

## Supplementary Information

### Unidirectional and Bidirectional Causation between Smoking and Blood DNA Methylation: Evidence from Twin-based Mendelian Randomisation

Madhurbain Singh<sup>1,2,3\*</sup>, Conor V. Dolan<sup>3,4,11</sup>, Dana M. Lapato<sup>1,2</sup>, Jouke-Jan Hottenga<sup>3,4</sup>, René Pool<sup>3,4</sup>, Brad Verhulst<sup>5</sup>, Dorret I. Boomsma<sup>3,4,12</sup>, Charles E. Breeze<sup>6,7</sup>, Eco J. C. de Geus<sup>3,4</sup>, Gibran Hemani<sup>8</sup>, Josine L. Min<sup>8</sup>, Roseann E. Peterson<sup>9,10,1</sup>, Hermine H. M. Maes<sup>1,2</sup>, Jenny van Dongen<sup>3,4,11\*</sup> and Michael C. Neale<sup>1,2,3,11\*</sup>

1. Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA
2. Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, USA
3. Department of Biological Psychology, Vrije Universiteit (VU) Amsterdam, Amsterdam, The Netherlands
4. Amsterdam Public Health Research Institute, Amsterdam, The Netherlands
5. Department of Psychiatry and Behavioral Sciences, Texas A&M University, College Station, TX, USA
6. Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department Health and Human Services, Bethesda, MD, USA
7. UCL Cancer Institute, University College London, London, UK.
8. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK
9. Department of Psychiatry and Behavioral Sciences, SUNY Downstate Health Sciences University, Brooklyn, NY, USA
10. Institute for Genomics in Health, SUNY Downstate Health Sciences University, Brooklyn, NY, USA
11. These authors jointly supervised this work.
12. Current address: Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Vrije Universiteit (VU) Amsterdam, Amsterdam, The Netherlands

\*Corresponding authors:

Madhurbain Singh. Email: [singhm18@vcu.edu](mailto:singhm18@vcu.edu). Address: Virginia Institute for Psychiatric and Behavioral Genetics, 800 E. Leigh St., Suite 100, Richmond, VA 23298, USA

Jenny van Dongen. Email: [j.van.dongen@vu.nl](mailto:j.van.dongen@vu.nl). Address: Department of Biological Psychology, Vrije Universiteit Amsterdam, van der Boeorchorststraat 7, 1081 BT Amsterdam, The Netherlands

Michael C. Neale. Email: [michael.neale@vcuhealth.org](mailto:michael.neale@vcuhealth.org). Address: Virginia Institute for Psychiatric and Behavioral Genetics, 800 E. Leigh St., Suite 100, Richmond, VA 23298, USA

## 41 Table of Contents

|    |  |           |
|----|--|-----------|
| 42 | <b><i>Supplementary Methods</i></b> .....  | <b>4</b>  |
| 43 | <b>Genotypic Data, Principal Components Analysis, and Ancestry Outlier Pruning</b> ..... | <b>4</b>  |
| 44 | <b>Peripheral Blood DNA Methylation Data</b> .....                                       | <b>4</b>  |
| 45 | <b>Smoking Assessment at Blood Sampling</b> .....  | <b>5</b>  |
| 46 | <b>mQTL Allelic Score</b> .....  | <b>5</b>  |
| 47 | <b>Polygenic Risk Score of Smoking</b> .....   | <b>7</b>  |
| 48 | <b>Univariate Twin Models</b> .....  | <b>7</b>  |
| 49 | <b>MR-DoC Models</b> .....   | <b>8</b>  |
| 50 | <b>Functional Enrichment Analyses</b> .....  | <b>9</b>  |
| 51 | <b><i>Supplementary Figures</i></b> .....  | <b>11</b> |
| 52 | <b>Figure S1</b> .....   | <b>11</b> |
| 53 | <b>Figure S2</b> .....   | <b>13</b> |
| 54 | <b>Figure S3</b> .....   | <b>13</b> |
| 55 | <b>Figure S4</b> .....   | <b>14</b> |
| 56 | <b>Figure S5</b> .....   | <b>14</b> |
| 57 | <b>Figure S6</b> .....   | <b>15</b> |
| 58 | <b>Figure S6</b> .....   | <b>17</b> |
| 59 | <b>Figure S8</b> .....   | <b>18</b> |
| 60 | <b>Figure S9</b> .....   | <b>19</b> |
| 61 | <b>Figure S10</b> .....  | <b>20</b> |
| 62 | <b>Figure S11</b> .....  | <b>21</b> |
| 63 | <b>Figure S12</b> .....  | <b>22</b> |
| 64 | <b>Figure S13</b> .....  | <b>24</b> |
| 65 | <b>Figure S14</b> .....  | <b>25</b> |

|    |                                |           |
|----|--------------------------------|-----------|
| 66 | <b>Figure S15</b> .....        | <b>26</b> |
| 67 | <b>Figure S16</b> .....        | <b>27</b> |
| 68 | <b>Figure S17</b> .....        | <b>28</b> |
| 69 | <b>Figure S18</b> .....        | <b>28</b> |
| 70 | <b>Figure S19</b> .....        | <b>29</b> |
| 71 | <b>Figure S20</b> .....        | <b>30</b> |
| 72 | <b>Figure S21</b> .....        | <b>31</b> |
| 73 | <b>Figure S22</b> .....        | <b>32</b> |
| 74 | <b>Figure S23</b> .....        | <b>33</b> |
| 75 | <b>Figure S24</b> .....        | <b>34</b> |
| 76 | <b>Figure S25</b> .....        | <b>35</b> |
| 77 | <b>Figure S26</b> .....        | <b>36</b> |
| 78 | <b>Figure S27</b> .....        | <b>37</b> |
| 79 | <b>Figure S28</b> .....        | <b>38</b> |
| 80 | <b>Figure S29</b> .....        | <b>39</b> |
| 81 | <b>Figure S30</b> .....        | <b>40</b> |
| 82 | <b>Figure S31</b> .....        | <b>41</b> |
| 83 | <b>Figure S32</b> .....        | <b>42</b> |
| 84 | <b>Figure S33</b> .....        | <b>43</b> |
| 85 | <b>Figure S34</b> .....        | <b>44</b> |
| 86 | <b>Figure S35</b> .....        | <b>45</b> |
| 87 | <b>Figure S36</b> .....        | <b>46</b> |
| 88 | <b><i>References</i></b> ..... | <b>47</b> |
| 89 |                                |           |
| 90 |                                |           |

## 91 **Supplementary Methods**

92 In this study, we analyzed data from the Netherlands Twin Register (NTR) [1] to examine the  
93 causal influences between smoking status and blood DNA methylation (DNAm) using MR-DoC  
94 models [2,3]. In the current analyses, we included data from European-ancestry adult twins with  
95 both genotypic and DNAm data, comprising 2,577 individuals (67% female).  
96

### 97 **Genotypic Data, Principal Components Analysis, and Ancestry Outlier Pruning**

98 The DNA samples included in the current study were genotyped on 3 SNP (single nucleotide  
99 polymorphism) microarray platforms: Affymetrix 6.0 (N= 2,399), Affymetrix Axiom (N= 83),  
100 and Illumina GSA NTR array (N= 95). Genotype calling was done following the manufacturer's  
101 protocols. Sample and variant quality control (QC), imputation, genetic principal component  
102 analysis (PCA), and ancestry assignment have been previously described [4]. Briefly, after QC  
103 and harmonizing variants across the three platforms, the data were aligned to the positive strand  
104 of *Genome Reference Consortium Human Build 37* (GRCh37) and then imputed against the  
105 European (EUR) super-population of the 1000 Genomes Project Phase-3 (KGP3) [5], the  
106 Haplotype Reference Consortium (HRC) [6] 1.1 (Ega version), and the Genome of the  
107 Netherlands Consortium (GoNL) [7] reference panels. Using SmartPCA in EIGENSTRAT [8],  
108 the first 20 PCs for the genotypic data were calculated in the KGP3 data, and the NTR samples  
109 were then projected onto the PC space based on the SNP weights. Samples identified as outliers  
110 in the PC space were then excluded.  
111

### 112 **Peripheral Blood DNA Methylation Data**

113 Epigenome-wide DNAm in peripheral whole blood was measured with the Infinium  
114 HumanMethylation450 BeadChip Kit (i.e., the Illumina 450k microarray), following the  
115 manufacturer's protocol [9]. QC and normalization of the DNAm data were performed using a  
116 custom pipeline developed by the BIOS (Biobank-based Integrative Omics Study) Consortium,  
117 as previously described [10]. Briefly, sample QC was done using MethylAid [11], followed by  
118 probe QC with DNAmArray [12]. The latter removed the probes with a raw signal intensity of  
119 zero, bead number <3, or a detection p-value >0.01, as well as the ambiguously mapped probes.  
120 Next, samples and probes with >5% missingness were removed. The resulting DNAm data were  
121 normalized using the Functional normalization algorithm [13] implemented in DNAmArray [12],  
122 with the first four PCs (with eigenvalue >1) derived from control probes. Finally, the probes  
123 containing a SNP within the CpG site (at C or G nucleotide) were removed regardless of the  
124 minor allele frequency. These SNPs were previously identified using DNA sequencing data from  
125 the Dutch population in GoNL [7]. For the current analyses, only autosomal probes were  
126 included, yielding 411,169 CpG sites that passed all QC metrics.  
127

128 **Smoking Assessment at Blood Sampling**

129 Self-reported cigarette smoking status was recorded through an interview during the home visit  
 130 for blood sample collection in 2004-2008 and 2010-2011. Participants were asked, “Do you  
 131 smoke?” with one of three possible answers: “No, I never smoked” (N = 1,492), “No, but I did in  
 132 the past” (N = 549), and “Yes” (N = 528). Those endorsing current smoking were asked how  
 133 many years they had been smoking and how many cigarettes or rolling tobacco they smoked per  
 134 day. Those endorsing former smoking were asked how many years ago they quit smoking, how  
 135 many years they had smoked before quitting, and the maximum number of cigarettes or rolling  
 136 tobacco they used to smoke per day. The original wording in Dutch is shown below.  
 137

|   |                                |  |
|---|--------------------------------|--|
| Rookt u?  | 1. Ja                          | 1a: hoelang rookt u al? .....jaar                                    |
|   |                                | 1b: hoeveel sigaretten/ shagjes per dag?<br>.....sigaretten/ shagjes |
|   | 2. Nee, wel in het<br>verleden | 2a: hoelang is dat geleden?.....jaar                                 |
|   |                                | 2b: hoeveel jaren heeft u gerookt?.....jaar                          |
| 2c: hoeveel rookte u per dag (max)?<br>.....sigaretten/ shagjes |                                |  |
| 3. Nooit  |                                |  |
| Gerookt binnen<br>laatste uur voor<br>bloedafname?              | 1. Ja<br>2. Nee<br>3. Nvt      |  |

138  
 139 The responses were checked for consistency with the information from the NTR longitudinal  
 140 surveys filled out closest to blood sampling. As previously described [14], potential  
 141 misclassification of smoking status through self-reports was evaluated based on plasma cotinine  
 142 levels (a metabolite of nicotine and a biomarker of smoking exposure), measured in a subset of  
 143 the sample. Of the 591 individuals with self-reported never smoking and measured plasma  
 144 cotinine, only five (0.8%) had cotinine levels indicative of smoking ( $\geq 15$  ng/ml), thus suggesting  
 145 low misclassification of smoking status. The number of individuals endorsing current or former  
 146 smoking was too small to evaluate a dose-response relationship of the causal effects in MR-DoC  
 147 models restricted to currently or formerly smoking individuals. Likewise, the sample with former  
 148 smoking was too small to examine the effect of “time since quitting smoking” on DNAm.  
 149

150  
 151 **mQTL Allelic Score**

152 We identified 12,940 smoking-associated CpGs with *cis*-mQTL summary statistics available  
 153 from GoDMC [15] (excluding NTR), using GoDMC’s definition of “*cis*” interval (within 1Mb of  
 154 the CpG). In GoDMC, the contributing cohorts performed genome-wide mQTL analyses, testing  
 155 the associations of ~480,000 CpG sites with ~12 million SNPs. However, before the meta-  
 156 analysis, the cohort-level results were filtered to retain the SNP-CpG pairs with  $p < 1 \times 10^{-5}$

157 within the cohort. Thus, since the summary statistics were already partly thresholded, we  
 158 computed the mQTL allelic scores by applying clumping and thresholding in *PLINK1.9* [16],  
 159 using summary statistics from the Genetics of DNA Methylation Consortium (GoDMC;  
 160 excluding NTR) [15]. Linkage disequilibrium (LD)-based clumping was performed using --  
 161 clump-p1 1 --clump-kb 250, with two levels of LD  $r^2$  (0.5 and 0.1) specified for --  
 162 clump-r2, thus yielding two sets of LD-clumped *cis*-SNPs. Using either set of SNPs, we  
 163 computed the allelic score with --score at a threshold of 0.05 (applied with --q-score-  
 164 range). If none of the SNPs had  $p < 0.05$ , no threshold was applied for score calculation. An  
 165 additional allelic score was calculated using the top *cis*-mQTL (with the minimum association  $p$ -  
 166 value) for each CpG. Thus, for every CpG, three scores were calculated (two LD-clumped  
 167 mQTL allelic scores, plus the top-mQTL), though these scores were not necessarily distinct; for  
 168 example, if a CpG had only one *cis*-SNP, all three criteria yielded the same score. Likewise, for  
 169 some CpGs, the two LD-clumping cut-offs resulted in the same set of SNPs and, hence, identical  
 170 mQTL allelic scores.

171  
 172 To assess the strength of an mQTL allelic score, we first estimated its incremental  $R^2$  by fitting  
 173 generalized estimating equations (GEE), controlling for the standard EWAS covariates (as  
 174 above), genotyping platform, and the first ten genetic PCs. For each CpG, the mQTL allelic  
 175 score with the highest incremental  $R^2$  was retained for further filtering based on F-statistic.  
 176 For each CpG, the effective GEE sample size ( $N_{Eff}$ ) was computed using the following  
 177 formulae:

$$178 \quad N_{Eff}^{MZ} = \frac{2 * N_{MZ}}{1 + r_{MZ}}$$

$$179 \quad N_{Eff}^{DZ} = \frac{2 * N_{DZ}}{1 + r_{DZ}}$$

$$180 \quad N_{Eff} = N_{Eff}^{MZ} + N_{Eff}^{DZ} + N_{Ind}$$

181 where,  $N_{Eff}^{MZ}$  and  $N_{Eff}^{DZ}$  are the estimated effective sample sizes of MZ and DZ twins,  $N_{MZ}$  and  
 182  $N_{DZ}$  are the numbers of complete MZ and DZ twin pairs, while  $r_{MZ}$  and  $r_{DZ}$  are the twin  
 183 phenotypic (DNAm) correlations in MZ and DZ twin pairs, respectively.  $N_{Ind}$  is the number of  
 184 individuals without the co-twin.

185 The estimated effective sample size was then used to transform the incremental  $R^2$  value into an  
 186 F-statistic as:

$$187 \quad F = \frac{R^2}{1 - R^2} \times \frac{N_{Eff} - K}{K - 1}$$

188  
 189  
 190  
 191  
 192  
 193

194 where  $K = 2$ , given two parameter estimates: the intercept and the regression coefficient of the  
195 mQTL allelic score.

196  
197

### 198 **Polygenic Risk Score of Smoking**

199 The PRS of smoking was based on the European-ancestry summary statistics from the genome-  
200 wide association study (GWAS) of smoking initiation (lifetime regular smoking) by GSCAN  
201 (GWAS & Sequencing Consortium of Alcohol and Nicotine use)[17], excluding the NTR from  
202 the meta-analysis.

203 As described in a previous study using the same PRS in the NTR[4], the post-imputation SNPs  
204 from the merged best-guess three-platform data were QCed to satisfy the following criteria:  
205 MAF  $>0.01$ , HWE  $p >0.00001$ , Mendel error rate  $< 1\%$ , and genotype call rate over 98%.

206 Furthermore, the imputation info for the three platforms needed to be above 0.10, and the allele  
207 frequency between platforms after imputation could not differ more than 2%, leaving a total of  
208 7,551,860 post-QC SNPs for analysis. The PRS was calculated using *LDpred* v0.9[18], with  
209 HRC+GoNL as the LD (linkage disequilibrium) reference panel. For estimating the target LD  
210 structure, we used a subset of unrelated individuals and a set of well-imputed variants in the  
211 NTR. The parameter `ld_radius` was set by dividing the number of variants in common (from  
212 the output of the coordination step) by 12000. For the coordination step, the median sample size  
213 was used as the input value for `N`. For the *LDpred* step, we applied the following thresholds for  
214 the fraction of variants with non-zero effects (in addition to the default infinitesimal model): --  
215 `PS=0.5,0.3,0.2,0.1,0.05,0.01`.

216 To determine the *LDpred* threshold that yielded the PRS with the highest predictive power for  
217 the variables of interest (current vs. never and former vs. never smoking), we fitted logistic  
218 regression models in R (v4.3.2) to estimate incremental  $R^2$  on a liability scale. We first fitted a  
219 null logistic regression model using the `glm()` function with  
220 `family=binomial(link='logit')` and a standard set of covariates comprising age  
221 (linear and quadratic), sex, SNP microarray platform (dummy variables), and the first ten genetic  
222 PCs (without including the PRS). Then, we fitted a full model with the PRS as an additional  
223 independent variable. We estimated the liability-scale  $R^2$  in both models and then the difference  
224 in the two  $R^2$  estimates as the variance in the outcome variable explained by the PRS (controlling  
225 for the covariates). For both outcome variables (current and former smoking), the PRS with the  
226 highest incremental  $R^2$  was based on a threshold of 0.1 and thus retained for further analyses.  
227 The PRS was residualized for the SNP microarray platform and the first ten genetic PCs using  
228 linear regression models. The residuals were then standardized to have a mean of zero and an  
229 S.D. of one before using it as an IV in the MR-DoC models.

230

### 231 **Univariate Twin Models**

232 Before fitting the MR-DoC models, we examined univariate ACE twin models of smoking status  
233 to estimate the additive genetic (A), shared environmental (C), and unique environmental (E)

234 variance components of the latent liability scale, with age and sex as covariates. Maximum-  
 235 likelihood tetrachoric correlation estimates for current versus never smoking were:  $r_{MZ} = 0.925$   
 236 ( $S.E. = 0.021$ ) in MZ pairs, and  $r_{DZ} = 0.533$  ( $S.E. = 0.083$ ) in DZ pairs. Likewise, former  
 237 versus never smoking had  $r_{MZ} = 0.822$  ( $S.E. = 0.038$ ) and  $r_{DZ} = 0.474$  ( $S.E. = 0.096$ ). Based  
 238 on likelihood-ratio tests (LRT), an AE twin model was the most parsimonious model for both  
 239 current versus never (AE versus ACE LRT  $p = 0.417$ ) and former versus never smoking (AE  
 240 versus ACE LRT  $p = 0.530$ ) (**Supplementary Table S31**). The estimated variance components  
 241 of current versus never smoking liability were  $A = 0.927$  (maximum-likelihood 95% confidence  
 242 interval: 0.879, 0.959) and  $E = 0.073$  (0.041, 0.121). The corresponding estimates of former  
 243 versus never smoking were  $A = 0.827$  (0.745, 0.888) and  $E = 0.173$  (0.112, 0.255).  
 244 Prior twin analyses of DNAm at CpG sites in NTR [10] showed that, of the 411,169 autosomal  
 245 post-QC CpG sites, the AE twin model was the best fitting model at all but 426 sites, with  
 246 significant (after multiple-testing correction of LRT p-values) C variance at 185 sites and  
 247 significant non-additive genetic (D) variance at 241 sites. Of the smoking-associated CpGs [19],  
 248 only two CpGs had significant estimates of C, while only seven CpGs had significant estimates  
 249 of D. Thus, in the MR-DoC models, we specified an AE variance decomposition of DNAm at all  
 250 smoking-associated CpGs. Note that, in the results presented in the main text, none of the CpG  
 251 sites with consistent, nominally significant estimates of causal effects in either direction (525  
 252 sites with *current smoking*  $\rightarrow$  DNAm; 64 sites with DNAm  $\rightarrow$  *current smoking*) have significant  
 253 C or D estimates per the previous univariate twin analyses [10]. Moreover, since smoking status  
 254 liability also has an AE variance decomposition, including a C or D variance component of  
 255 DNAm in the model would not change the possible sources of covariance between smoking  
 256 status and DNAm in the model.

257

## 258 MR-DoC Models

259 We used the *OpenMx* (version 2.21.8) [20] package in R (version 4.3.2) to fit the MR-DoC  
 260 models using the code provided in the original publications [2,3]. Binary smoking status was  
 261 examined under the liability threshold model [21], assuming a latent liability distribution with its  
 262 mean fixed at zero and variance fixed at one, while the threshold was freely estimated.

263 In each MR-DoC model, the residual variance of smoking status liability is decomposed into  $a_S^2$   
 264 (A) and  $e_S^2$  (E), while that of DNAm is decomposed into  $a_D^2$  (A) and  $e_D^2$  (E). The correlation  
 265 between the latent A factors of smoking and DNAm ( $r_A$ ) represents the confounding due to  
 266 additive genetic factors. The correlation between the latent E factors ( $r_E$ ) represents the  
 267 confounding due to unique environmental factors. Across all models, the causal path from  
 268 smoking to DNAm is labeled  $g_1$ , while that from DNAm to smoking is labeled  $g_2$ . The  
 269 residualized PRS and mQTL allelic scores are regressed on respective latent factors, representing  
 270 the underlying “true” standardized scores with mean fixed at zero and variance fixed at one. The  
 271 coefficient of the path from the latent score to the observed score estimates the standard  
 272 deviation of the observed score ( $SD_{PRS}$  and  $SD_{mQTL}$ , respectively).

273 We fitted five sets of MR-DoC models with current versus never smoking and similar sets with  
 274 former versus never smoking (**Figure 1**): (1) *Smoking*  $\rightarrow$  DNAm MR-DoC1 with horizontal



275 pleiotropy, (2) *Smoking* → *DNAm* MR-DoC1 with unique environmental confounding, (3)  
276 *DNAm* → *Smoking* MR-DoC1 with horizontal pleiotropy, (4) *DNAm* → *Smoking* MR-DoC1  
277 with unique environmental confounding, and (5) bidirectional MR-DoC2. Each model included  
278 age and sex as covariates of smoking status. Thus, for each CpG site included in the analyses,  
279 three causal estimates were obtained in either direction (*Smoking* → *DNAm*, or *DNAm* →  
280 *Smoking*) from (1) MR-DoC1 with horizontal pleiotropy, (2) MR-DoC1 with unique  
281 environmental confounding, and (3) MR-DoC2. For each set of causal estimates across CpG  
282 sites, we calculated the Bayesian inflation factor ( $\lambda$ ) using the R package *bacon* [22], made QQ  
283 plots using the R package *GWASTools* [23], and then applied Benjamini-Hochberg FDR  
284 correction [24] to the p-values using the R package *qvalue* [25]. For Bonferroni multiple-testing  
285 correction, the significance level was defined as  $\alpha = 0.05/16940 = 2.95 \times 10^{-6}$  for *Current*  
286 *Smoking* → *DNAm* MR-DoC1 models and  $\alpha = 0.05/11124 = 4.49 \times 10^{-6}$  for *DNAm* →  
287 *Current Smoking* MR-DoC1 and bidirectional current-smoking MR-DoC2 models.  
288

### 289 **Functional Enrichment Analyses**

290 We used Metascape [26] (v3.5.20240101; <https://metascape.org/gp/index.html#/main/step1>, with  
291 the default settings for “Express” analyses) to perform gene-set annotation and functional  
292 enrichment analyses of the CpGs with potential causal effects in either direction. The input list of  
293 gene IDs was selected based on proximity to the CpGs with consistent and nominally significant  
294 ( $p < 0.05$ ) estimates in all three models; i.e., 64 CpGs with potential *DNAm* → *Current Smoking*  
295 effects (“Nearest Gene” in **Supplementary Table S3**) and 525 CpGs with potential *Current*  
296 *Smoking* → *DNAm* effects (“Nearest Gene” in **Supplementary Table S1**). None of the sites with  
297 potential *DNAm* → *Current Smoking* effects are located in the MHC region. For *Current*  
298 *Smoking* → *DNAm* effects, 21 additional sites in the MHC region showed consistent, nominally  
299 significant estimates. There was no significant relationship between a CpG site having consistent  
300 causal estimates and its being located in the MHC region (Fisher’s exact test p-value = 0.5455).  
301 However, out of an abundance of caution, the sites located in this region were not included in the  
302 enrichment analyses to avoid sites with potentially unreliable results due to its complex LD  
303 structure.  
304

305 As described in the Metascape manuscript [26], the program performed integrated enrichment  
306 analyses against multiple reference ontology knowledgebases, including GO processes [27],  
307 KEGG pathways [28], canonical pathways [29], and Reactome gene sets [30]. The significant  
308 terms with a hypergeometric p-value  $< 0.01$  and  $> 1.5$ -fold enrichment were clustered into a  
309 hierarchical tree based on Kappa-statistical similarities among their gene memberships. The tree  
310 was then cast into clusters based on a threshold of 0.3 kappa score to obtain enriched, non-  
311 redundant ontology terms.  
312

### 313 **eFORGE (experimentally derived Functional element Overlap analysis of ReGions from** 314 **EWAS)**

315 We performed *eFORGE 2.0* [31–33] analyses of the selected CpG probe IDs with consistent and  
316 nominally significant ( $p < 0.05$ ) estimates in either direction (from **Supplementary Tables S1,**  
317 **S3**). Using the web-based tool (<https://eforge.altiusinstitute.org/>), we examined the overlap  
318 between the implicated CpGs and multiple comprehensive reference sets of genomic and  
319 epigenomic features that regulate gene expression in different tissues and cell types. The  
320 platform was set as “Illumina 450k array”, with default analysis options: proximity = 1kb  
321 window, background repetitions = 1000, and significance thresholds of FDR  $< 0.01$  (strict) and  
322 FDR  $< 0.05$  (marginal). Three sets of analyses were performed for each list of probe IDs,  
323 selecting the reference data from “Consolidated Roadmap Epigenomics - Chromatin - All 15-  
324 state marks”, “Consolidated Roadmap Epigenomics - DHS”, and “Consolidated Roadmap  
325 Epigenomics - All H3 marks”.

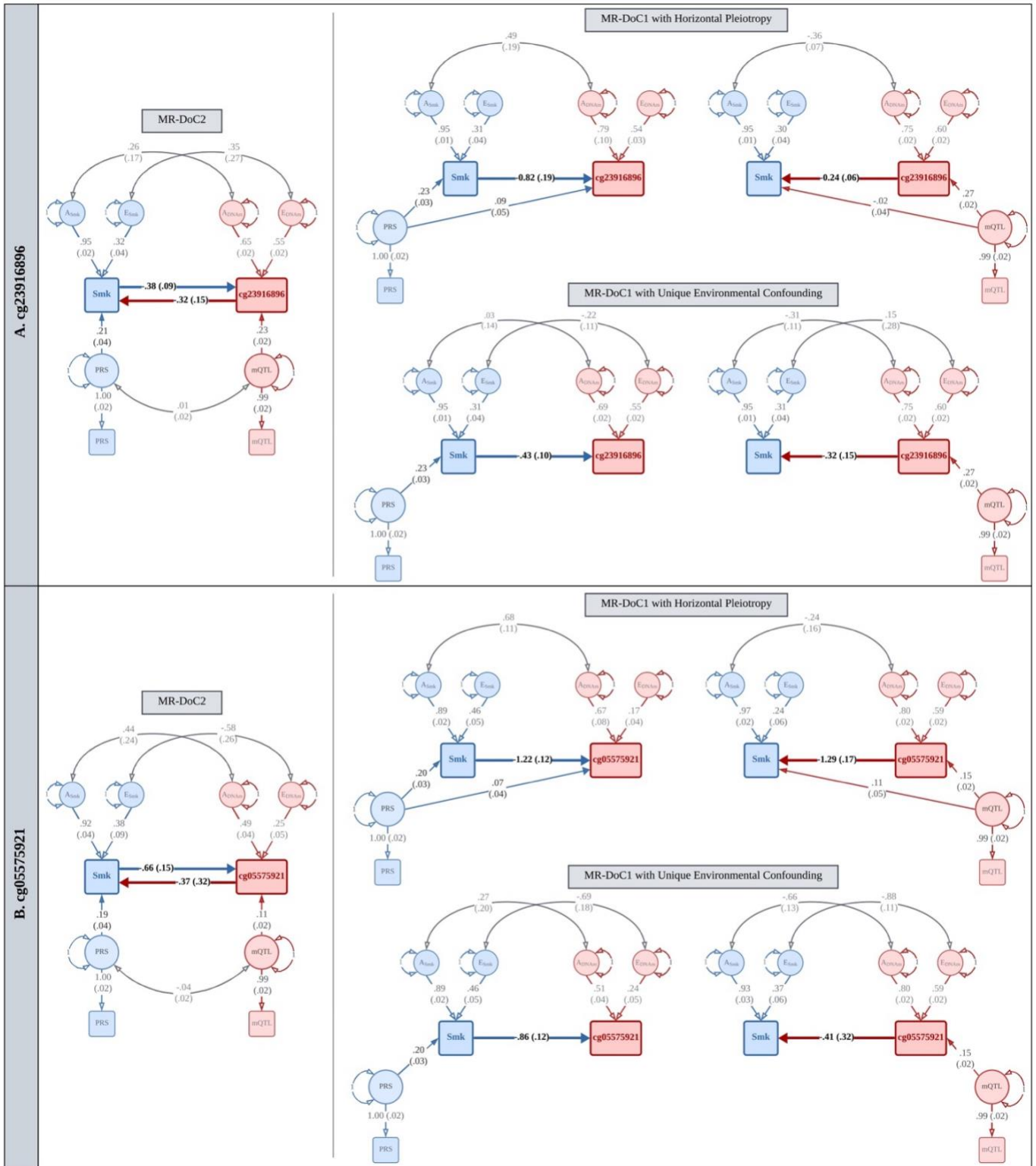
326 The eFORGE results include the specific probe IDs overlapping between the input set and the  
327 reference sample. We performed iterative follow-up analyses for the CpGs with potential *DNAm*  
328  $\rightarrow$  *Current Smoking* effects, based on the overlapping probe IDs to examine the specificity of  
329 significant (FDR  $< 0.01$ ) enrichment in tissues of interest. Analyses restricted to the 21 CpGs  
330 overlapping with enhancers in the fetal brain (**Supplementary Figure S18, Table S12**) showed  
331 significant enrichment only for enhancers in the fetal brain samples, suggesting high specificity  
332 (**Supplementary Figure S21**). The histone mark analyses also showed enrichment in the fetal  
333 brain (though not specific to the brain), wherein all 21 CpGs overlapped with H3K4me1, while a  
334 subset of 17 CpGs overlapped with H3K4me3 (**Supplementary Figure S22**). Finally, we  
335 performed analyses restricted to these 17 CpGs.

336 We performed similar follow-up analyses with probe IDs showing overlap with enhancers in the  
337 lung (potentially etiologically relevant tissue) and the primary B-cells in cord blood (the tissue  
338 type with the most significant enrichment) (from **Supplementary Figure S18, Table S12**).

339

340 Enrichment in blood cell types may be influenced by residual cell-composition effects in whole  
341 blood analyses [31]. So, we also examined the overlap between the CpGs with potential *DNAm*  
342  $\rightarrow$  *Current Smoking* effects and the genes implicated in the GWAS of blood cell counts [34] to  
343 probe the potential impact of the cell-count GWAS associations on the causal inference and cell-  
344 type enrichment. Similar overlap was examined for the subset of CpGs overlapping with  
345 enhancers in cord blood primary B cells.

346



350 **Illustrative MR-DoC models of causality between current smoking and blood DNAm at (A)**  
351 **cg23916896 and (B) cg05575921 in the AHR gene.**

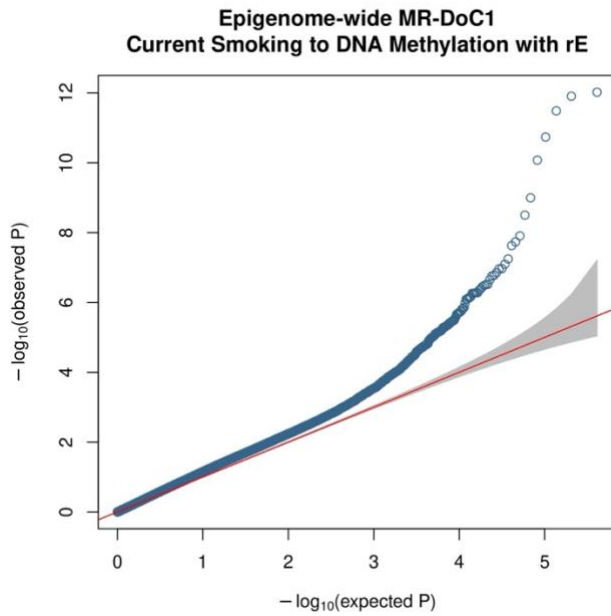
352 *We fitted five MR-DoC models at each CpG: (1) Smoking  $\rightarrow$  DNAm MR-DoC1 with horizontal*  
353 *pleiotropy, (2) Smoking  $\rightarrow$  DNAm MR-DoC1 with unique environmental confounding, (3) DNAm*  
354  *$\rightarrow$  Smoking MR-DoC1 with horizontal pleiotropy, (4) DNAm  $\rightarrow$  Smoking MR-DoC1 with unique*  
355 *environmental confounding, and (5) bidirectional MR-DoC2. Thus, for each CpG, three causal*  
356 *estimates were obtained in either direction of causation.*

357 *In the path diagrams, squares/rectangles indicate observed variables, circles indicate latent*  
358 *(unobserved variables), single-headed arrows indicate regression paths, and double-headed*  
359 *curved arrows indicate (co-)variance. The residual variance of smoking status liability is*  
360 *partitioned into additive genetic ( $A_{Smk}$ ) and unique environmental ( $E_{Smk}$ ) components. Likewise,*  
361 *the residual variance of DNAm is partitioned into  $A_{DNAm}$  and  $E_{DNAm}$ . The correlation between*  
362  *$A_{Smk}$  and  $A_{DNAm}$  represents the confounding between smoking and DNAm due to latent*  
363 *(unobserved) additive genetic factors, while the correlation between  $E_{Smk}$  and  $E_{DNAm}$  represents*  
364 *confounding due to latent unique environmental factors. Each model included age and sex as*  
365 *covariates of smoking status (not shown). DNAm  $\beta$ -values were residualized for standard*  
366 *biological and technical covariates used in EWAS (see Methods). The smoking PRS and the*  
367 *mQTL allelic scores were residualized for standard GWAS covariates, including genetic*  
368 *principal components and genotyping platform. In the path diagrams, the residualized PRS and*  
369 *mQTL allelic scores are regressed on respective latent factors, representing the underlying*  
370 *“true” standardized scores (mean = zero; variance = one). The coefficient of the path from the*  
371 *latent score to the observed score estimates the standard deviation of the observed score.*  
372 *Note. The paths are labeled by the point estimate and its S.E. in parentheses. For better*  
373 *readability, the path diagrams show only the within-individual part of the models fitted to data*  
374 *from twin pairs.*

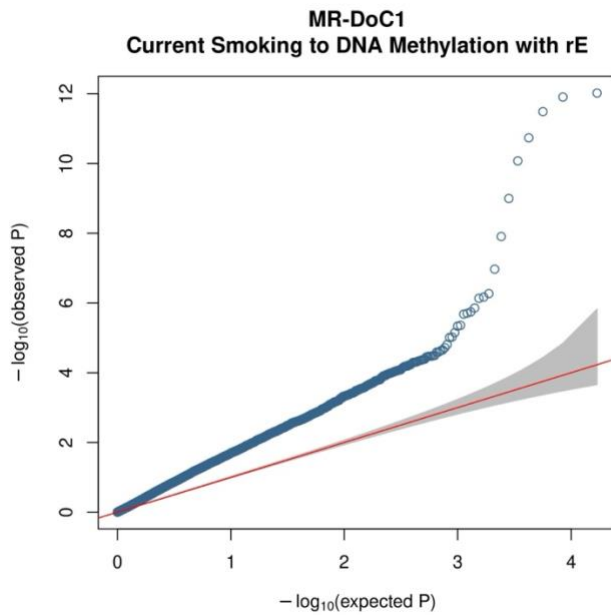
375

376

377 **Figure S2**  
378 *QQ Plot of MR-DoC1 models (with unique environmental confounding, rE) of Current Smoking*  
379 *→ DNAm at 411,169 epigenome-wide CpGs (Bayesian genomic inflation factor,  $\lambda = 1.09$ ).*

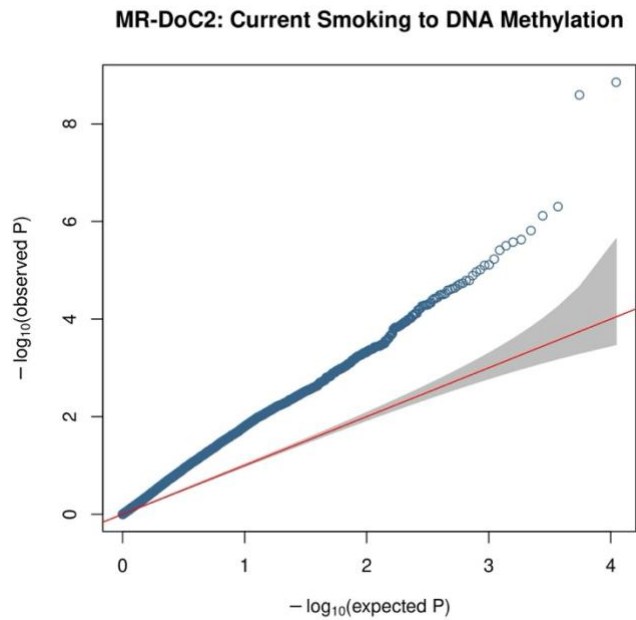


380  
381 **Figure S3**  
382 *QQ Plot of MR-DoC1 models (with unique environmental confounding, rE) of Current Smoking*  
383 *→ DNAm at 16,940 smoking-associated CpGs (Bayesian genomic inflation factor,  $\lambda = 1.20$ ).*

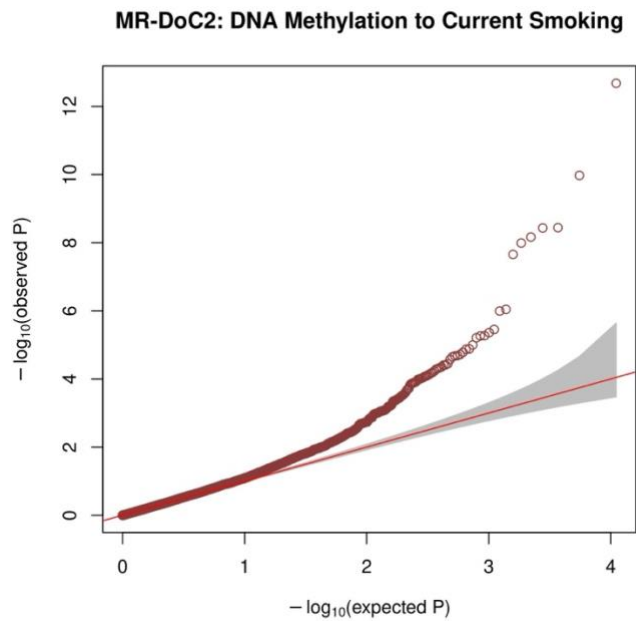


384  
385

386 **Figure S4**  
387 *QQ Plot of the Current Smoking → DNAm causal estimates in MR-DoC2 models across 11,124*  
388 *smoking-associated CpGs (Bayesian genomic inflation factor,  $\lambda = 1.20$ ).*



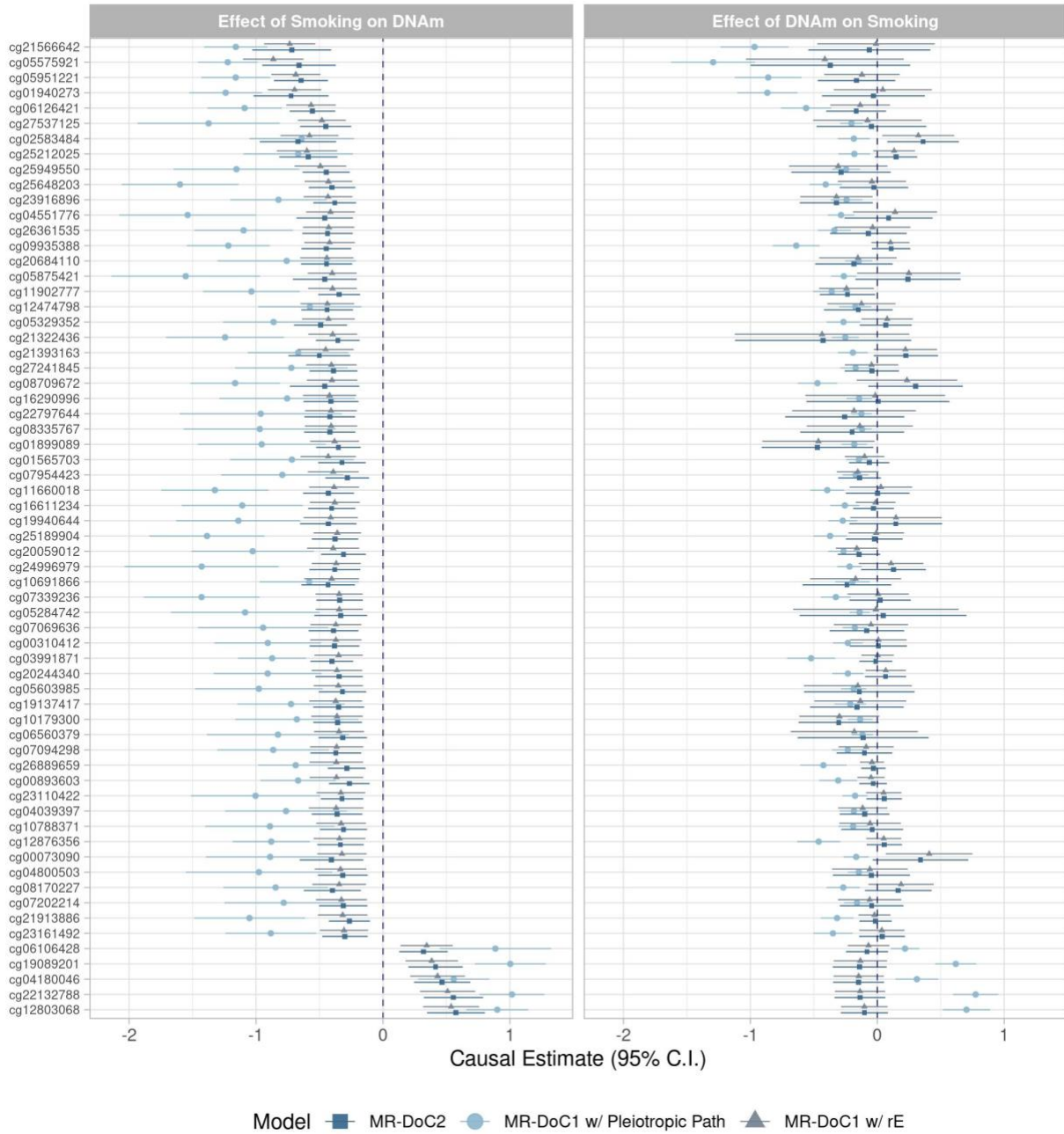
389  
390 **Figure S5**  
391 *QQ Plot of the DNAm → Current Smoking causal estimates in MR-DoC2 models across 11,124*  
392 *smoking-associated CpGs (Bayesian genomic inflation factor,  $\lambda = 1.01$ ).*



393  
394

395 **Figure S6**  
 396 *Bidirectional Causal Estimates at the 64 CpGs with Robust Evidence of the Causal Effects of*  
 397 *Current Smoking on DNA methylation*

Bidirectional Causal Estimates between Current Smoking and DNAm  
 At 64 CpGs where Current Smoking Likely Affects DNAm

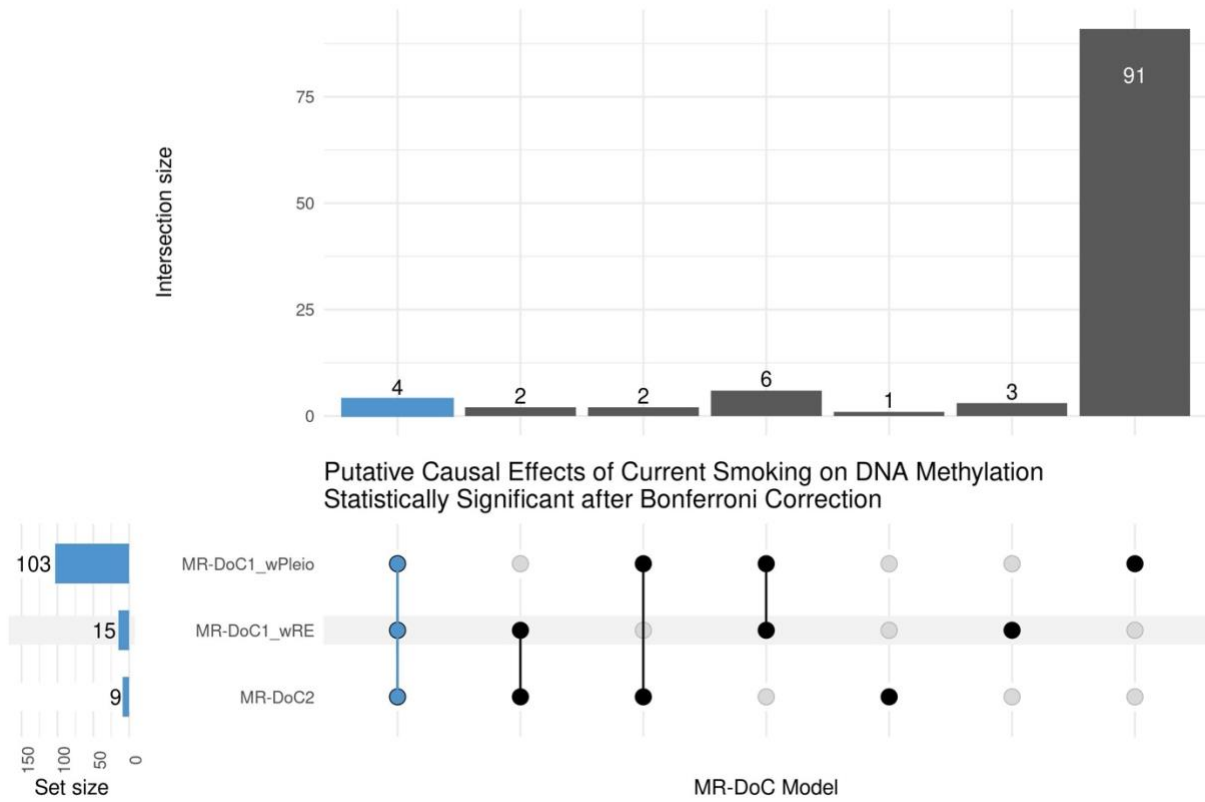


398  
 399 These CpGs did not show robust evidence for the reverse effects of DNAm on current smoking.  
 400 Please refer to **Supplementary Tables S1** (*Current Smoking* → *DNAm*) and **S2** (*DNAm* →  
 401 *Current Smoking*) for the corresponding data.  
 402



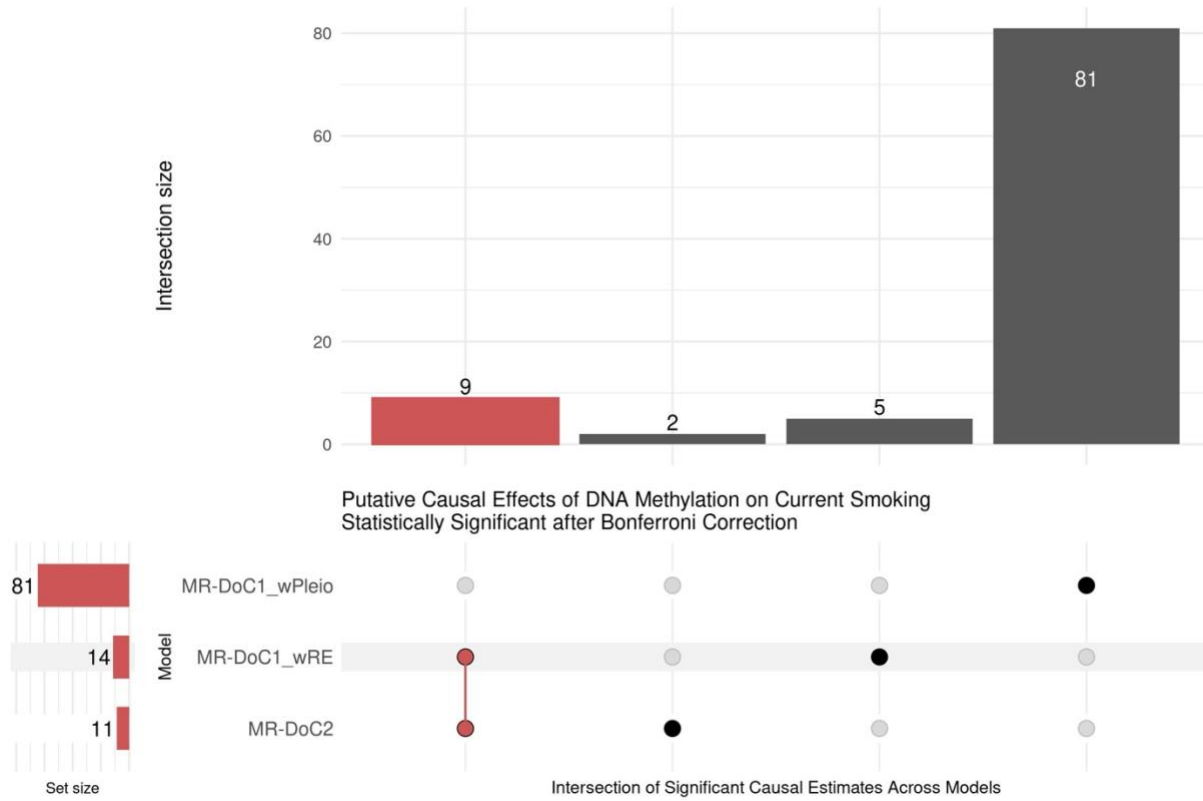


404 **Figure S6**  
 405 *Upset plot of the intersection of CpGs with statistically significant Current Smoking → DNAm*  
 406 *effects after Bonferroni correction in each of the three MR-DoC models*



407  
 408  
 409 *Note.* Please refer to **Supplementary Table S1** for the corresponding data  
 410  
 411

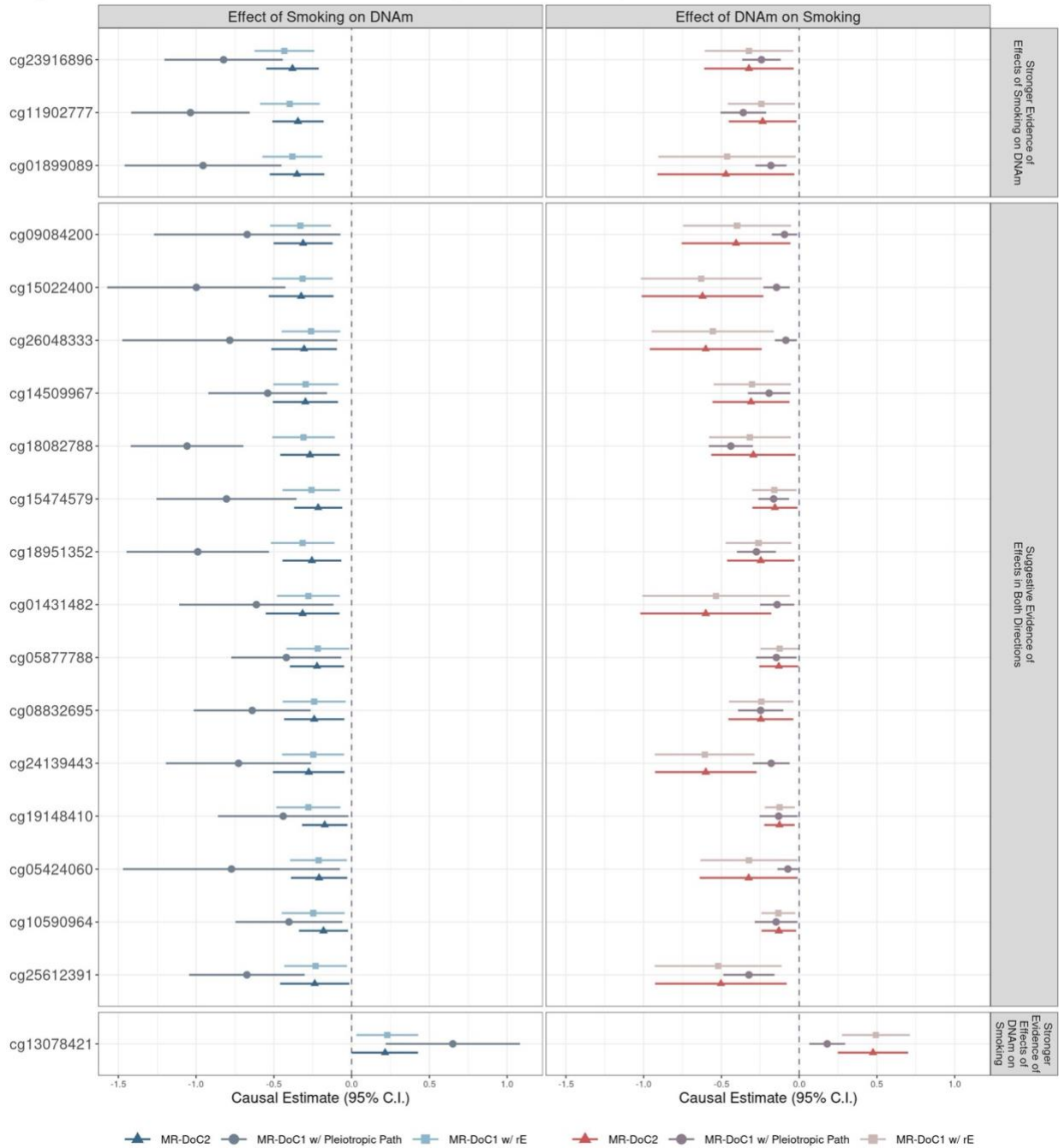
412 **Figure S8**  
 413 *Upset plot of the intersection of CpGs with statistically significant DNAm → Current Smoking*  
 414 *effects after Bonferroni correction in each of the three MR-DoC models*  
 415



416  
 417  
 418 *Note.* Please refer to **Supplementary Table S3** for the corresponding data  
 419  
 420  
 421

422 **Figure S9**  
 423 *19 CpGs with potential bidirectional causal effects between current smoking and DNA*  
 424 *methylation*

Suggestive Bidirectional Causal Effects between Smoking and DNAm at 19 CpGs



425 *Note.* Three CpGs had more robust evidence of *Current Smoking* → *DNAm* causal effects than  
 426 *vice versa*. One CpG had more robust evidence of *DNAm* → *Current Smoking* causal effects than  
 427 *vice versa*. The rest 15 CpGs had only suggestive evidence (consistent, nominally significant  
 428

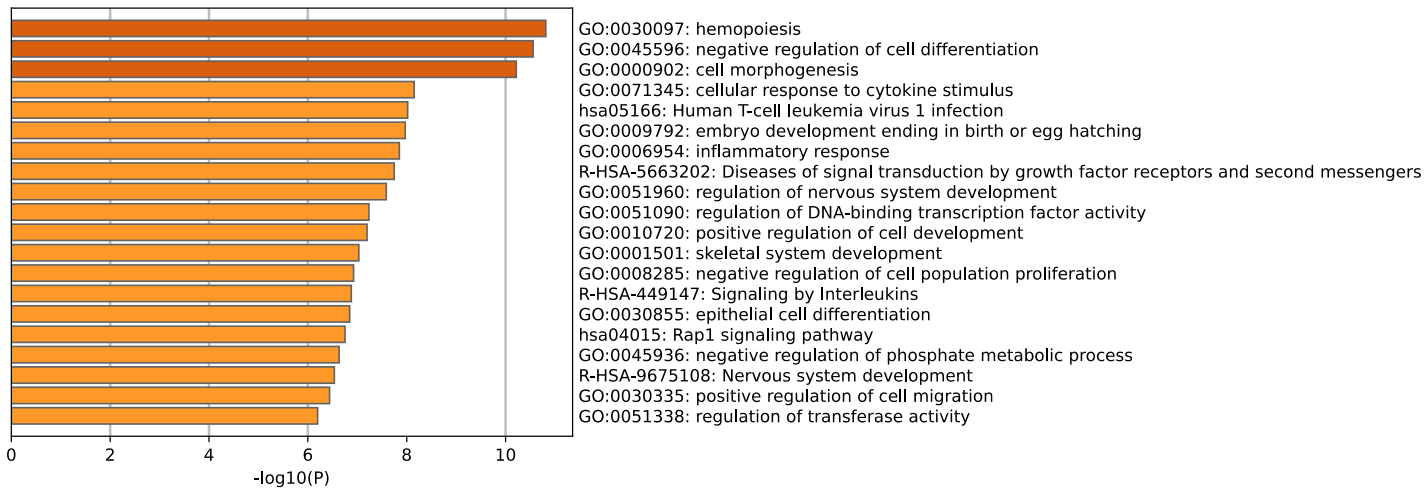
429 estimates across models) in both directions. Please refer to **Supplementary Tables S1-S4** for the  
430 corresponding data.

431  
432  
433

434 **Figure S10**

435 *Top Enriched Ontology Clusters in Metascape’s Gene Annotation and Functional Enrichment*  
436 *Analyses of the 525 CpGs (outside the MHC region) with Potential Current Smoking → DNAm*  
437 *effects*

438



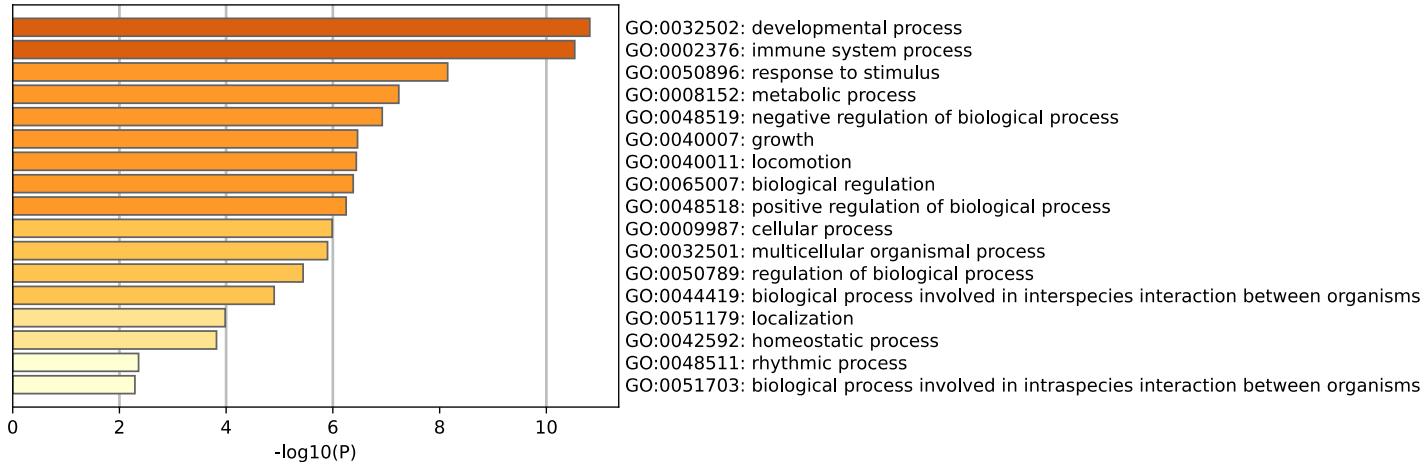
439  
440

441 *Note.* The “NearestGene” IDs from *Supplementary Table S1* were used as the input data for  
442 Metascape[26]. Please refer to **Supplementary Tables S5-S6** for the corresponding annotation  
443 and enrichment results.

444 As detailed in the Metascape manuscript[26], the program first identified all significant ontology  
445 terms, including GO/KEGG terms, canonical pathways, and hallmark gene sets. The significant  
446 terms (based on hypergeometric p-value <0.01 and >1.5-fold enrichment) were then clustered  
447 into a hierarchical tree based on Kappa-statistical similarities among their gene memberships.  
448 The tree was then cast into term clusters based on a threshold of 0.3 kappa score. The enrichment  
449 clusters and their underlying terms are marked as “Summary” and “Membership”, respectively,  
450 under the column *GroupID* in *Supplementary Table S24*. The “Summary” terms provide an  
451 overview of enriched, non-redundant ontology terms.

452  
453

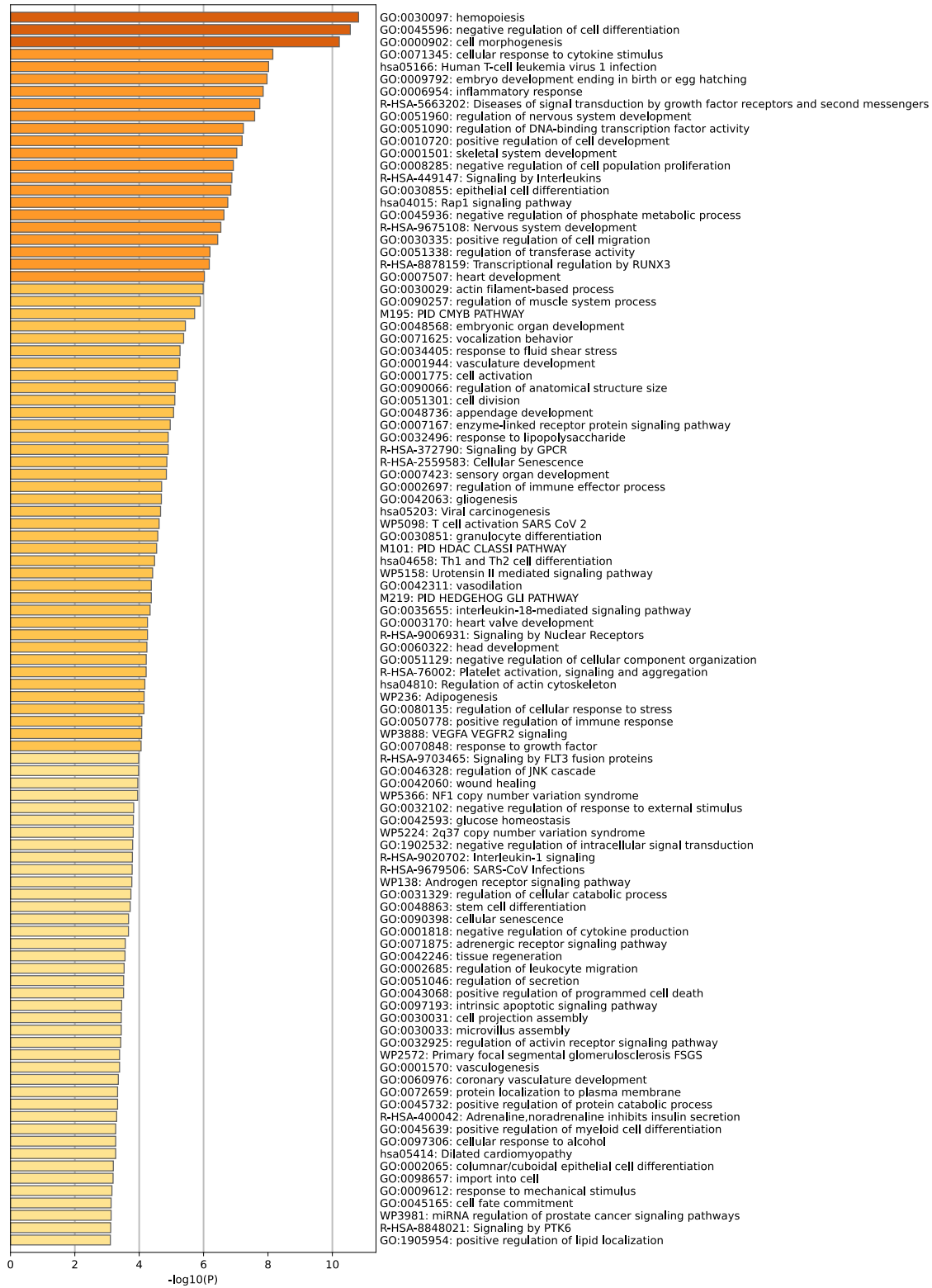
454 **Figure S11**  
 455 *Enrichment Results for Gene-Ontology (GO) Processes in Metascape’s Gene Annotation and*  
 456 *Functional Enrichment Analyses of the 525 CpGs (outside the MHC region) with Potential*  
 457 *Current Smoking → DNAm effects*



458  
 459  
 460 *Note.* The “NearestGene” IDs from *Supplementary Table S1* were used as the input data for  
 461 Metascape[26]. Please refer to **Supplementary Table S6** for all enrichment results.

462  
 463  
 464

465 **Figure S12**  
 466 *Top 100 Ontology Terms in Metascape's Gene Annotation and Functional Enrichment Analyses*  
 467 *of the 525 CpGs (outside the MHC region) with Potential Current Smoking → DNAm effects*



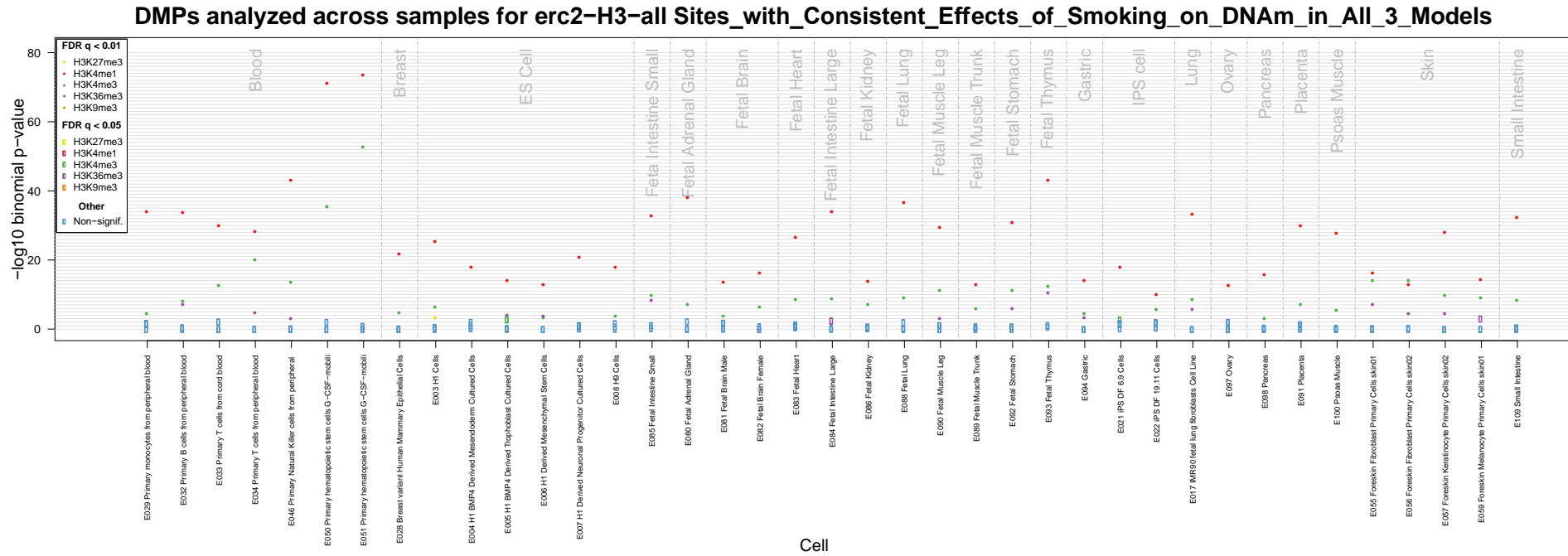
468

469 *Note.* The “NearestGene” IDs from *Supplementary Table S1* were used as the input data for  
470 Metascape[26]. Please refer to **Supplementary Table S6** for all enrichment results.  
471



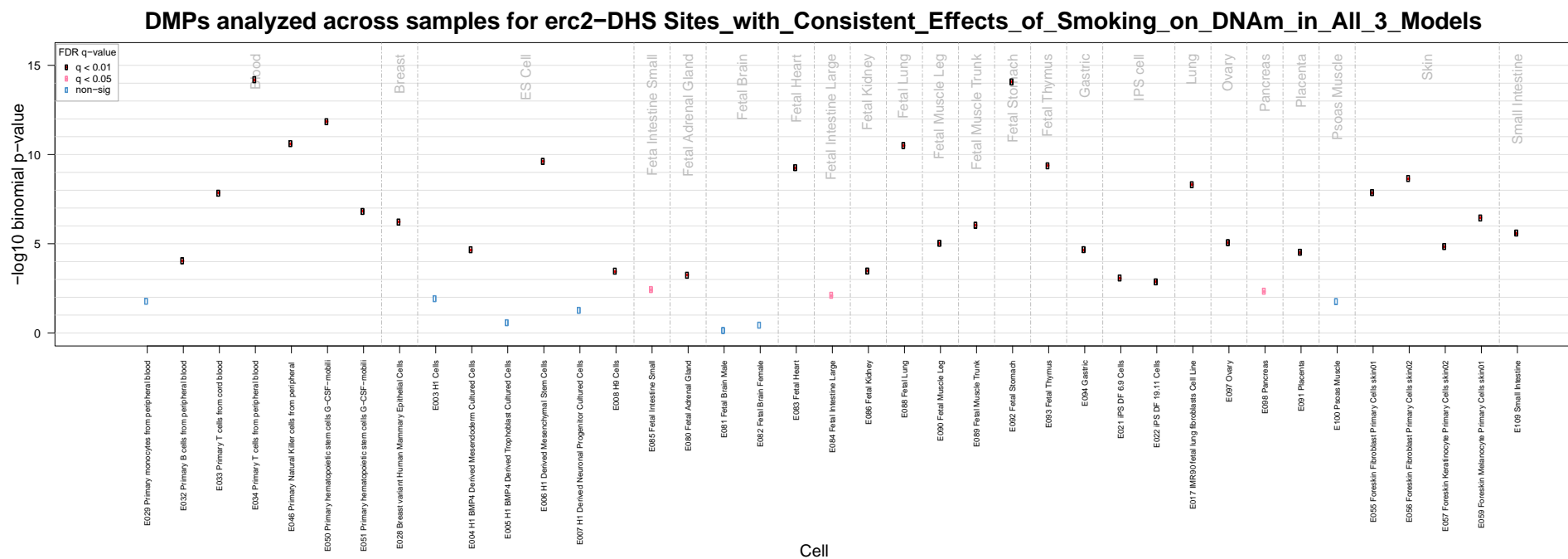


480 **Figure S14**  
 481 *eFORGE* analyses of overlap between histone-mark modifications and the 525 CpGs (outside the MHC region) with potential Current  
 482 Smoking → DNAm effects  
 483



484  
 485  
 486 *Note.* Please refer to **Supplementary Table S8** for the corresponding data.  
 487

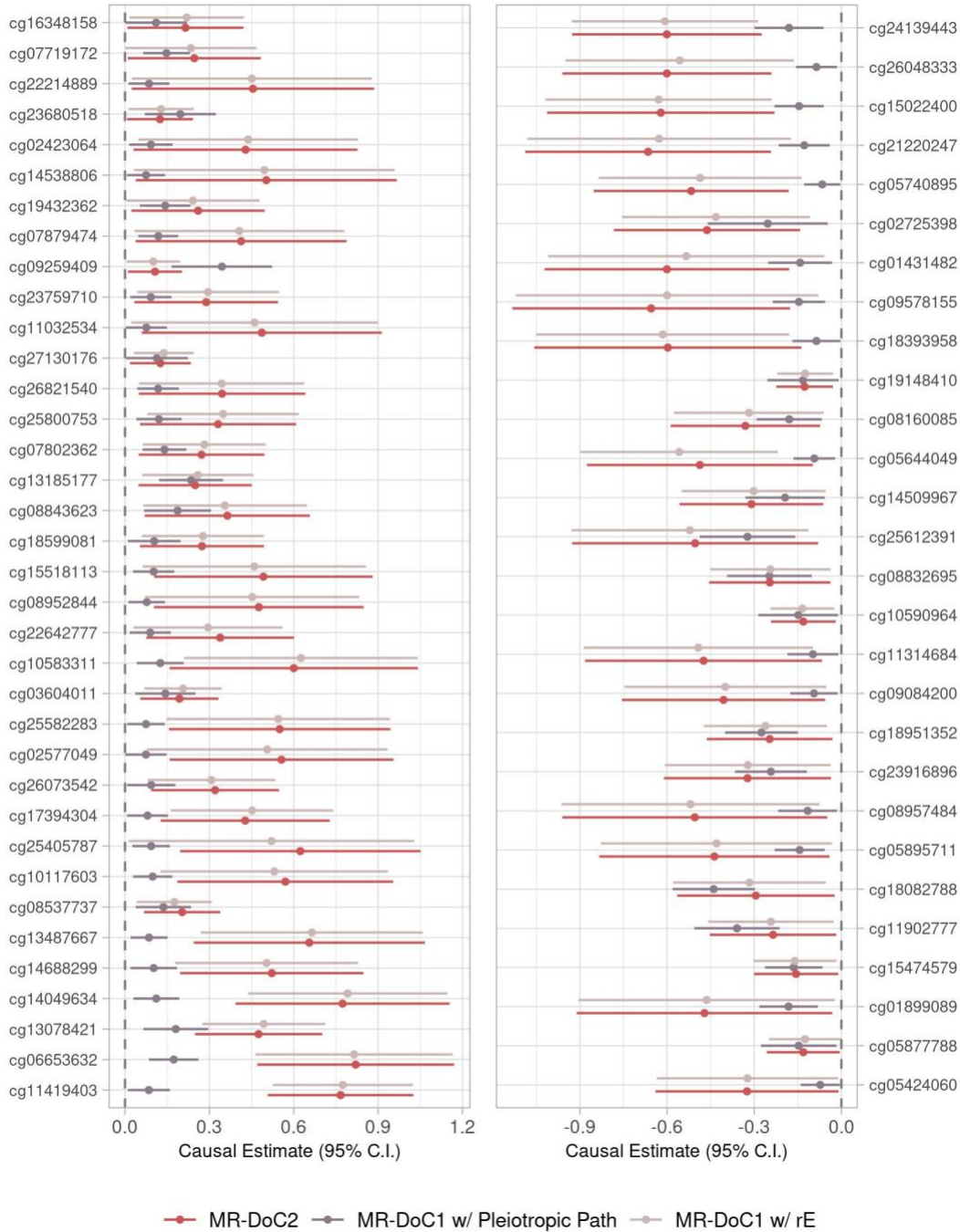
488 **Figure S15**  
 489 *eFORGE* analyses of overlap between DNase hypersensitivity (DHS) sites and the 546 CpGs with potential Current Smoking →  
 490 DNAm effects  
 491



492  
 493  
 494 *Note.* Please refer to **Supplementary Table S9** for the corresponding data.  
 495

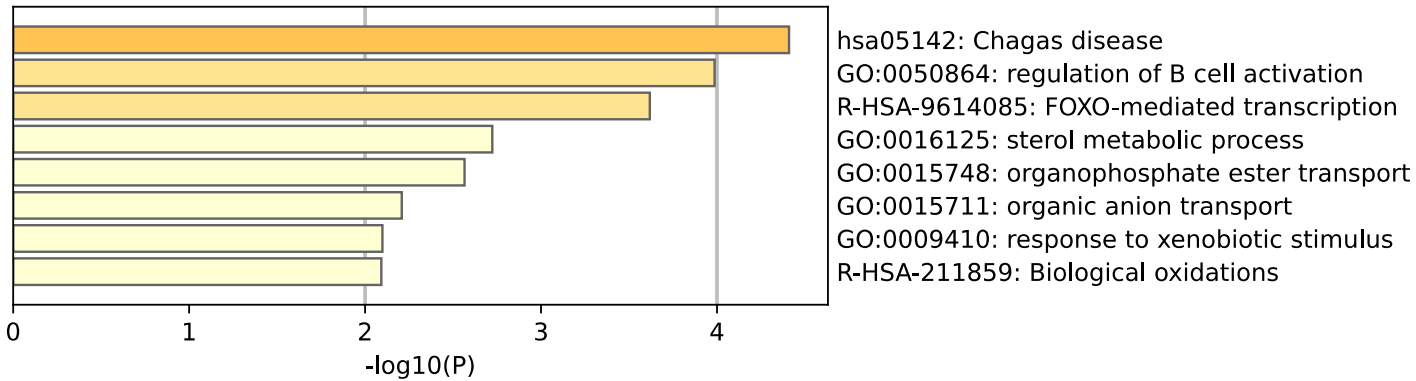
496 **Figure S16**  
 497 *64 CpGs with potential DNAm → Current Smoking effects, based on consistent, nominally*  
 498 *significant estimates across models*

CpGs with Consistent Estimates of the Effects of DNAm on Current Smoking



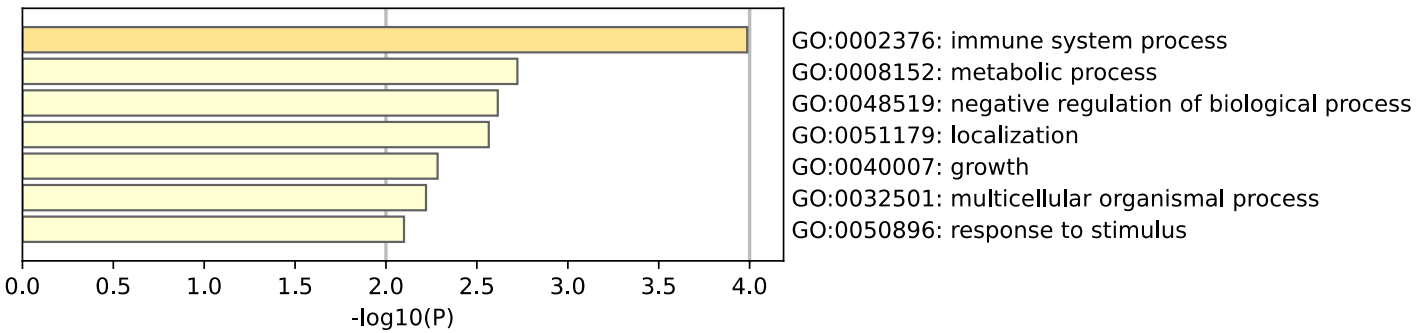
499 *Note.* These CpGs were used for the follow-up enrichment analyses with eFORGE[31] and  
 500 Metascape[26]. None of these sites are in the MHC region. Please refer to **Supplementary**  
 501 **Table S3** for the corresponding data.

503 **Figure S17**  
 504 *Top Ontology Clusters in Metascape’s Gene Annotation and Functional Enrichment Analyses of*  
 505 *the 64 CpGs with Potential DNAm → Current Smoking effects*  
 506



507  
 508 *Note.* The “NearestGene” IDs from *Supplementary Table S3* were used as the input data for  
 509 Metascape[26]. None of the ontology terms were significant after multiple-testing correction.  
 510 Please refer to **Supplementary Tables S10 and S11** for all annotation and enrichment results.  
 511

512 **Figure S18**  
 513 *Enrichment Results for Gene-Ontology (GO) Processes in Metascape’s Gene Annotation and*  
 514 *Functional Enrichment Analyses of the 64 CpGs with Potential DNAm → Current Smoking*  
 515 *effects*  
 516

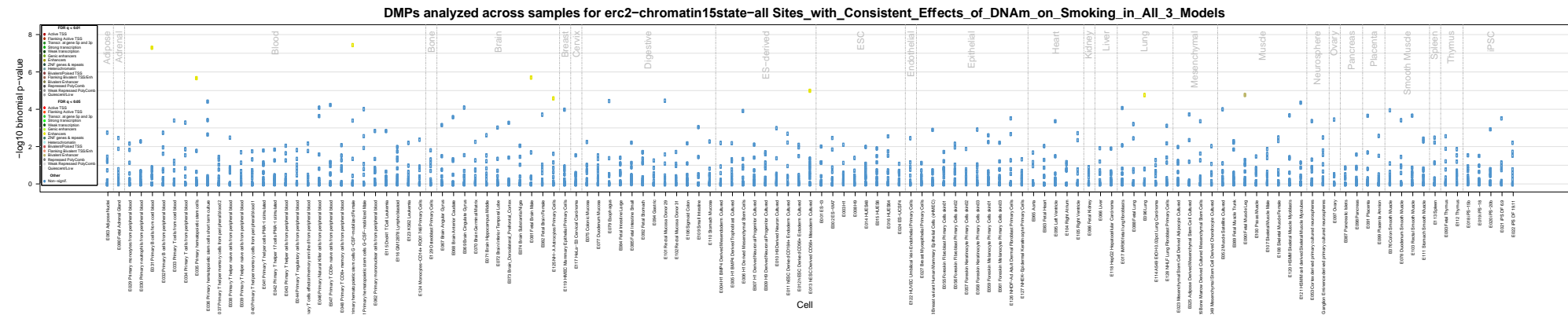


517  
 518 *Note.* The “NearestGene” IDs from *Supplementary Table S3* were used as the input data for  
 519 Metascape[26]. None of the ontology terms were significant after multiple-testing correction.  
 520 Please refer to **Supplementary Table S11** for all enrichment results.  
 521  
 522

523  
524

**Figure S19**

*e*FORGE analyses of overlap between gene-regulatory chromatin states and the 64 CpGs with potential DNAm → Current Smoking effects



525  
526  
527

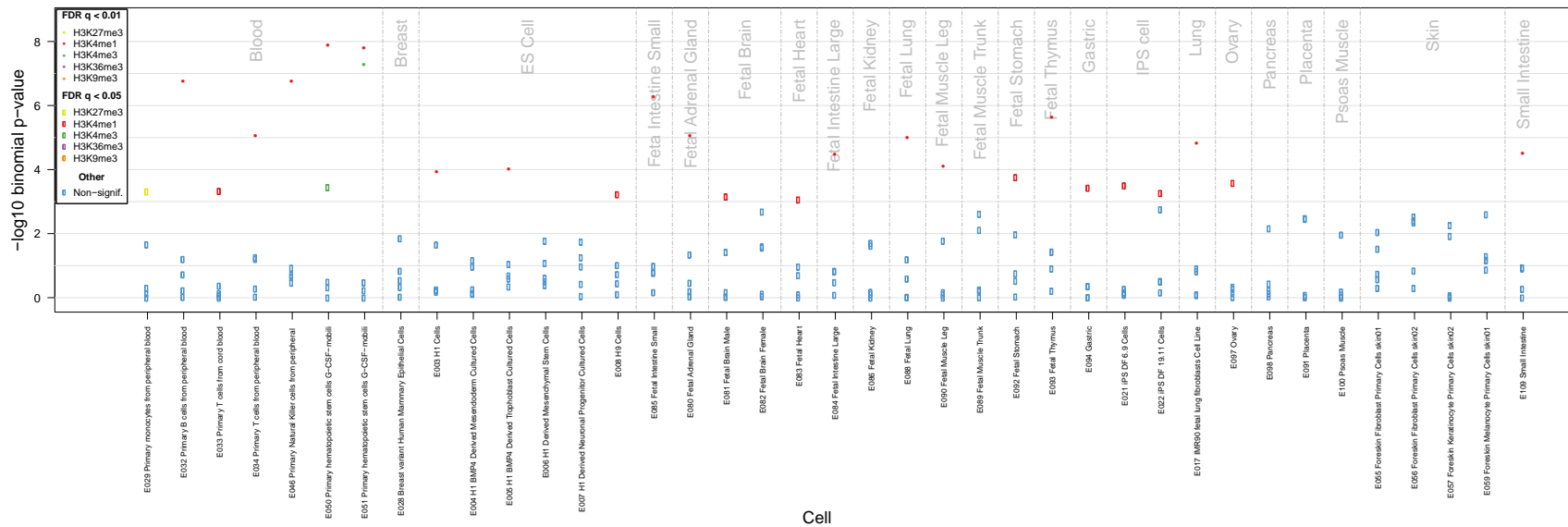
Note. Please refer to **Supplementary Table S12** for the corresponding data.

528  
529

**Figure S20**

*eFORGE* analyses of overlap between histone-mark modifications and the 64 CpGs with potential DNAm → Current Smoking effects

**DMPs analyzed across samples for erc2-H3-all Sites\_with\_Consistent\_Effects\_of\_DNAm\_on\_Smoking\_in\_All\_3\_Models**



530  
531  
532  
533

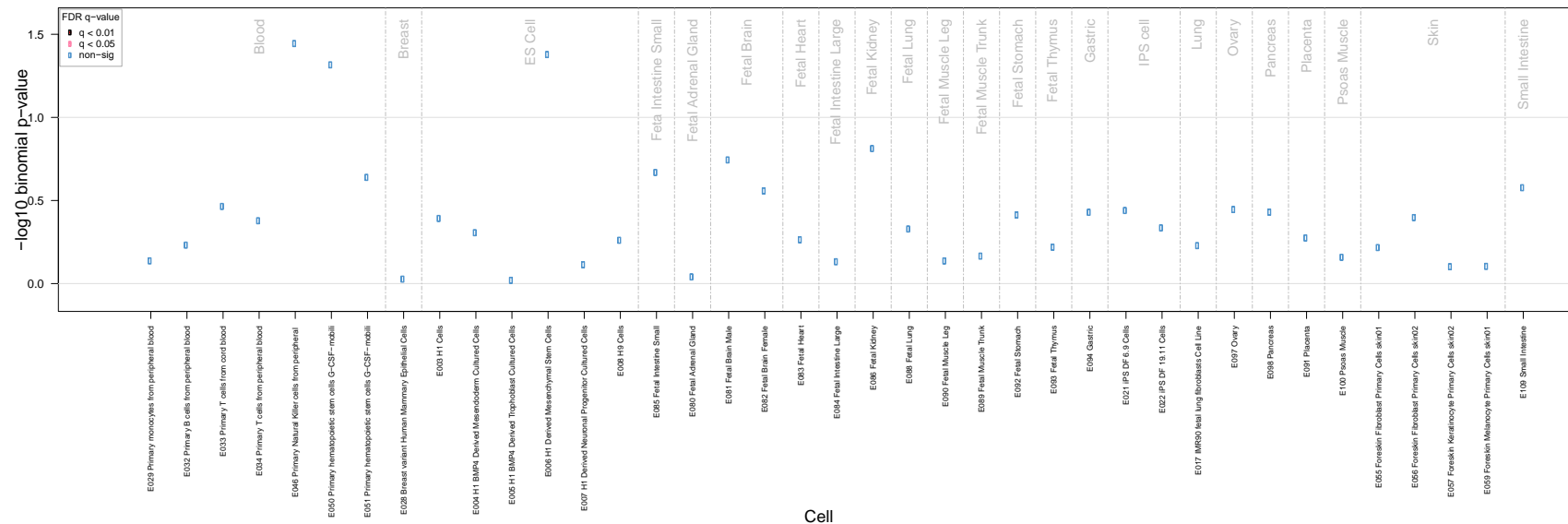
*Note.* Please refer to **Supplementary Table S13** for the corresponding data.

534  
535

**Figure S21**

*eFORGE* analyses of overlap between DNase hypersensitivity (DHS) sites and the 64 CpGs with potential DNAm → Current Smoking effects

**DMPs analyzed across samples for *erc2*-DHS Sites\_with\_Consistent\_Effects\_of\_DNAm\_on\_Smoking\_in\_All\_3\_Models**



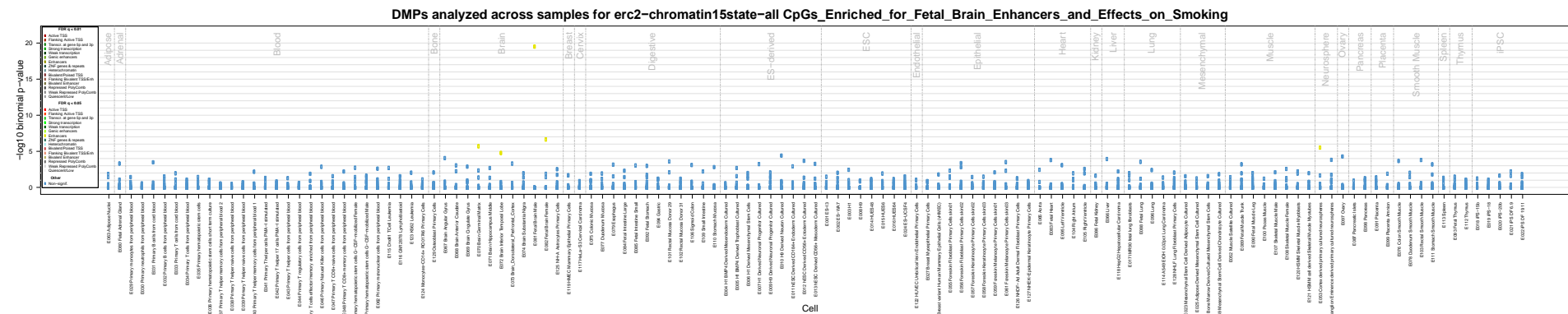
536  
537  
538  
539  
540

*Note.* Please refer to **Supplementary Table S14** for the corresponding data.

541  
542  
543  
  
544  
545  
546  
547  
548

### Figure S22

Follow-up eFORGE analyses of overlap between gene-regulatory chromatin states and the 21 CpGs enriched for overlap with Enhancers in the “Fetal Brain Male” sample in Figure S19/Table S5



Note. Please refer to **Supplementary Table S15** for the corresponding data.

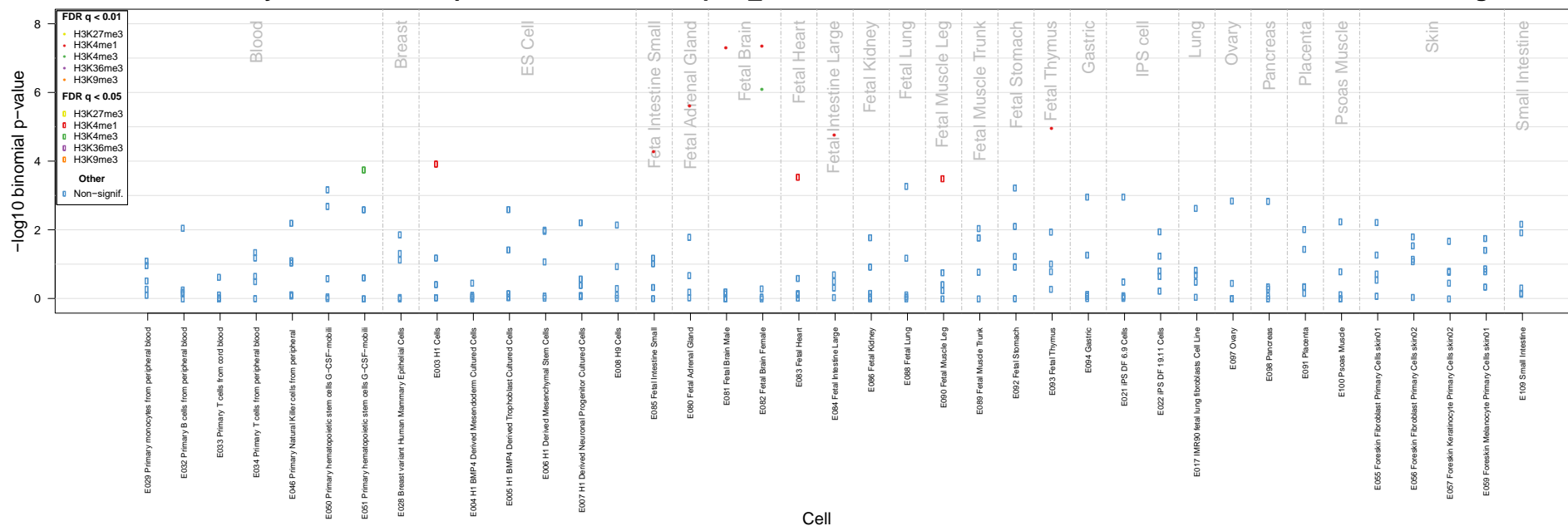


549  
550  
551

**Figure S23**

Follow-up eFORGE analyses of overlap between histone-mark modifications and the 21 CpGs enriched for overlap with Enhancers in the “Fetal Brain Male” sample in Figure S19/Table S5

**DMPs analyzed across samples for erc2-H3-all CpGs\_Enriched\_for\_Fetal\_Brain\_Enhancers\_and\_Effects\_on\_Smoking**



552  
553  
554  
555  
556

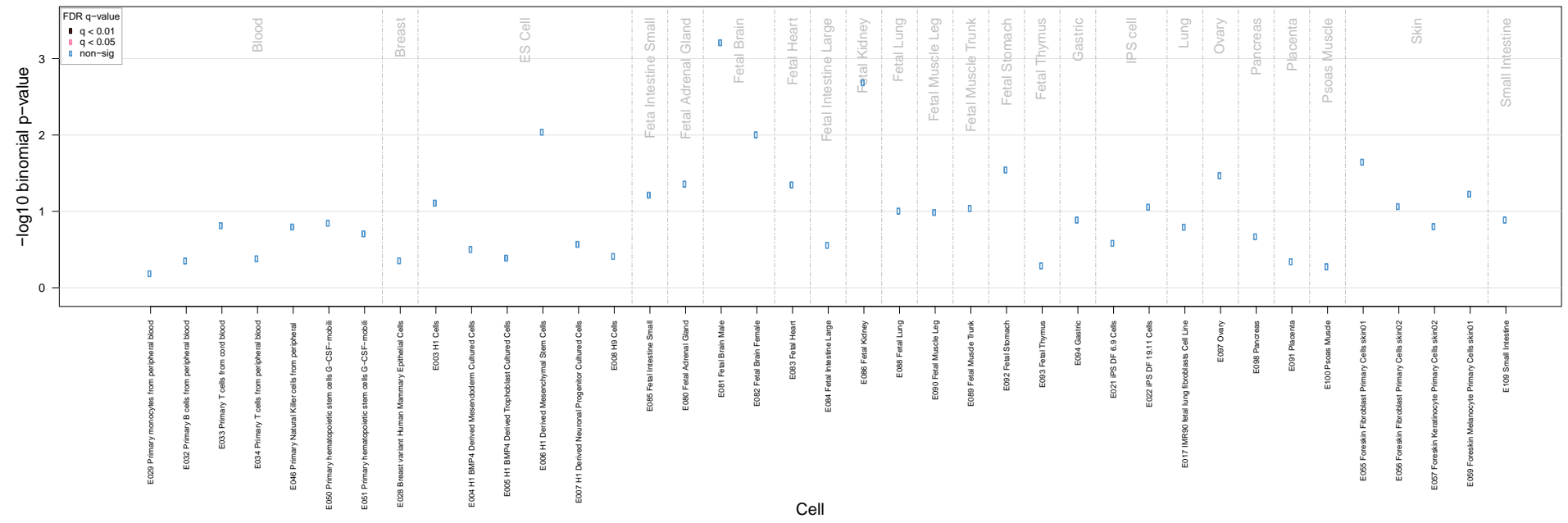
Note. Please refer to **Supplementary Table S16** for the corresponding data.

557  
558  
559

**Figure S24**

Follow-up eFORGE analyses of overlap between DNase hypersensitivity (DHS) sites and the 21 CpGs enriched for overlap with Enhancers in the “Fetal Brain Male” sample in Figure S19/Table S5

**DMPs analyzed across samples for erc2–DHS CpGs\_Enriched\_for\_Fetal\_Brain\_Enhancers\_and\_Effects\_on\_Smoking**



560  
561  
562  
563  
564

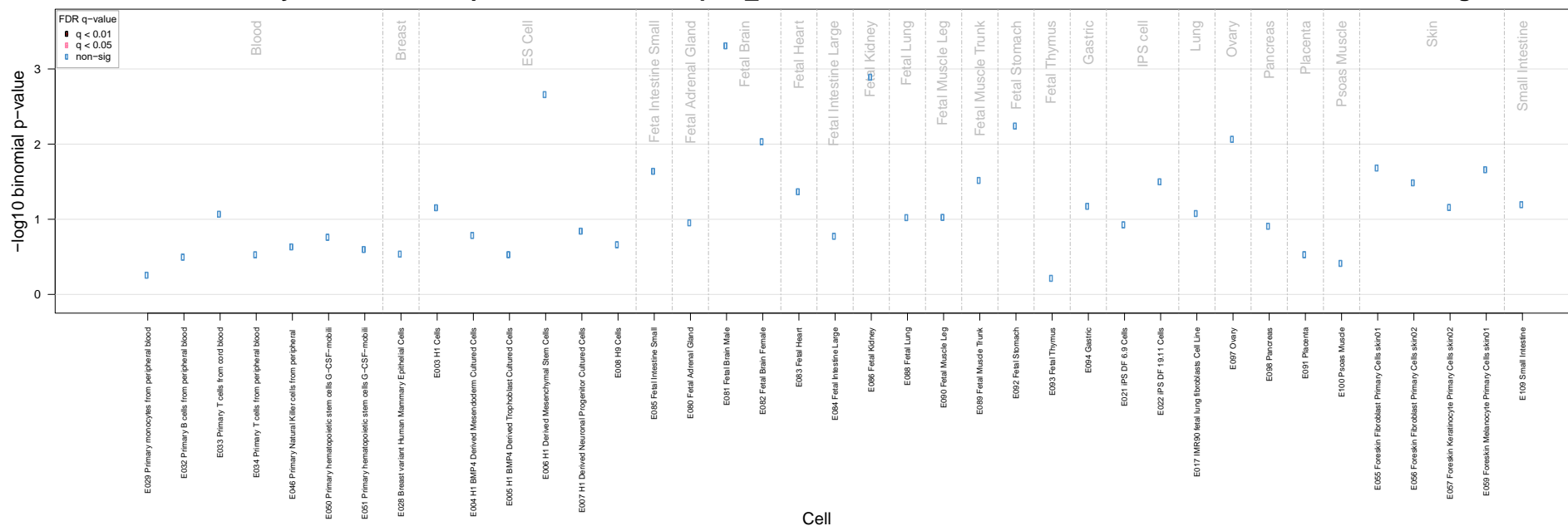
Note. Please refer to **Supplementary Table S17** for the corresponding data.

565  
566  
567

**Figure S25**

Follow-up eFORGE analyses of overlap between DNase hypersensitivity (DHS) sites and the 17 CpGs enriched for overlap with H3K4me3 modifications in the “Fetal Brain Female” samples in Supplementary Figure S23/Table S9

**DMPs analyzed across samples for erc2-DHS CpGs\_Enriched\_for\_Fetal\_Brain\_H3K4me3\_and\_Effects\_on\_Smoking**



568  
569  
570

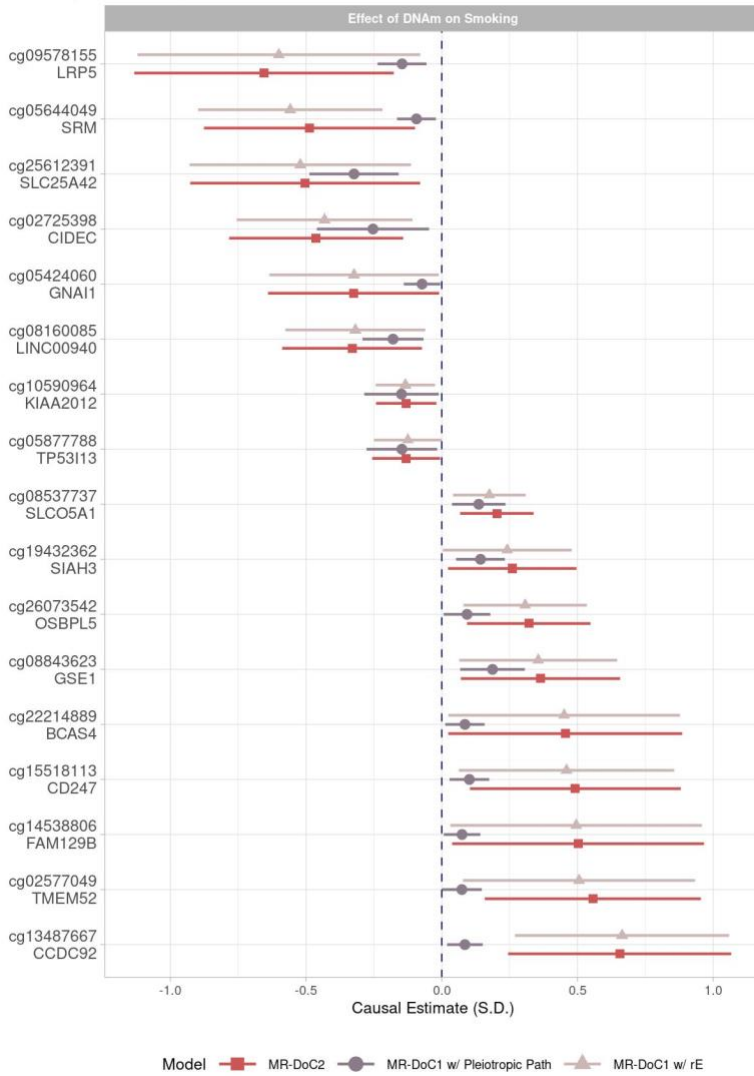
Note. Please refer to **Supplementary Table S20** for the corresponding data.

571  
572  
573  
574

**Figure S26**

*Estimated DNAm → Current Smoking effects at the 17 CpGs showing highly specific enrichment for overlap with gene-regulatory elements in the brain in Figure 5*

Estimated Effects of DNAm on Current Smoking  
At 17 CpGs Showing Enrichment for Functional Elements in the Brain



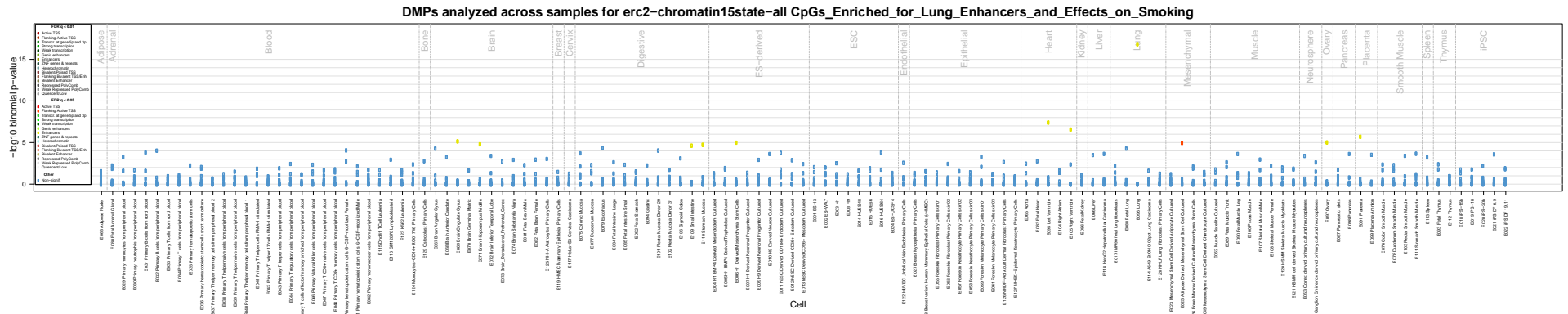
575  
576  
577  
578

*Note.* The Y-axis shows the probe ID and the “Nearest Gene”. For the corresponding data, please refer to **Supplementary Table S3**.

579  
580  
581

### Figure S27

*e*FORGE analyses of overlap between gene-regulatory chromatin states and the 18 CpGs underlying the enriched overlap with Enhancers in the “Lung” sample in Figure S19/Table S5



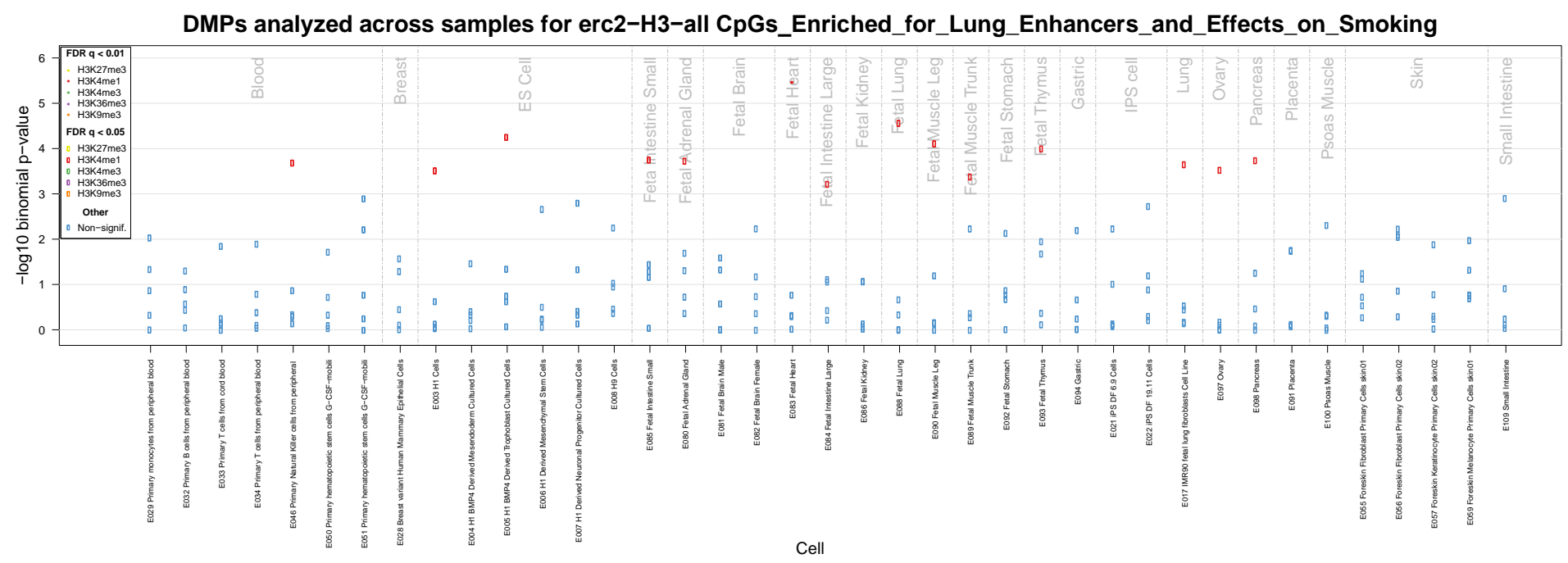
582  
583  
584  
585  
586

Note. Please refer to **Supplementary Table S21** for the corresponding data.

587  
588  
589  
590

**Figure S28**

*eFORGE* analyses of overlap between histone-mark modifications and the 18 CpGs underlying the enriched overlap with Enhancers in the “Lung” sample in Figure S19/Table S5



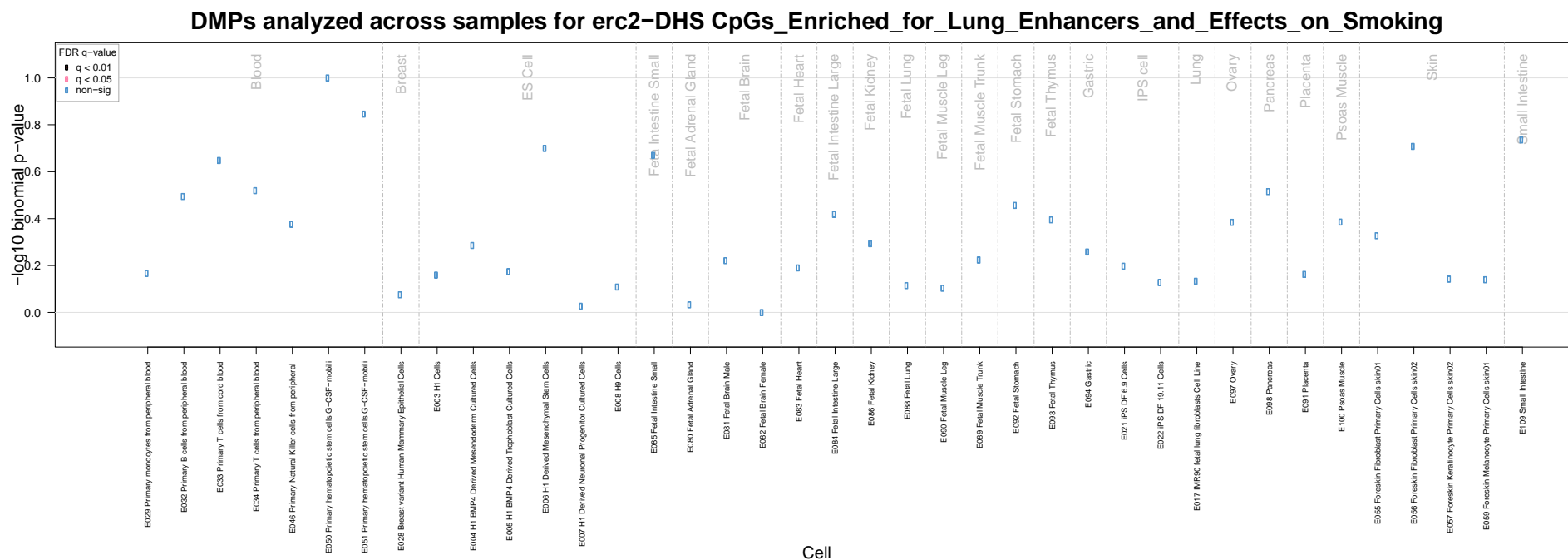
591  
592  
593

Note. Please refer to **Supplementary Table S22** for the corresponding data.

594  
595  
596  
597

**Figure S29**

*eFORGE* analyses of overlap between DNase hypersensitivity (DHS) sites and the 18 CpGs underlying the enriched overlap with Enhancers in the “Lung” sample in Figure S19/Table S5



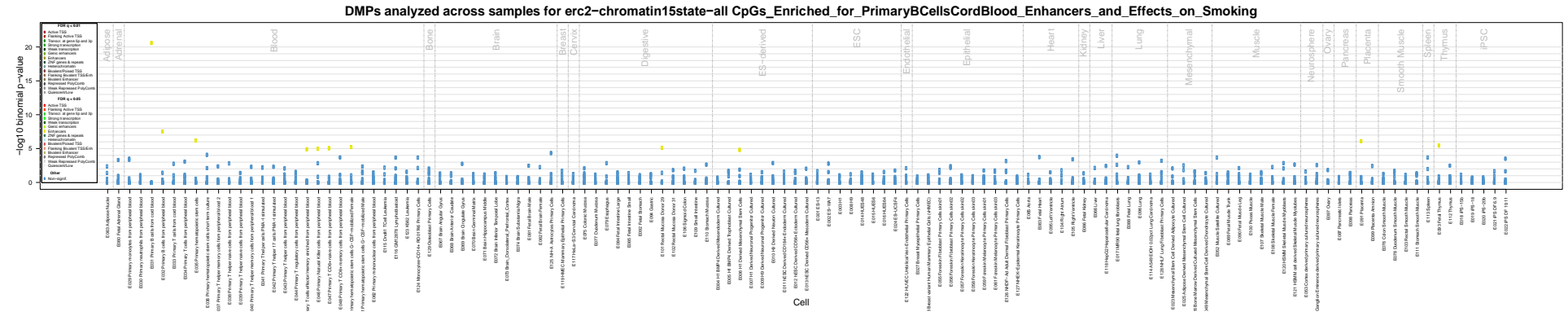
598  
599  
600  
601  
602  
603

Note. Please refer to **Supplementary Table S23** for the corresponding data.

604  
605  
606

**Figure S30**

*eFORGE* analyses of overlap between gene-regulatory chromatin states and the 18 CpGs underlying the enriched overlap with Enhancers in the “Primary B cells from cord blood” sample in Figure S19/Table S5



607  
608  
609  
610  
611

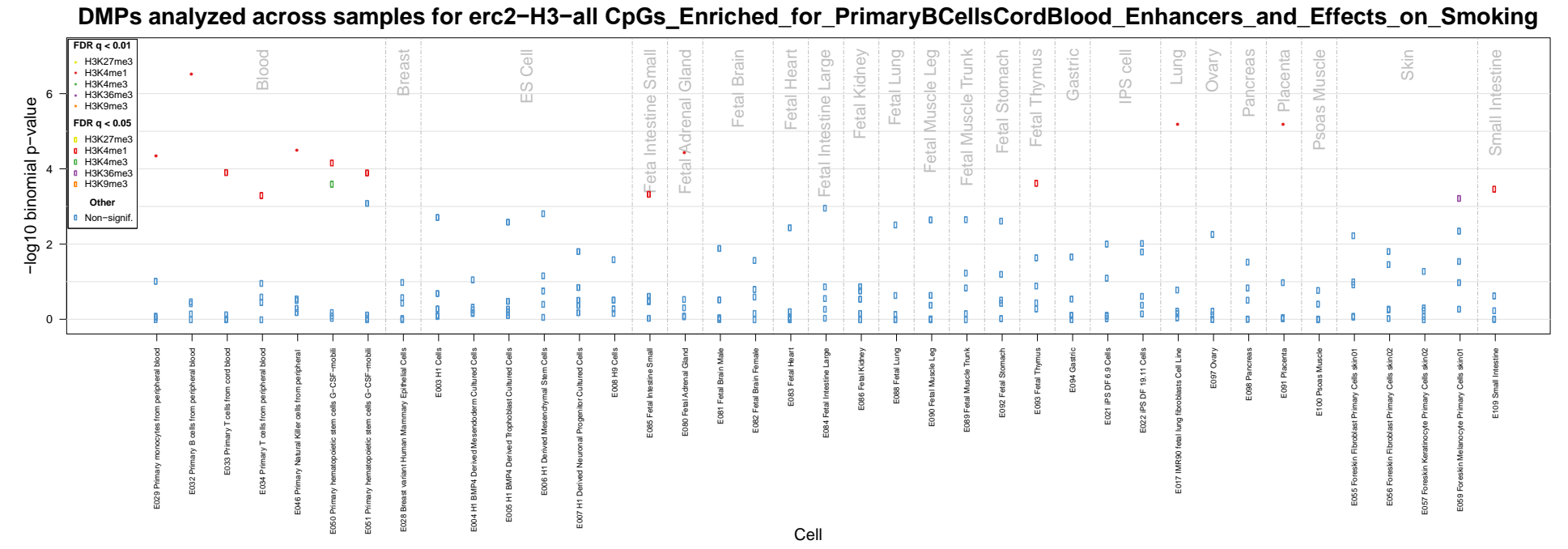
Note. Please refer to **Supplementary Table S24** for the corresponding data.



612  
613  
614  
615

**Figure S31**

*eFORGE* analyses of overlap between histone-mark modifications and the 18 CpGs underlying the enriched overlap with Enhancers in the “Primary B cells from cord blood” sample in Figure S19/Table S5



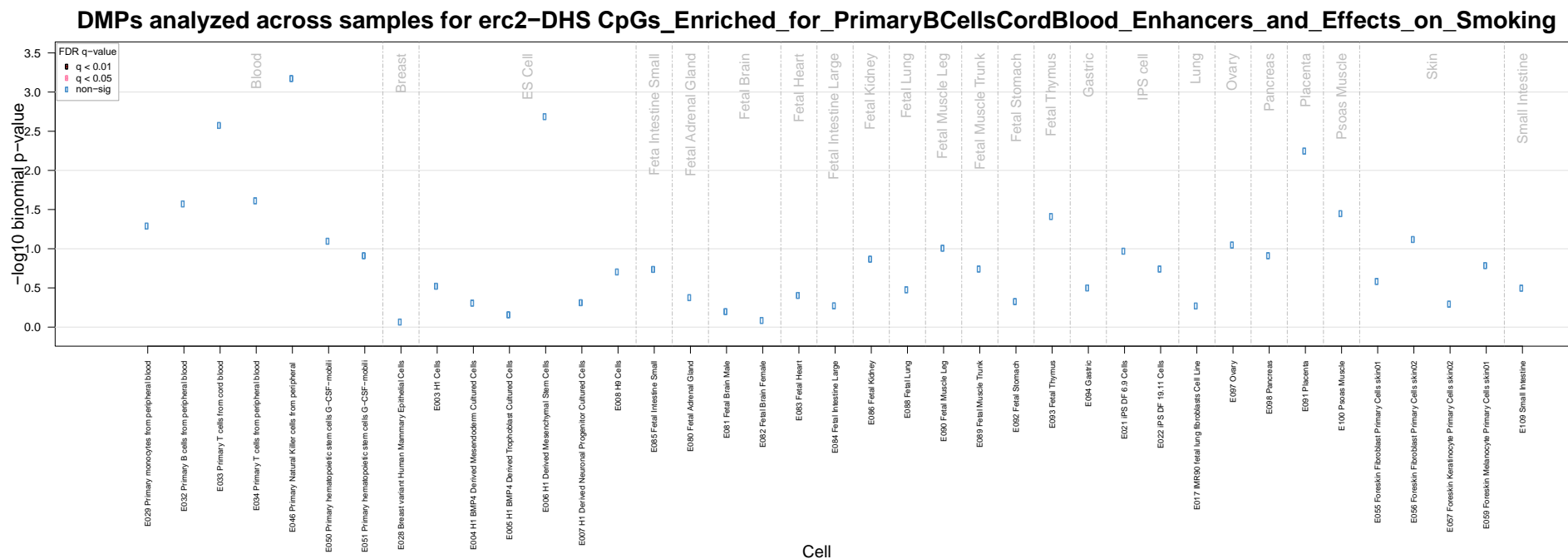
616  
617  
618  
619  
620  
621  
622

*Note.* Please refer to **Supplementary Table S25** for the corresponding data.

623  
624  
625  
626

**Figure S32**

*eFORGE analyses of overlap between DNase hypersensitivity (DHS) sites and the 18 CpGs underlying the enriched overlap with Enhancers in the “Primary B cells from cord blood” sample in Figure S19/Table S5*



627  
628  
629  
630  
631

*Note. Please refer to **Supplementary Table S26** for the corresponding data.*

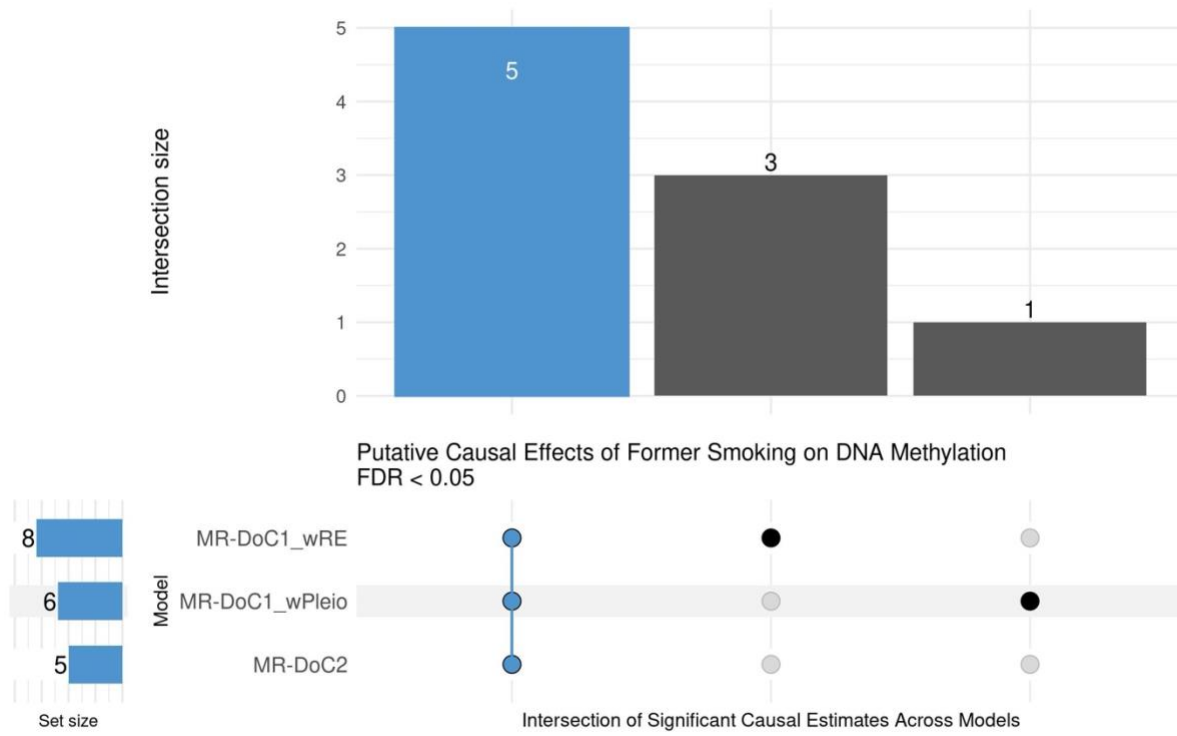
632

**Figure S33**

633

*Upset plot of the intersection of CpGs with statistically significant (FDR <0.05) Former Smoking → DNAm effects in each of the three MR-DoC models*

634



635

636

637

*Note.* Please refer to **Supplementary Table S27** for the corresponding data.

638

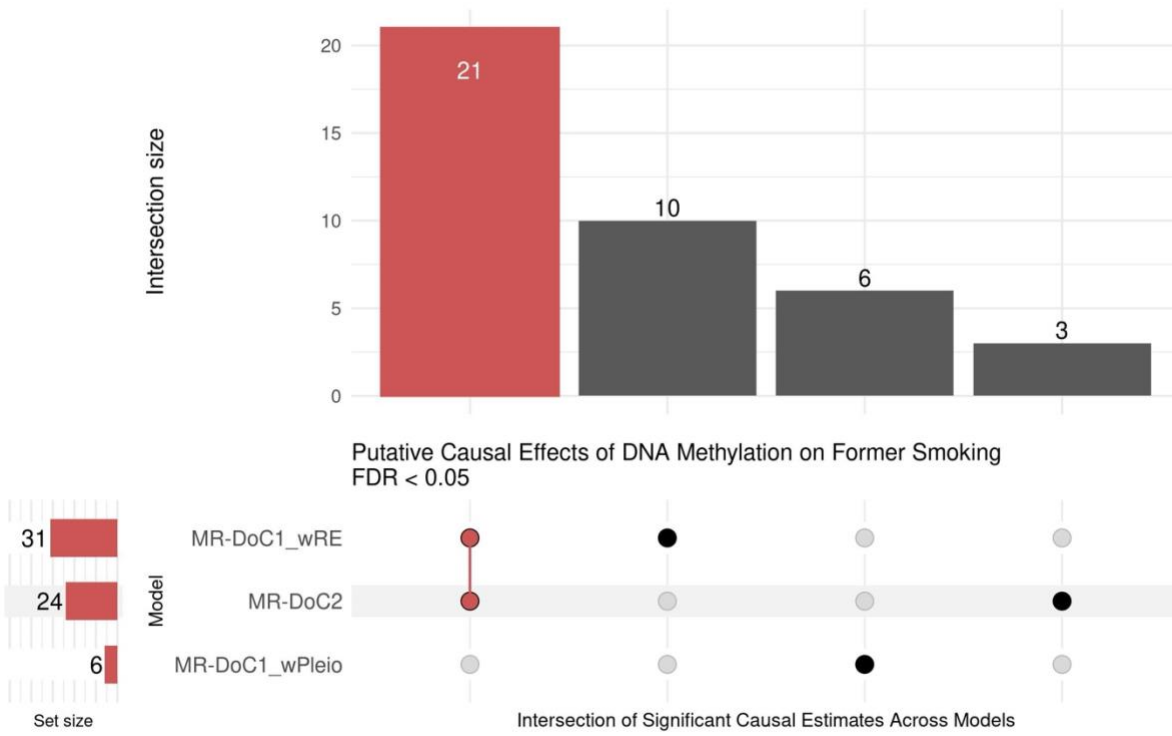
639

640

641  
642  
643

**Figure S34**

*Upset plot of the intersection of CpGs with statistically significant (FDR < 0.05) DNAm → Former Smoking effects in each of the three MR-DoC models*



644  
645  
646  
647  
648

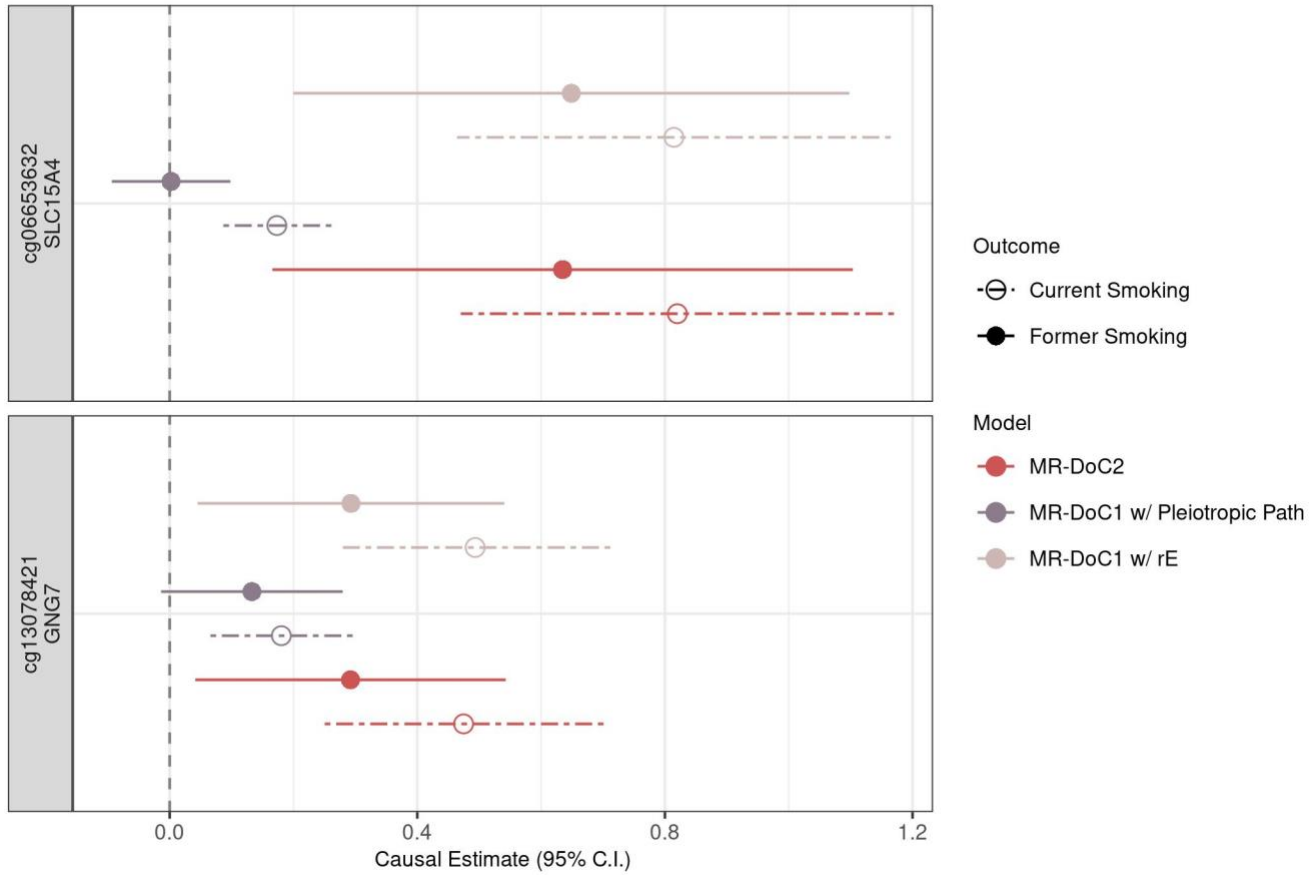
*Note.* Please refer to **Supplementary Table S29** for the corresponding data.

649  
650  
651  
652

**Figure S35**

*Estimated DNAm → Former Smoking effects at the two former-smoking-associated CpGs that showed robust evidence of DNAm → Current Smoking effects*

Putative Effects of DNA Methylation on Current Smoking  
Compared to the Estimated Effects on Former Smoking



653  
654  
655  
656  
657

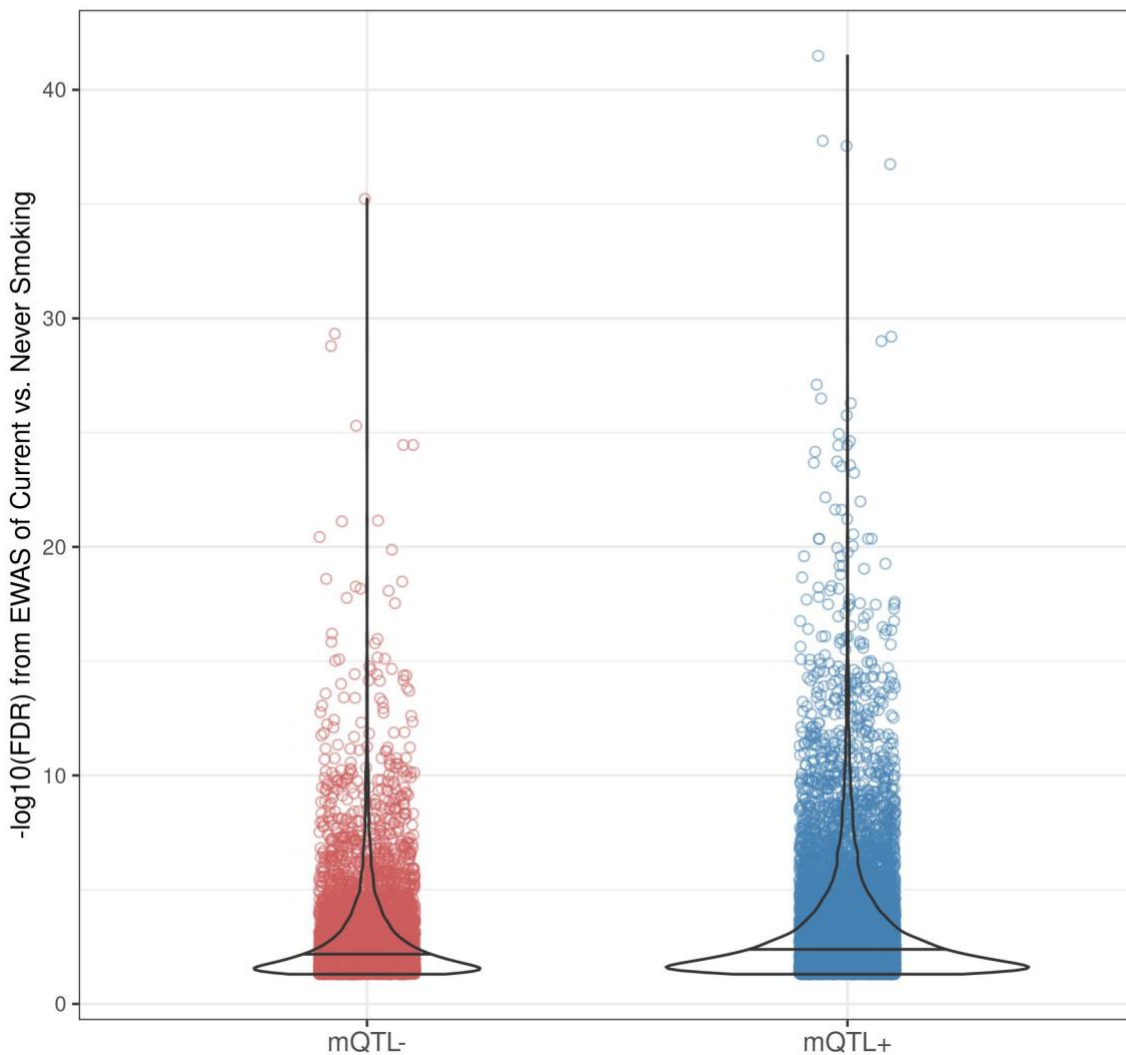
*Note.* Please refer to **Supplementary Tables S3 and S29** for the corresponding data.

658  
659  
660  
661

**Figure S36**

*Prior EWAS association statistics of smoking-associated CpGs stratified by whether the CpG was identified as having an mQTL allelic score with F-statistic >10 in the current study*

EWAS Meta-Analysis Association Statistics of Smoking-Associated CpGs  
With and Without an mQTL Allelic Score with F-statistic >10



662  
663  
664  
665  
666  
667

*On the X-axis, “mQTL-” indicates the CpGs without an mQTL allelic score with  $F > 10$  (5,816 CpGs), and “mQTL+” indicates the CpGs with an mQTL allelic score with  $F > 10$  (11,124 CpGs). The Y-axis shows the  $-\log_{10}(FDR)$  values of the association results from the previous EWAS meta-analysis of current vs. never smoking[19]. The “mQTL-” CpGs were not tested for DNAm  $\rightarrow$  Smoking causal effects in the current study.*

668 **References**

- 669 1. Ligthart L, van Beijsterveldt CEM, Kevenaar ST, de Zeeuw E, van Bergen E, Bruins S, et al. The  
670 Netherlands Twin Register: Longitudinal Research Based on Twin and Twin-Family Designs. *Twin*  
671 *Research and Human Genetics*. 2019;22(6):623–36.
- 672 2. Minică CC, Dolan CV, Boomsma DI, De Geus E, Neale MC. Extending Causality Tests with Genetic  
673 Instruments: An Integration of Mendelian Randomization with the Classical Twin Design. *Behavior*  
674 *Genetics*. 2018;48(4):337–49.
- 675 3. Castro-de-Araujo LFS, Singh M, Zhou Y, Vinh P, Verhulst B, Dolan CV, et al. MR-DoC2:  
676 Bidirectional Causal Modeling with Instrumental Variables and Data from Relatives. *Behavior*  
677 *Genetics*. 2023 Feb 1;53(1):63–73.
- 678 4. Singh M, Verhulst B, Vinh P, Zhou Y (Daniel), Castro-de-Araujo LFS, Hottenga JJ, et al. Using  
679 Instrumental Variables to Measure Causation over Time in Cross-Lagged Panel Models.  
680 *Multivariate Behavioral Research*. 2024 Feb 15;59(2):342–70.
- 681 5. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global  
682 reference for human genetic variation. *Nature*. 2015 Oct 1;526(7571):68–74.
- 683 6. Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation.  
684 *Nature Genetics*. 2016 Oct 1;48(10):1279–83.
- 685 7. Francioli LC, Menelaou A, Pulit SL, van Dijk F, Palamara PF, Elbers CC, et al. Whole-genome  
686 sequence variation, population structure and demographic history of the Dutch population. *Nature*  
687 *Genetics*. 2014 Aug 1;46(8):818–25.
- 688 8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components  
689 analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006 Aug  
690 1;38(8):904–9.
- 691 9. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array  
692 with single CpG site resolution. *Genomics*. 2011 Oct 1;98(4):288–95.
- 693 10. van Dongen J, Nivard MG, Willemsen G, Hottenga JJ, Helmer Q, Dolan CV, et al. Genetic and  
694 environmental influences interact with age and sex in shaping the human methylome. *Nature*  
695 *Communications*. 2016 Sep 1;7(1):11115.
- 696 11. van Iterson M, Tobi EW, Sliker RC, den Hollander W, Luijk R, Slagboom PE, et al. MethylAid:  
697 visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics*. 2014 Dec  
698 1;30(23):3435–7.
- 699 12. Sinke L, van Iterson M, Cats D, Sliker R, Heijmans B. DNAMArray: Streamlined workflow for the  
700 quality control, normalization, and analysis of Illumina methylation array data [Internet]. Zenodo;  
701 2019. Available from: <https://doi.org/10.5281/zenodo.3355292>
- 702 13. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of  
703 450k methylation array data improves replication in large cancer studies. *Genome Biology*. 2014  
704 Dec 3;15(11):503.

- 705 14. van Dongen J, Bonder MJ, Dekkers KF, Nivard MG, van Iterson M, Willemsen G, et al. DNA  
706 methylation signatures of educational attainment. *npj Science of Learning*. 2018 Mar 23;3(1):7.
- 707 15. Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, et al. Genomic and  
708 phenotypic insights from an atlas of genetic effects on DNA methylation. *Nature Genetics*. 2021  
709 Sep 1;53(9):1311–21.
- 710 16. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK:  
711 rising to the challenge of larger and richer datasets. *GigaScience*. 2015 Dec 1;4(1).
- 712 17. Saunders GRB, Wang X, Chen F, Jang SK, Liu M, Wang C, et al. Genetic diversity fuels gene  
713 discovery for tobacco and alcohol use. *Nature*. 2022 Dec 22;612(7941):720–4.
- 714 18. Vilhjálmsson J, Yang J, Finucane K, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage  
715 Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human  
716 Genetics*. 2015 Oct 1;97(4):576–92.
- 717 19. Joehanes R, Just AC, Marioni R, Pilling L, Reynolds L, Mandaviya PR, et al. Epigenetic Signatures  
718 of Cigarette Smoking. *Circulation Cardiovascular genetics*. 2016;9(5):436–47.
- 719 20. Neale MC, Hunter MD, Pritikin JN, Zahery M, Brick TR, Kirkpatrick RM, et al. OpenMx 2.0:  
720 Extended Structural Equation and Statistical Modeling. *Psychometrika*. 2016;81(2):535–49.
- 721 21. Verhulst B, Neale MC. Best Practices for Binary and Ordinal Data Analyses. *Behavior Genetics*.  
722 2021;51(3):204–14.
- 723 22. van Iterson M, van Zwet EW, Heijmans BT, the BIOS Consortium. Controlling bias and inflation in  
724 epigenome- and transcriptome-wide association studies using the empirical null distribution.  
725 *Genome Biology*. 2017 Jan 27;18(1):19.
- 726 23. Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, et al. GWASTools: an  
727 R/Bioconductor package for quality control and analysis of genome-wide association studies.  
728 *Bioinformatics*. 2012 Dec 1;28(24):3329–31.
- 729 24. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful  
730 Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*.  
731 1995 Jan 1;57(1):289–300.
- 732 25. Storey J, Bass A, Dabney A, Robinson D. qvalue: Q-value estimation for false discovery rate  
733 control. doi:10.18129/B9.bioc.qvalue [Internet]. 2023. Available from:  
734 <https://doi.org/10.18129/B9.bioc.qvalue>
- 735 26. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides  
736 a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*.  
737 2019 Apr 3;10(1).
- 738 27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for  
739 the unification of biology. *Nature Genetics*. 2000 May 1;25(1):25–9.
- 740 28. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*.  
741 2000 Jan 1;28(1):27–30.



- 742 29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set  
743 enrichment analysis: A knowledge-based approach for interpreting genome-wide expression  
744 profiles. *Proceedings of the National Academy of Sciences*. 2005 Oct 25;102(43):15545–50.
- 745 30. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome  
746 Pathway Knowledgebase. *Nucleic Acids Research*. 2018 Jan 4;46(D1):D649–55.
- 747 31. Breeze CE, Paul DS, van Dongen J, Butcher LM, Ambrose JC, Barrett JE, et al. eFORGE: A Tool  
748 for Identifying Cell Type-Specific Signal in Epigenomic Data. *Cell Reports*. 2016 Nov  
749 15;17(8):2137–50.
- 750 32. Breeze CE, Reynolds AP, van Dongen J, Dunham I, Lazar J, Neph S, et al. eFORGE v2.0: updated  
751 analysis of cell type-specific signal in epigenomic data. *Bioinformatics*. 2019 Nov 15;35(22):4767–  
752 9.
- 753 33. Breeze CE. Cell Type-Specific Signal Analysis in Epigenome-Wide Association Studies. In: Guan  
754 W, editor. *Epigenome-Wide Association Studies: Methods and Protocols* [Internet]. New York, NY:  
755 Springer US; 2022. p. 57–71. Available from: [https://doi.org/10.1007/978-1-0716-1994-0\\_5](https://doi.org/10.1007/978-1-0716-1994-0_5)
- 756 34. Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The Polygenic and  
757 Monogenic Basis of Blood Traits and Diseases. *Cell*. 2020 Sep 3;182(5):1214-1231.e11.
- 758
- 759