjects (N=6) in the NINFEA dataset (**Table**
305 **3**). The two top ranked CpG sites that passed Benjamini-Hochberg correction (both p-values=0.02)
306 remained associated with child's sex also after more conservative, Bonferroni correction (both p-
307 values=0.04), and the only CpG site associated with maternal smoking at 0.05 FDR level remained
308 associated also after Bonferroni correction (p=0.04).

309 One of the approaches that would be correct for a replication study is the FDR-based replication p-
310 value (r-value). For the analyses on sex differences in methylation levels, only one CpG site was
311 replicated (cg03168896) with the directional FDR r-value=0.04, and it remained replicated in the
312 NINFEA cohort also after considering the underlying correlation (modified r-value=0.04). It should
313 be noted that the methylation level at the replicated cg03168896 was positively associated with
314 female sex both in CHAMACOS and in the NINFEA study, and had a Benjamini-Hochberg FDR p-
315 value=0.02. Other thirteen CpG sites that passed the Benjamini-Hochberg FDR correction, despite
316 having the same direction of the effect in the CHAMACOS and the NINFEA study, were not
317 replicated (**Table 3**). No CpG site was replicated for maternal smoking during pregnancy.

318 **Discussion**

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

325 As shown in this paper, a certain degree of correlation already affects the discovery phase of
326 EWAS, when analyses are carried out at the genome-wide level. This underlying correlation
327 structure is enhanced in large sample size studies of exposures/outcomes that broadly affect DNA
328 methylation, in which thousands of candidate CpG sites are selected for replication. The increase in
329 correlation can be substantial: in one of the examples that we evaluated in this paper the mean pair-
330 wise correlation coefficient increased from 0.12 at the genome-wide level to 0.26 among the
331 selected CpG sites. Thus, the independency assumption of standard multiple testing procedures can
332 be seriously violated, resulting in spurious replication findings. It should be noted that we analyzed
333 the correlation structure using only two datasets, one with child saliva DNA methylation, and one
334 with cord blood DNA methylation. The underlying correlation between the pre-selected CpG sites
335 was higher in both datasets compared to the genome-wide mean correlation coefficient of 0.12.
336 Average correlation at genome-wide level and that of pre-selected CpG sites might be different in
337 other data sets, populations, age groups or tissues/biofluids.

338 As the examples presented here [11,12], most of the EWAS studies use Benjamini-Hochberg FDR
339 method to adjust for multiple tests, both in the discovery and replication analysis [30, 31]. We argue
340 that in situations of high correlation it is important to explore its magnitude by conducting
341 permutations in which the exposure/outcome status is randomly shuffled. The so-called permutation
342 procedure that empirically generates a model-free p-value is based on this approach, and it is robust
343 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 Dudbridge 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 *et 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

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

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

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

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

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

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

406 **Conclusions**

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

418

419

420 **Executive summary**

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

433 **Ethical conduct of research**

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

436

437

438

439

440

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).

- 465 10. Sofer T, Heller R, Bogomolov M, et al. A powerful statistical framework for generalization
466 testing in GWAS, with application to the HCHS/SOL. *Genet Epidemiol* 41, 251–258 (2017).
- 467 11. Joubert BR, Felix JF, Yousefi P et al. DNA Methylation in newborns and maternal smoking
468 in pregnancy: genome-wide consortium meta-analysis. *Am J Hum Genet* 98(4), 680–696
469 (2016).
- 470 12. Yousefi P, Huen K, Davé V, Barcellos L, Eskenazi B, Holland N. Sex differences in DNA
471 methylation assessed by 450 K BeadChip in newborns. *BMC Genomics* 16, 911 (2015).
- 472 13. Rahmani E, Zaitlen N, Baran Y, et al. Sparse PCA corrects for cell type heterogeneity in
473 epigenome-wide association studies. *Nat Methods* 13(5), 443–445 (2016).
Formattato: Italiano (Italia)
- 474 14. Richiardi L, Baussano I, Vizzini L et al. Feasibility of recruiting a birth cohort through the
475 Internet: the experience of the NINFEA cohort. *Eur J Epidemiol* 22, 831–837 (2007).
- 476 15. Du P, Zhang X, Huang C-C et al. Comparison of Beta-value and M-value methods for
477 quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11, 587 (2010).
- 478 16. R Core Team. R: A language and environment for statistical computing. R Foundation for
479 Statistical Computing, Vienna, Austria. (2017). <https://www.R-project.org/>.
- 480 17. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples of
481 an indefinitely large population. *Biometrika* 10(4), 507–521 (1915).
- 482 18. Zeileis A. Econometric Computing with HC and HAC Covariance Matrix Estimators. *J Stat
483 Softw* 11(10), 1–17 (2004).
- 484 19. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of
485 DNA methylation data. *Bioinformatics* 30(10), 1431–1439 (2014).
- 486 20. McGregor K, Bernatsky S, Colmegna I, et al. An evaluation of methods correcting for cell-
487 type heterogeneity in DNA methylation studies. *Genome Biol.* 17, 84 (2016).
- 488 21. Massey FJJ. The Kolmogorov–Smirnov test for goodness of fit. *J Am Stat Assoc* 46, 68–78

- 489 (1951).
- 490 22. Razali NM, Wah YB. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov,
491 Lilliefors and Anderson–Darling tests. *J. Stat. Modeling Anal* 2, 21–33 (2011).
- 492 23. Anderson TW, Darling DA. A test of goodness of fit. *J Am Stat Assoc* 49(268), 765–769
493 (1954).
- 494 24. Stephens MA. EDF statistics for goodness of fit and some comparisons. *J Am Stat Assoc* 69,
495 730–737 (1974).
- 496 25. Warnes GR, Bolker B, Lumley T. gtools: Various R Programming Tools. (2015).
497 <https://CRAN.R-project.org/package=gtools>
- 498 26. Komsta L, Novomestky F. moments: Moments, cumulants, skewness, kurtosis and related
499 tests. (2015). <https://CRAN.R-project.org/package=moments>
- 500 27. Revolution Analytics, Weston S. foreach: Provides Foreach Looping Construct for R.
501 (2015). <https://CRAN.R-project.org/package=foreach>
- 502 28. Heller R, Bogomolov M, Benjamini Y. Deciding whether follow-up studies have replicated
503 findings in a preliminary large-scale "omics" study" Available at:
504 <http://www.rummycode.org/companion/view/542> [Accessed July 17, 2017]
- 505 29. Lin MF, Lucas HC, Shmueli G. Too big to fail: large samples and the p-value problem.
506 *Inform Syst Res* 24, 906-917 (2013).
- 507 30. Morales E, Vilahur N, Salas LA, et al. Genome-wide DNA methylation study in human
508 placenta identifies novel loci associated with maternal smoking during pregnancy. *Int J
509 Epidemiol* 45(5), 1644-1655 (2016).
- 510 31. Gruzieva O, Xu CJ, Breton CV, et al. Epigenome-Wide Meta-Analysis of Methylation in
511 Children Related to Prenatal NO₂ Air Pollution Exposure. *Environ Health Perspect* 125(1),
512 104-110 (2017).

Formattato: Inglese (Regno Unito)

Formattato: Italiano (Italia)

Formattato: Italiano (Italia)

- 513 32. Good P. Permutation tests: A practical guide to resampling methods for testing hypotheses,
514 2nd edition. Springer-Verlag, New York, NY (1994).
- 515 33. Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of P
516 values for multiple correlated tests. *Am J Hum Genet* 81(6), 1158–1168 (2007).
- 517 34. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in
518 linkage disequilibrium with each other. *Am J Hum Genet* 74(4), 765–769 (2004).
- 519 35. Dudbridge F, Koeleman BPC. Efficient computation of significance levels for multiple
520 associations in large studies of correlated data, including genomewide association studies.
521 *Am J Hum Genet* 75(3), 424–435 (2004).
- 522 36. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test
523 procedures for correlated test statistics. *J Stat Plan Infer* 82, 171–196 (1999).
- 524 37. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under
525 dependency. *Ann Stat* 29(4), 1165–1188 (2001).
- 526 38. van Iterson M, van Zwet EW, BIOS Consortium, Heijmans BT. Controlling bias and
527 inflation in epigenome- and transcriptome-wide association studies using the empirical null
528 distribution. *Genome Biol* 18(1), 19 (2017).
- 529
- 530
- 531
- 532
- 533
- 534
- 535

536

537

538

539

540

541

542

543

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

Set of CpG sites	N	3 rd percentile	Mean	Median	97 th 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

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567 **Table 2.** Kolmogorov-Smirnov test assessing the uniformity of the p-value distributions from
568 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.

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

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.

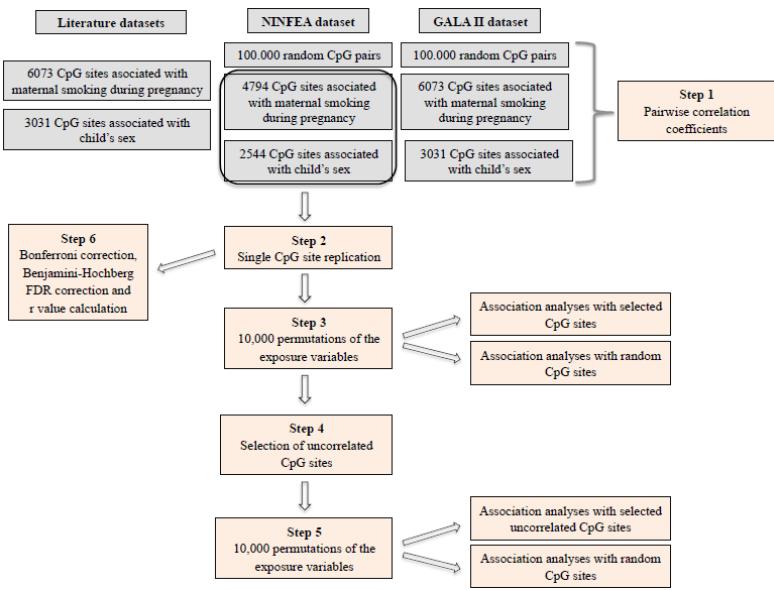
Smoking-associated CpG sites	Discovery study two-sided p-value	Replication study two-sided p-value	FWER	BH FDR p-value	FDR r-value^a	Modified r-value^b
cg12793610	4.42e-05	9.12e-06	0.04	0.04	0.91	1.00
Sex-associated CpG sites						
cg23092538	7.43e-05	1.69e-05	0.04	0.02	0.18	0.74
cg03168896	1.86e-08	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

591

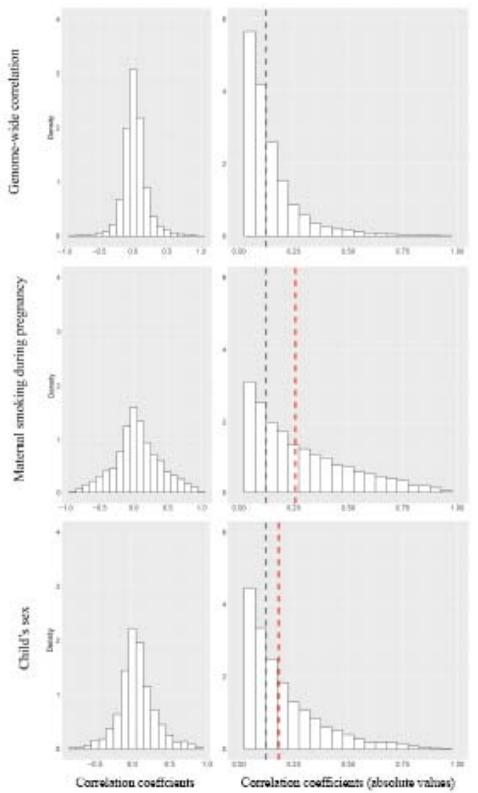
Figure 1. The main steps of the analysis



592

593

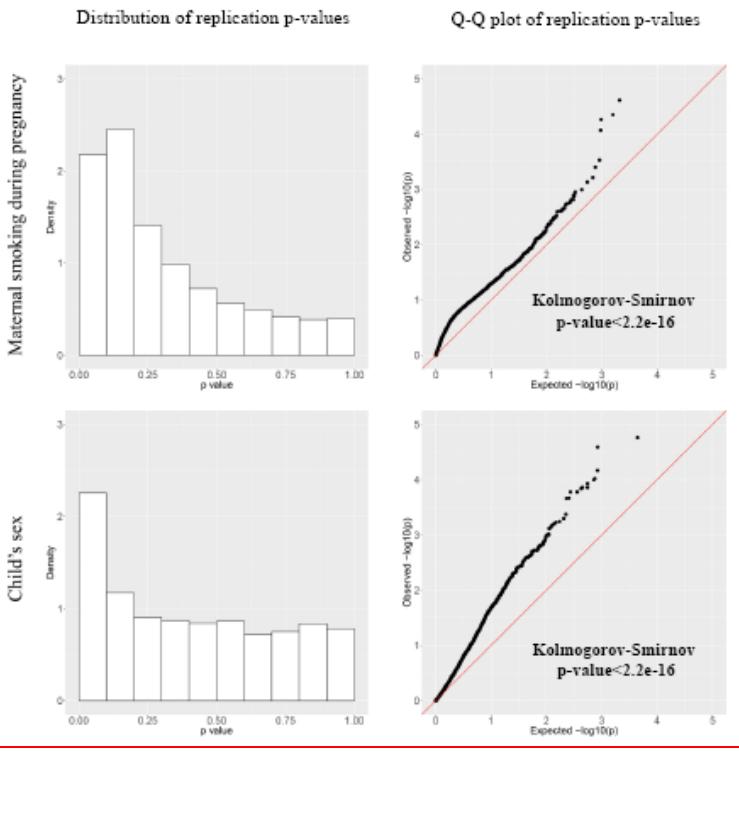
594 **Figure 2.** Distribution of correlation coefficients (left side) and their absolute values (right side) for
595 genome-wide CpG sites, 4794 CpG sites associated with maternal smoking during pregnancy, and
596 2544 CpG sites associated with child's sex. Vertical gray line indicates genome-wide mean
597 correlation coefficient (absolute values). Vertical red lines indicate mean correlations coefficients
598 (absolute values) for each set of the selected CpG sites.



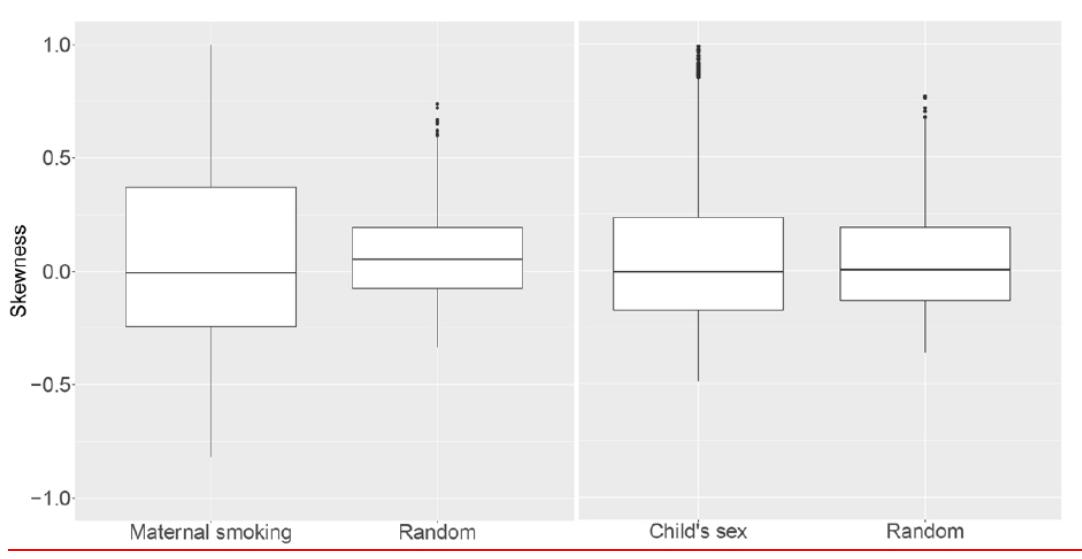
Note: Vertical gray line indicates genome-wide mean correlation coefficient.
Vertical red line indicate mean correlation coefficients of each set of selected CpG sites.

599

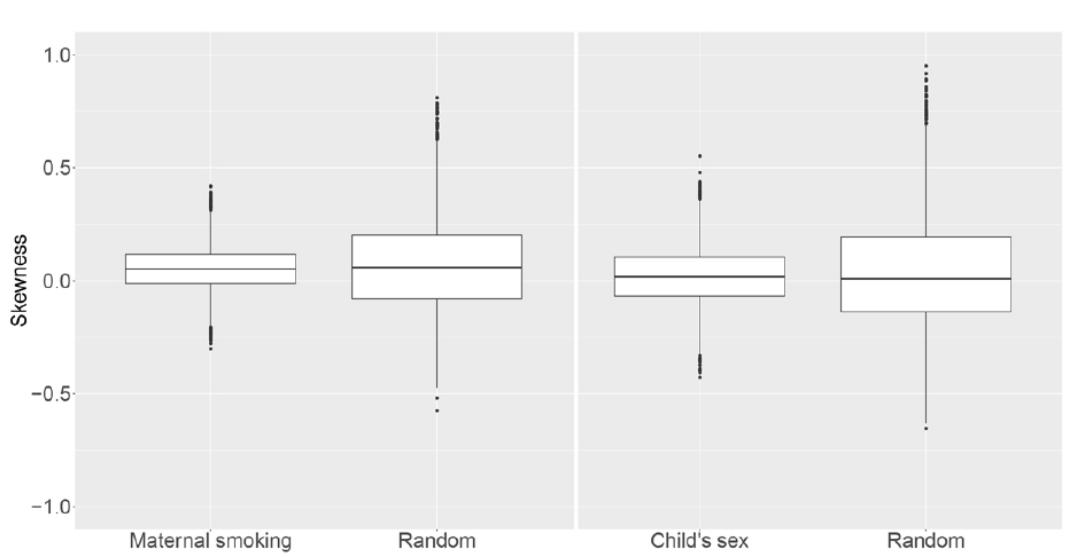
600 **Figure 3.** Replication in the NINFEA cohort: Distribution of replication p-values and Q-Q plots of
601 observed versus expected p-values for the associations between methylation levels at smoking-
602 related (N=4794) and sex-related (N=2544) CpG sites and maternal smoking during pregnancy, and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



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



612 **Figure 5.** Skewness of p-value distributions from the analyses of the association between smoking-
 613 related “low-correlated” ($N=256$) and sex-related “low-correlated” ($N=129$) pre-selected CpG sites
 614 and permutations of maternal smoking/child’s sex from 10,000 replications. “Random” indicates
 615 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	2
Technical note	2
Results	6
Table S1.....	6
Figure S1	7
Figure S2	8
Figure S3	9

Methods

Technical note

We assume that we are testing m independent null hypotheses $H_{01}, H_{02}, \dots, H_{0m}$. 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:

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

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 :

$$\begin{aligned}\mathcal{H}_j &= \{H_j = (h_{1j}, h_{2j}); h_{ij} \in \{0,1\}\} \\ &= \{(0,0), (0,1), (1,0), (1,1)\}.\end{aligned}$$

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

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 :

$$\begin{aligned}\mathcal{H}_j &= \{H_j = (h_{1j}, h_{2j}); h_{ij} \in \{-1, 0, 1\}\} \\ &= \{(-1, -1), (-1, 0), (-1, 1), (0, -1), (0, 0), (0, 1), (1, -1), (1, 0), (1, 1)\}.\end{aligned}$$

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 $\leq q$.

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	3 rd percentile	Mean	Median	97 th 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

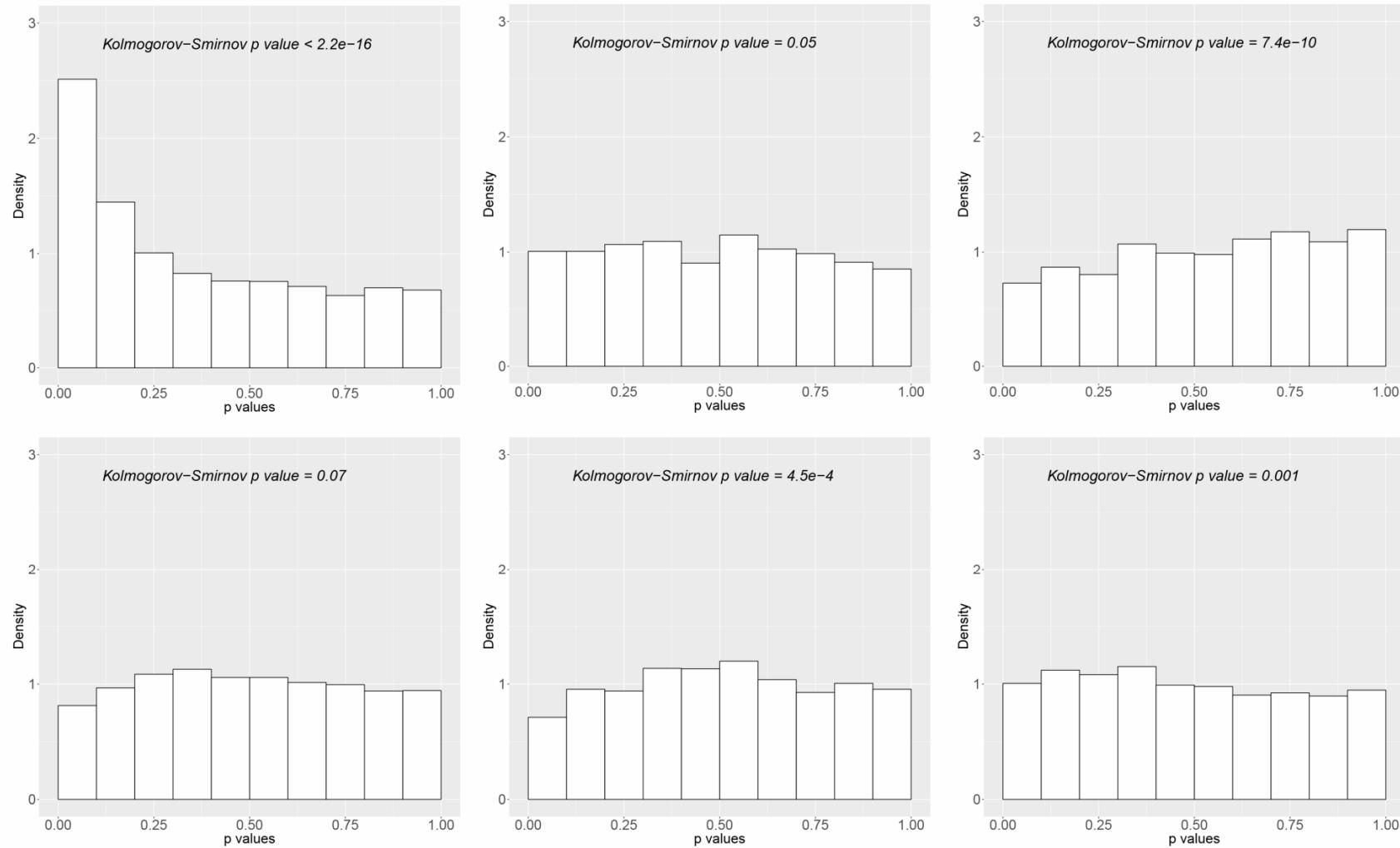


Figure S1. Histograms of p-value distributions from random permutations and Kolmogorov-Smirnov p value assessing whether the observed p-value distributions come from a hypothesized uniform distribution

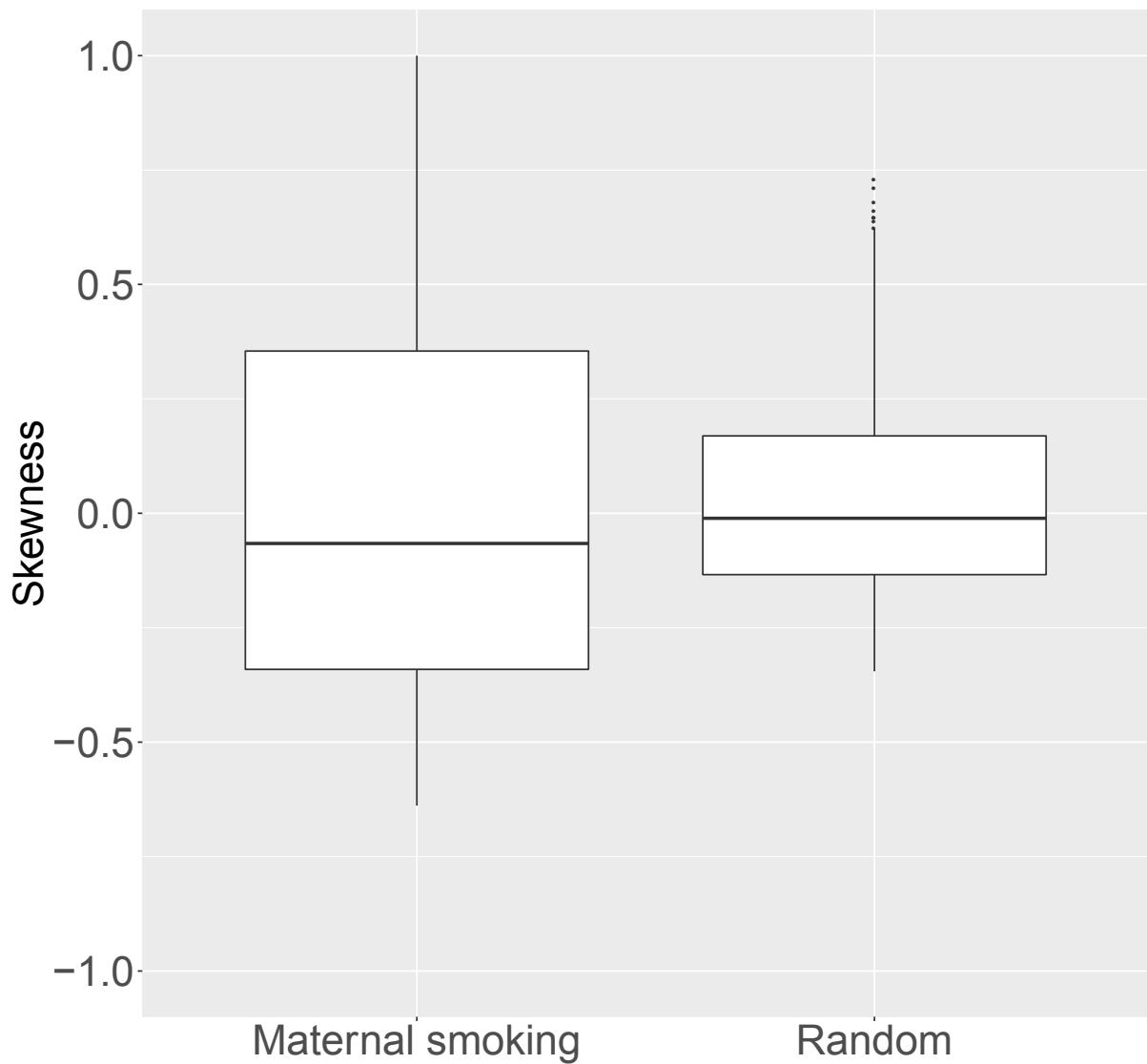


Figure S2. Skewness of p-value distributions from the analyses of the association between 4794 CpG sites associated with maternal smoking and 10,000 permutations of an imaginary exposure for 138 subjects from the NINFEA cohort. “*Random*” indicates random permutations of both CpG sites and exposure under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

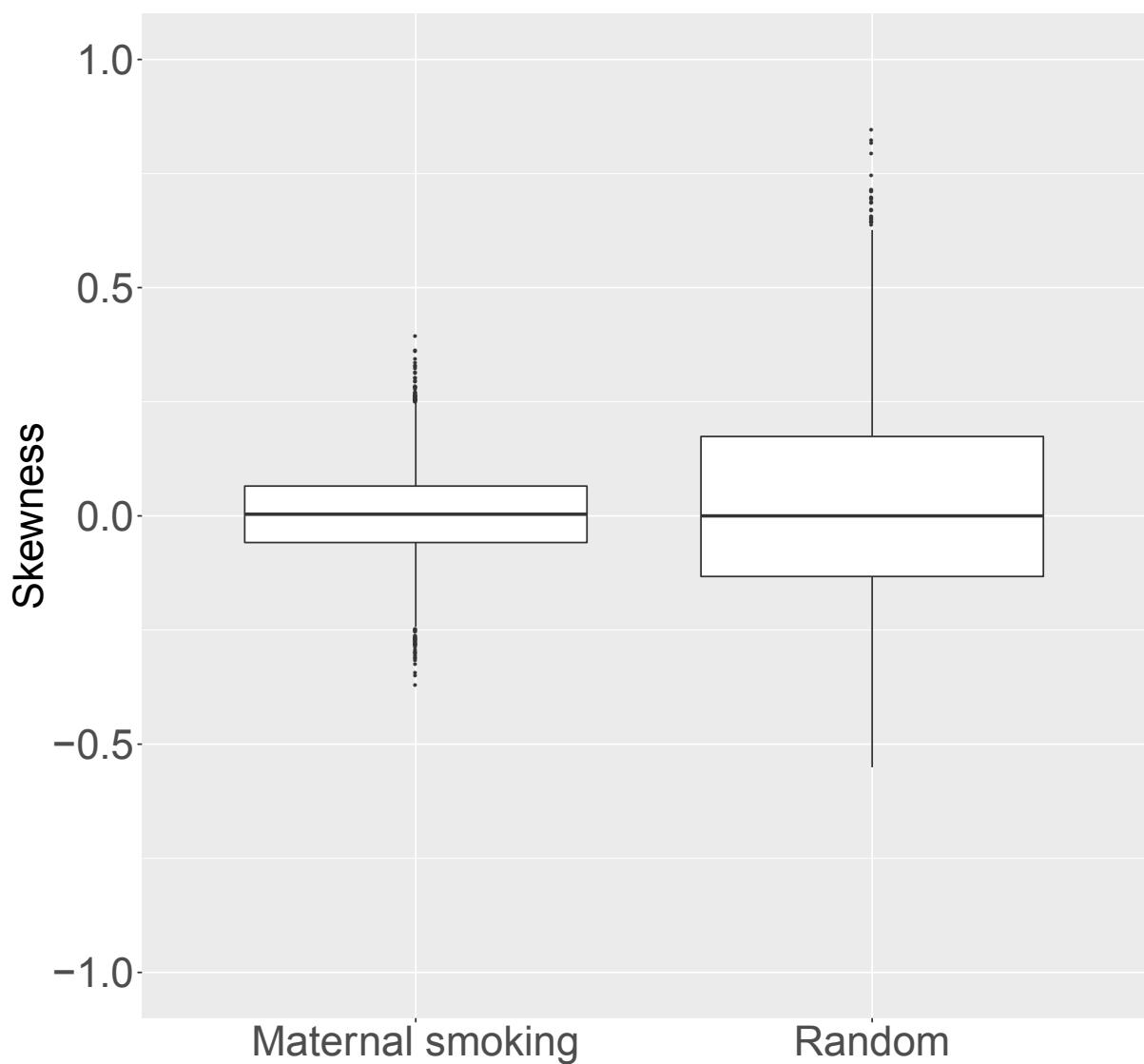


Figure S3. Skewness of p-value distributions from the analyses of the association between 256 low-correlated CpG sites associated with maternal smoking and 10,000 random permutations of an imaginary exposure for 138 subjects from the NINFEA cohort. “*Random*” indicates random permutations of both CpG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