

Supplementary Information

Twin-based Mendelian Randomization Analyses Highlight Smoking's Effects on Blood DNA Methylation, with Putative Reverse Causation

Madhurbain Singh^{1,2,3*}, Conor V. Dolan^{3,4}, Dana M. Lapato^{1,2}, Jouke-Jan Hottenga^{3,4}, René Pool^{3,4}, Brad Verhulst⁵, Dorret I. Boomsma^{3,4,12}, Charles E. Breeze^{6,7}, Eco J. C. de Geus^{3,4}, Gibran Hemani⁸, Josine L. Min⁸, Roseann E. Peterson^{9,10,1}, Hermine H. M. Maes^{1,2}, Jenny van Dongen^{3,4,11*}, and Michael C. Neale^{1,2,3,11*}

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. 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. Address: Department of Biological Psychology, Vrije Universiteit Amsterdam, van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands

Michael C. Neale. Email: 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 **Supplementary Methods.....4**

43 **Genotypic Data, Principal Components Analysis, and Ancestry Outlier Pruning4**

44 **Smoking Assessment at Blood Sampling.....4**

45 **Polygenic Risk Score of Smoking4**

46 **Supplementary Figures 6**

47 **Figure S16**

48 **Figure S26**

49 **Figure S37**

50 **Figure S47**

51 **Figure S58**

52 **Figure S69**

53 **Figure S710**

54 **Figure S811**

55 **Figure S912**

56 **Figure S1013**

57 **Figure S1114**

58 **Figure S1216**

59 **Figure S1317**

60 **Figure S1418**

61 **Figure S1519**

62 **Figure S1620**

63 **Figure S1720**

64 **Figure S1821**

65 **Figure S1922**

66 **Figure S20**23

67 **Figure S21**24

68 **Figure S22**25

69 **Figure S23**26

70 **Figure S24**27

71 **Figure S25**28

72 **Figure S26**29

73 **Figure S27**30

74 **Figure S28**31

75 **Figure S29**32

76 **Figure S30**33

77 **Figure S31**34

78 **Figure S32**35

79 **Figure S33**36

80 **Figure S34**37

81 **Figure S35**38

82 ***References***.....39

83

84

85 **Supplementary Methods**

86 In this study, we analyzed data from the Netherlands Twin Register (NTR)¹ to examine the causal
87 influences between smoking status and blood DNA methylation (DNAm) using MR-DoC models^{2,3}. In
88 the current analyses, we included data from European-ancestry adult twins with both genotypic and
89 DNAm data, comprising 2,577 individuals (67% female).

91 **Genotypic Data, Principal Components Analysis, and Ancestry Outlier Pruning**

92 The DNA samples included in the current study were genotyped on 3 SNP (single nucleotide
93 polymorphism) microarray platforms: Affymetrix 6.0 (N= 2,399), Affymetrix Axiom (N= 83), and
94 Illumina GSA NTR array (N= 95). Genotype calling was done following the manufacturer's protocols.
95 Sample and variant quality control (QC), imputation, genetic principal component analysis (PCA), and
96 ancestry assignment have been previously described⁴. Briefly, after QC and harmonizing variants across
97 the three platforms, the data were aligned to the positive strand of *Genome Reference Consortium Human*
98 *Build 37* (GRCh37) and then imputed against the European (EUR) super-population of the 1000 Genomes
99 Project Phase-3 (KGP3)⁵, the Haplotype Reference Consortium (HRC)⁶ 1.1 (Ega version), and the
100 Genome of the Netherlands Consortium (GoNL)⁷ reference panels. Using SmartPCA in EIGENSTRAT⁸,
101 the first 20 PCs for the genotypic data were calculated in the KGP3 data, and the NTR samples were then
102 projected onto the PC space based on the SNP weights. Samples identified as outliers in the PC space
103 were then excluded.

105 **Smoking Assessment at Blood Sampling**

106 At the time of blood sampling for DNAm assessment, the participants answered the following
107 questions about smoking in Dutch. The English translations are described in the main **Methods** section.

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

108

109

110 **Polygenic Risk Score of Smoking**

111 The PRS of smoking was based on the European-ancestry summary statistics from the genome-
112 wide association study (GWAS) of smoking initiation (lifetime regular smoking) by GSCAN (GWAS &
113 Sequencing Consortium of Alcohol and Nicotine use)⁹, excluding the NTR from the meta-analysis.

114 As described in a previous study using the same PRS in the NTR⁴, the post-imputation SNPs
115 from the merged best-guess three-platform data were QCed to satisfy the following criteria: MAF >0.01,

116 HWE $p > 0.00001$, Mendel error rate $< 1\%$, and genotype call rate over 98%. Furthermore, the imputation
117 info for the three platforms needed to be above 0.10, and the allele frequency between platforms after
118 imputation could not differ more than 2%, leaving a total of 7,551,860 post-QC SNPs for analysis. The
119 PRS was calculated using *LDpred* v0.9¹⁰, with HRC+GoNL as the LD (linkage disequilibrium) reference
120 panel. For estimating the target LD structure, we used a subset of unrelated individuals and a set of well-
121 imputed variants in the NTR. The parameter `ld_radius` was set by dividing the number of variants in
122 common (from the output of the coordination step) by 12000. For the coordination step, the median
123 sample size was used as the input value for N. For the LDpred step, we applied the following thresholds
124 for the fraction of variants with non-zero effects (in addition to the default infinitesimal model): --
125 `PS=0.5, 0.3, 0.2, 0.1, 0.05, 0.01`.

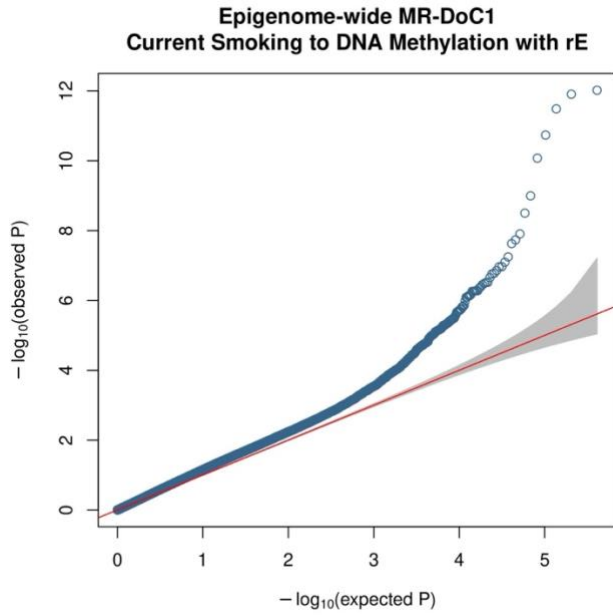
126 To determine the LDpred threshold that yielded the PRS with the highest predictive power for the
127 variables of interest (current vs. never and former vs. never smoking), we fitted logistic regression models
128 in R (v4.3.2) to estimate incremental R^2 on a liability scale. We first fitted a null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

139
140

141 **Supplementary Figures**

142 **Figure S1**

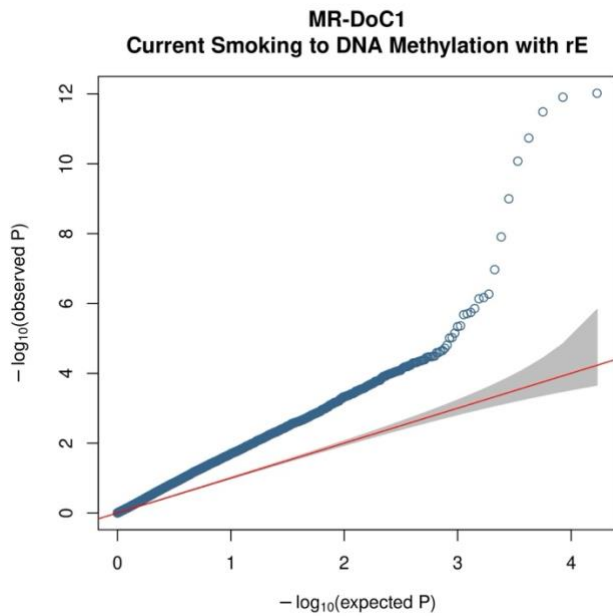
143 *QQ Plot of MR-DoC1 models (with unique environmental confounding, rE) of Current Smoking →*
144 *DNAm at 411,169 epigenome-wide CpGs (Bayesian genomic inflation factor, $\lambda = 1.09$).*



145

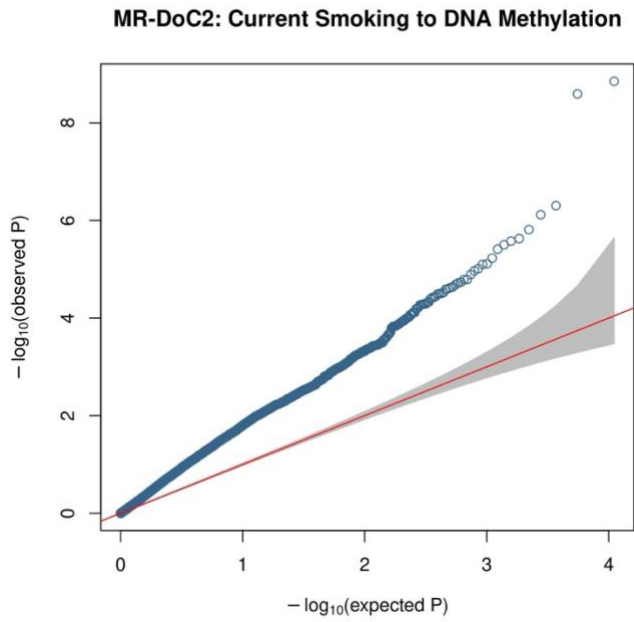
146 **Figure S2**

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

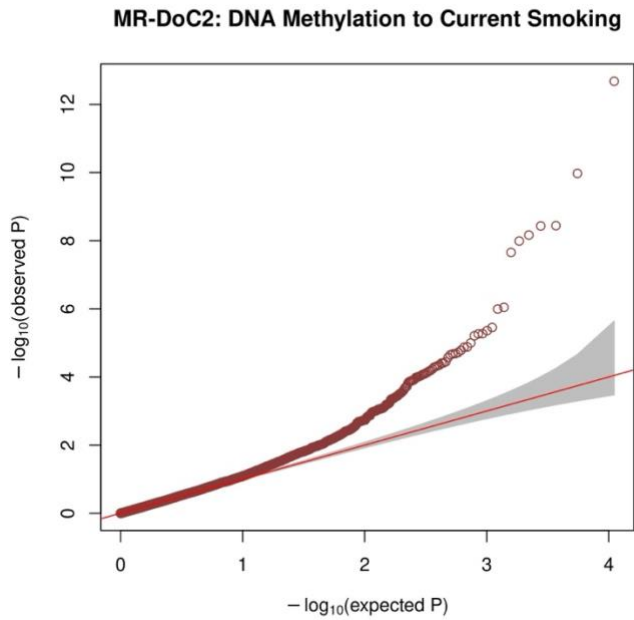


149

150 **Figure S3**
151 *QQ Plot of the Current Smoking → DNAm causal estimates in MR-DoC2 models across 11,124 smoking-*
152 *associated CpGs (Bayesian genomic inflation factor, $\lambda = 1.20$).*



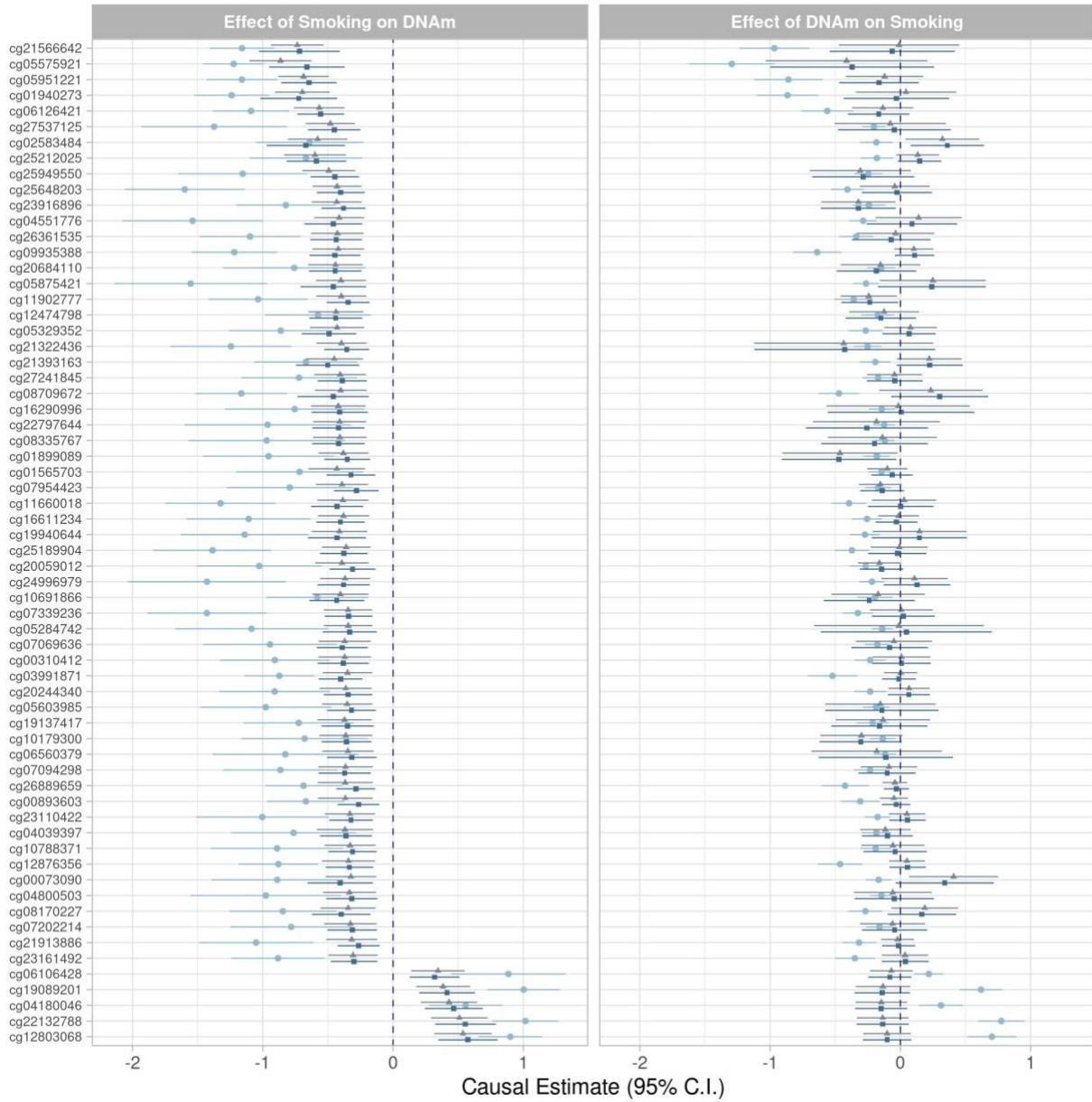
153
154 **Figure S4**
155 *QQ Plot of the DNAm → Current Smoking causal estimates in MR-DoC2 models across 11,124 smoking-*
156 *associated CpGs (Bayesian genomic inflation factor, $\lambda = 1.01$).*



157
158

159 **Figure S5**
 160 *Bidirectional Causal Estimates at the 64 CpGs with Robust Evidence of the Causal Effects of Current*
 161 *Smoking on DNA methylation*

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

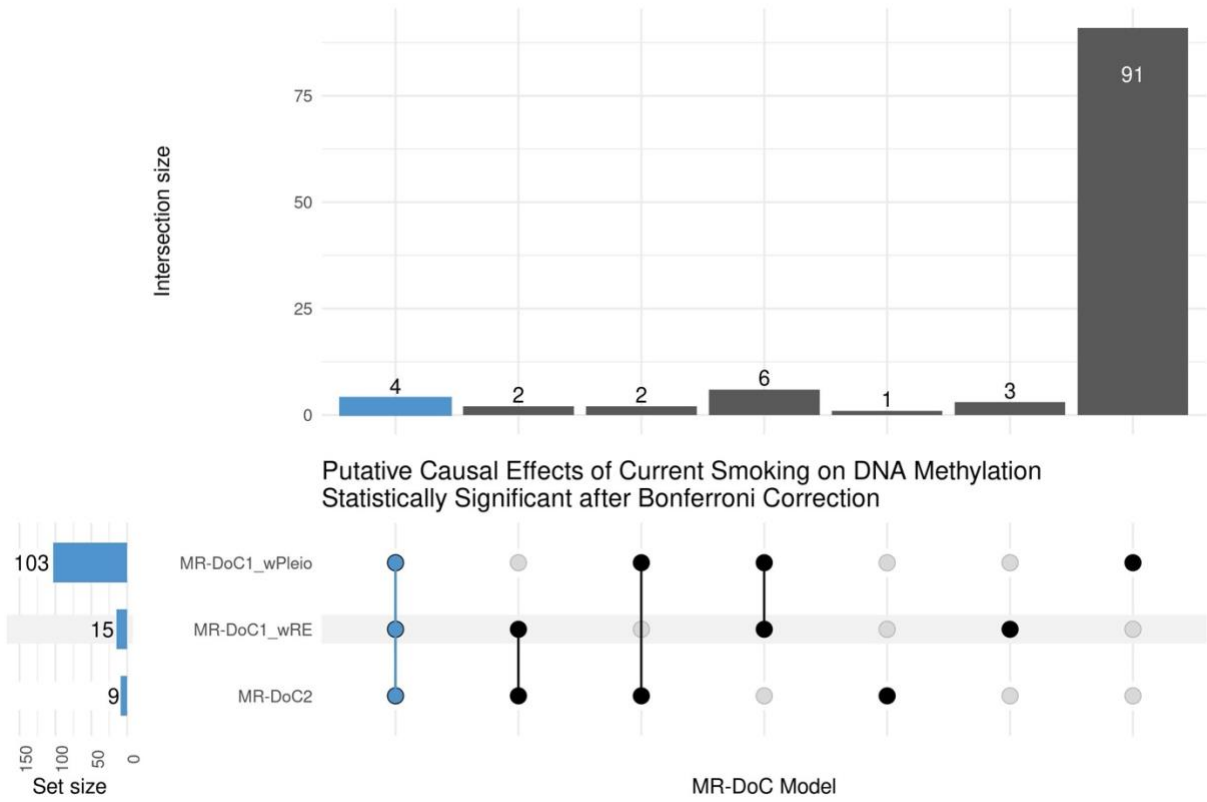


Model ■ MR-DoC2 ● MR-DoC1 w/ Pleiotropic Path ▲ MR-DoC1 w/ rE

162
 163 These CpGs did not show robust evidence for the reverse effects of DNAm on current smoking. Please
 164 refer to **Supplementary Tables S1** (*Current Smoking → DNAm*) and **S2** (*DNAm → Current Smoking*) for
 165 the corresponding data.

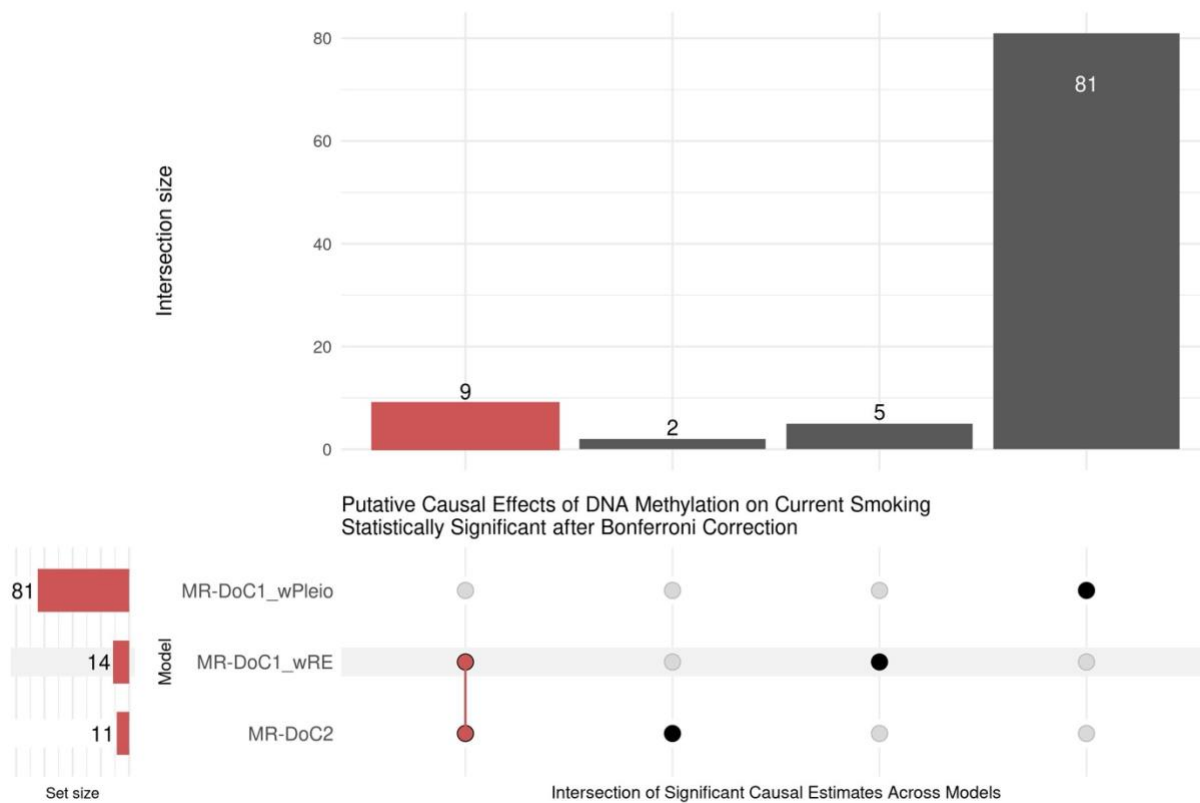
166
 167

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



171
 172
 173 *Note.* Please refer to **Supplementary Table S1** for the corresponding data
 174
 175

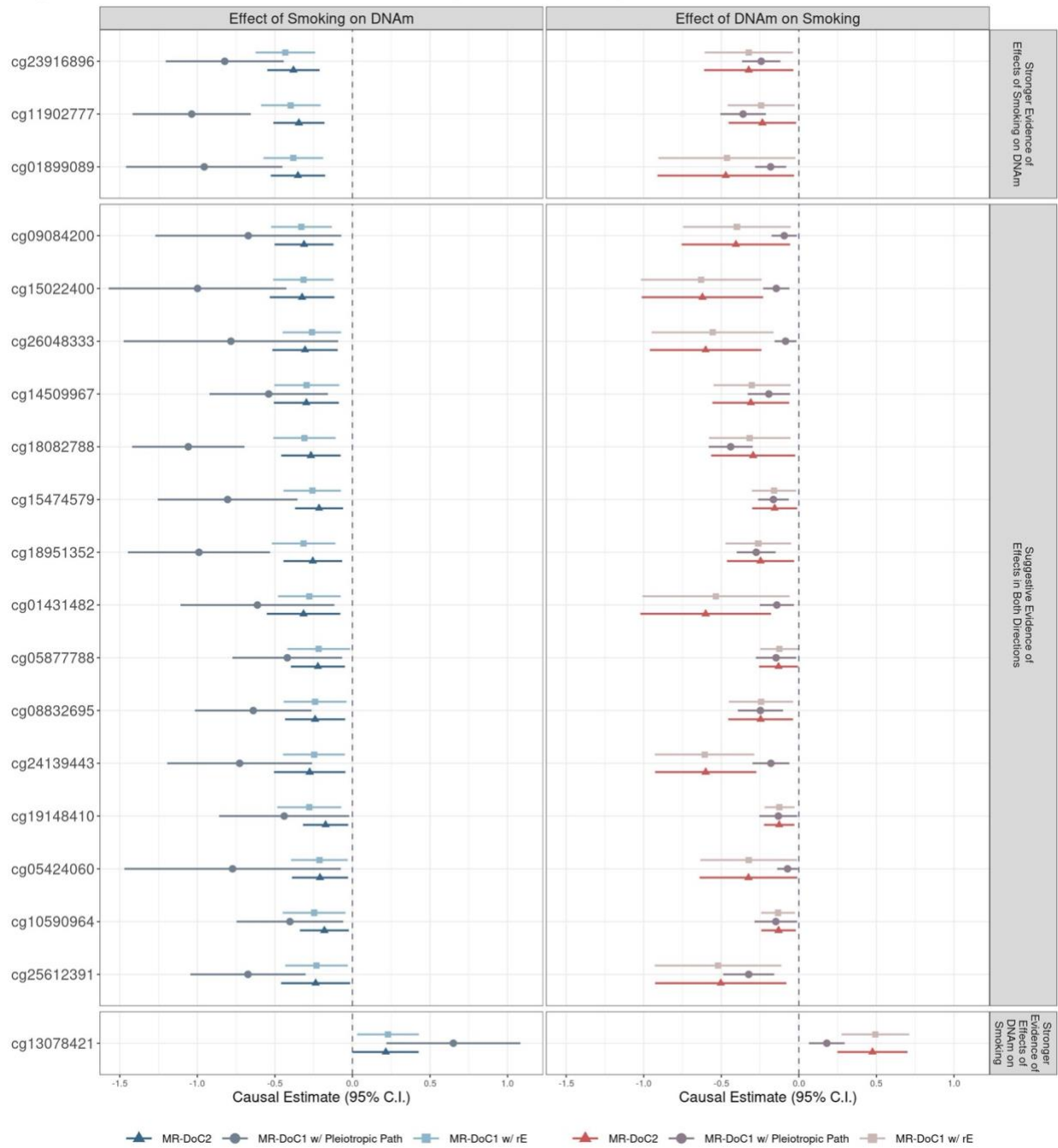
176 **Figure S7**
 177 *Upset plot of the intersection of CpGs with statistically significant DNAm → Current Smoking effects*
 178 *after Bonferroni correction in each of the three MR-DoC models*
 179



180
 181
 182 *Note.* Please refer to **Supplementary Table S3** for the corresponding data
 183
 184
 185

186
187

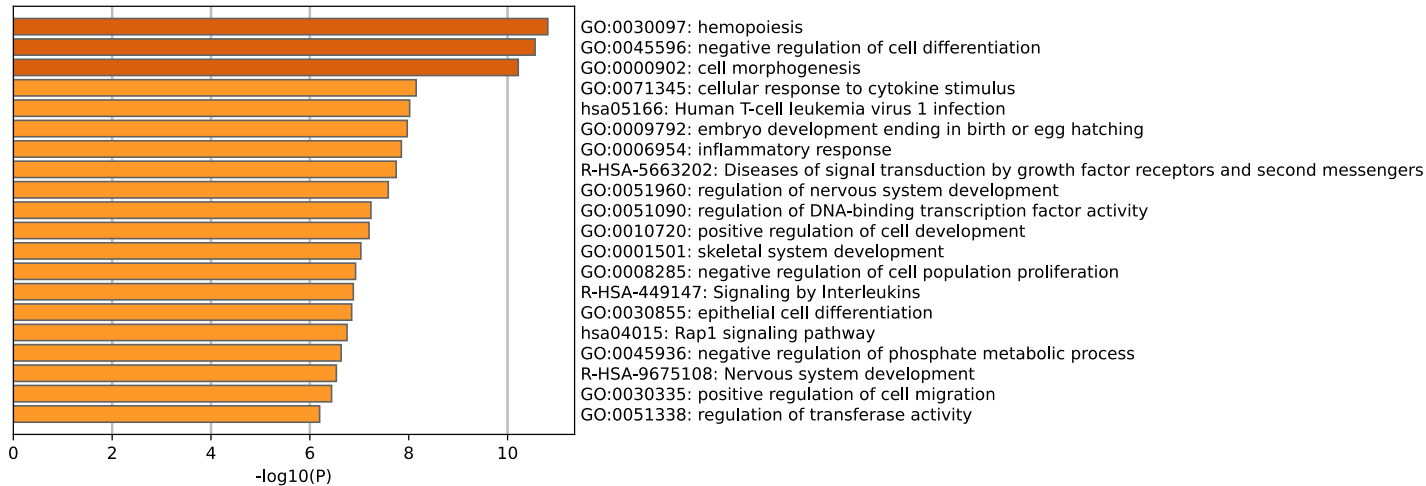
Figure S8
19 CpGs with potential bidirectional causal effects between current smoking and DNA methylation
Suggestive Bidirectional Causal Effects between Smoking and DNAm at 19 CpGs



188
189
190
191
192

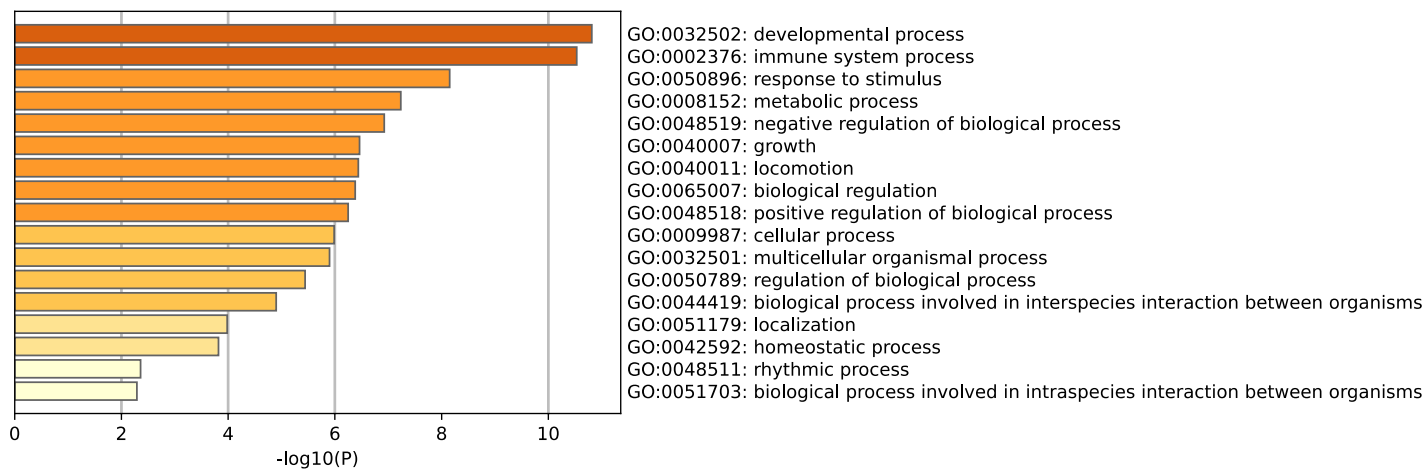
Note. Three CpGs had more robust evidence of *Current Smoking* \rightarrow *DNAm* causal effects than *vice versa*. One CpG had more robust evidence of *DNAm* \rightarrow *Current Smoking* causal effects than *vice versa*. The rest 15 CpGs had only suggestive evidence (consistent, nominally significant estimates across models) in both directions. Please refer to **Supplementary Tables S1-S4** for the corresponding data.

193 **Figure S9**
 194 *Top Enriched Ontology Clusters in Metascape’s Gene Annotation and Functional Enrichment Analyses*
 195 *of the 525 CpGs (outside the MHC region) with Potential Current Smoking → DNAm effects*
 196



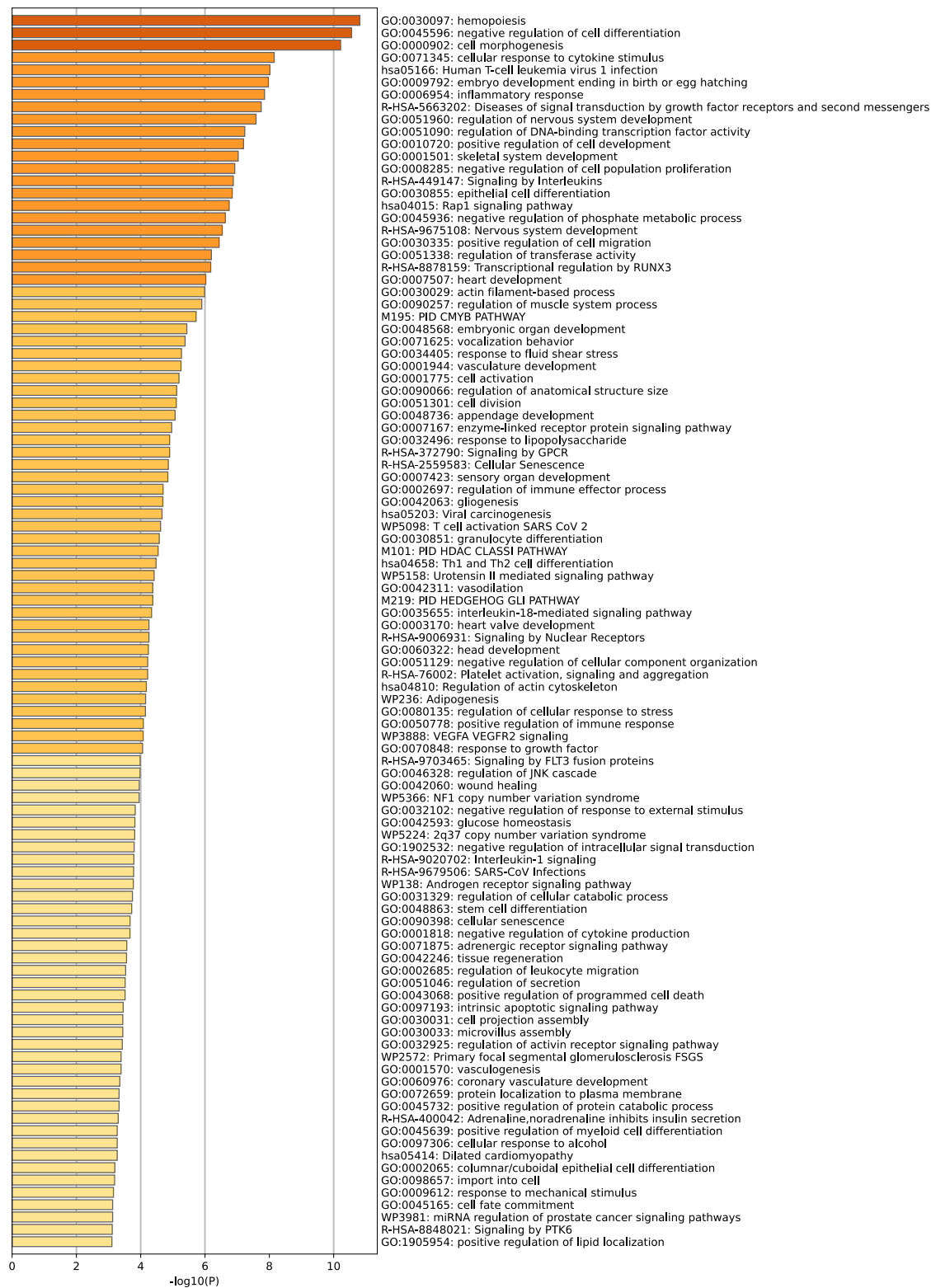
197
 198
 199 *Note.* The “NearestGene” IDs from *Supplementary Table S1* were used as the input data for Metascape¹².
 200 Please refer to **Supplementary Tables S5-S6** for the corresponding annotation and enrichment results.
 201 As detailed in the Metascape manuscript¹², the program first identified all significant ontology terms,
 202 including GO/KEGG terms, canonical pathways, and hallmark gene sets. The significant terms (based on
 203 hypergeometric p-value <0.01 and >1.5-fold enrichment) were then clustered into a hierarchical tree
 204 based on Kappa-statistical similarities among their gene memberships. The tree was then cast into term
 205 clusters based on a threshold of 0.3 kappa score. The enrichment clusters and their underlying terms are
 206 marked as “Summary” and “Membership”, respectively, under the column *GroupID* in *Supplementary*
 207 *Table S24*. The “Summary” terms provide an overview of enriched, non-redundant ontology terms.
 208
 209

210 **Figure S10**
 211 *Enrichment Results for Gene-Ontology (GO) Processes in Metascape’s Gene Annotation and Functional*
 212 *Enrichment Analyses of the 525 CpGs (outside the MHC region) with Potential Current Smoking →*
 213 *DNA_m effects*



214
 215
 216 *Note.* The “NearestGene” IDs from *Supplementary Table S1* were used as the input data for Metascape¹².
 217 Please refer to **Supplementary Table S6** for all enrichment results.
 218
 219
 220

221 **Figure S11**
 222 *Top 100 Ontology Terms in Metascape's Gene Annotation and Functional Enrichment Analyses of the*
 223 *525 CpGs (outside the MHC region) with Potential Current Smoking → DNAm effects*



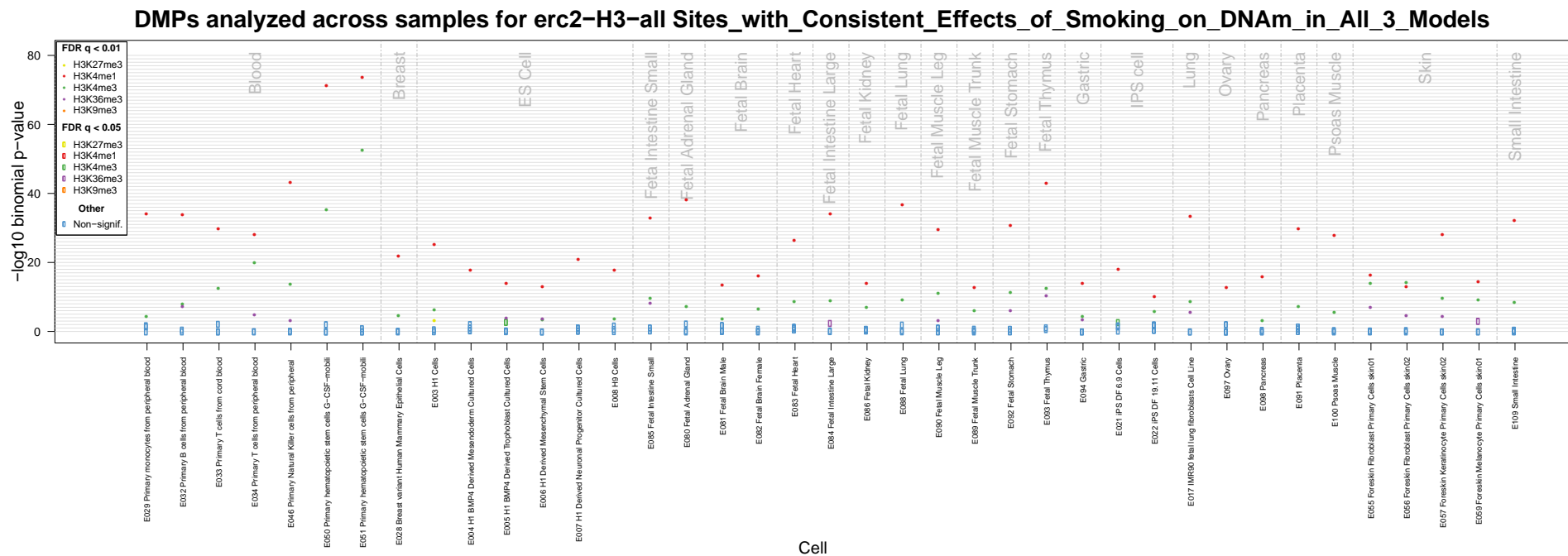
225 *Note.* The “NearestGene” IDs from *Supplementary Table S1* were used as the input data for Metascape¹².
226 Please refer to **Supplementary Table S6** for all enrichment results.
227

236 **Figure S13**

237 *eFORGE analyses of overlap between histone-mark modifications and the 525 CpGs (outside the MHC region) with potential Current Smoking →*

238 *DNAm effects*

239



240

241

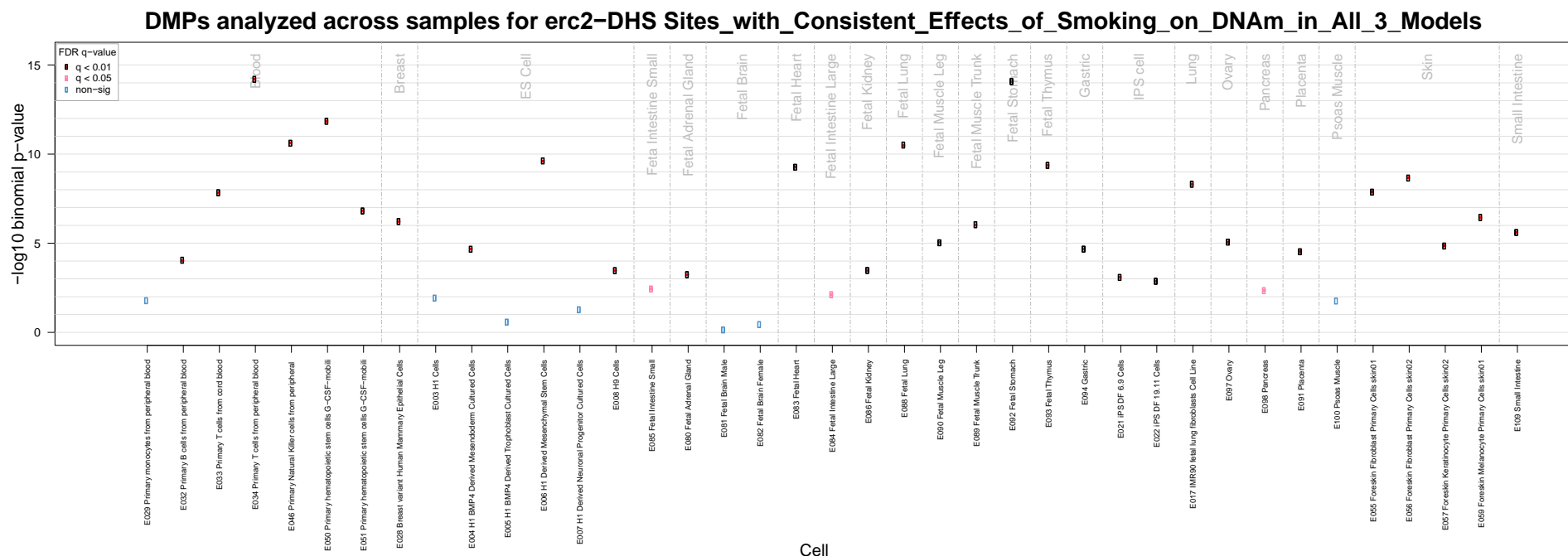
242 *Note.* Please refer to **Supplementary Table S8** for the corresponding data.

243

244 **Figure S14**

245 *eFORGE* analyses of overlap between DNase hypersensitivity (DHS) sites and the 546 CpGs with potential Current Smoking → DNAm effects

246



247

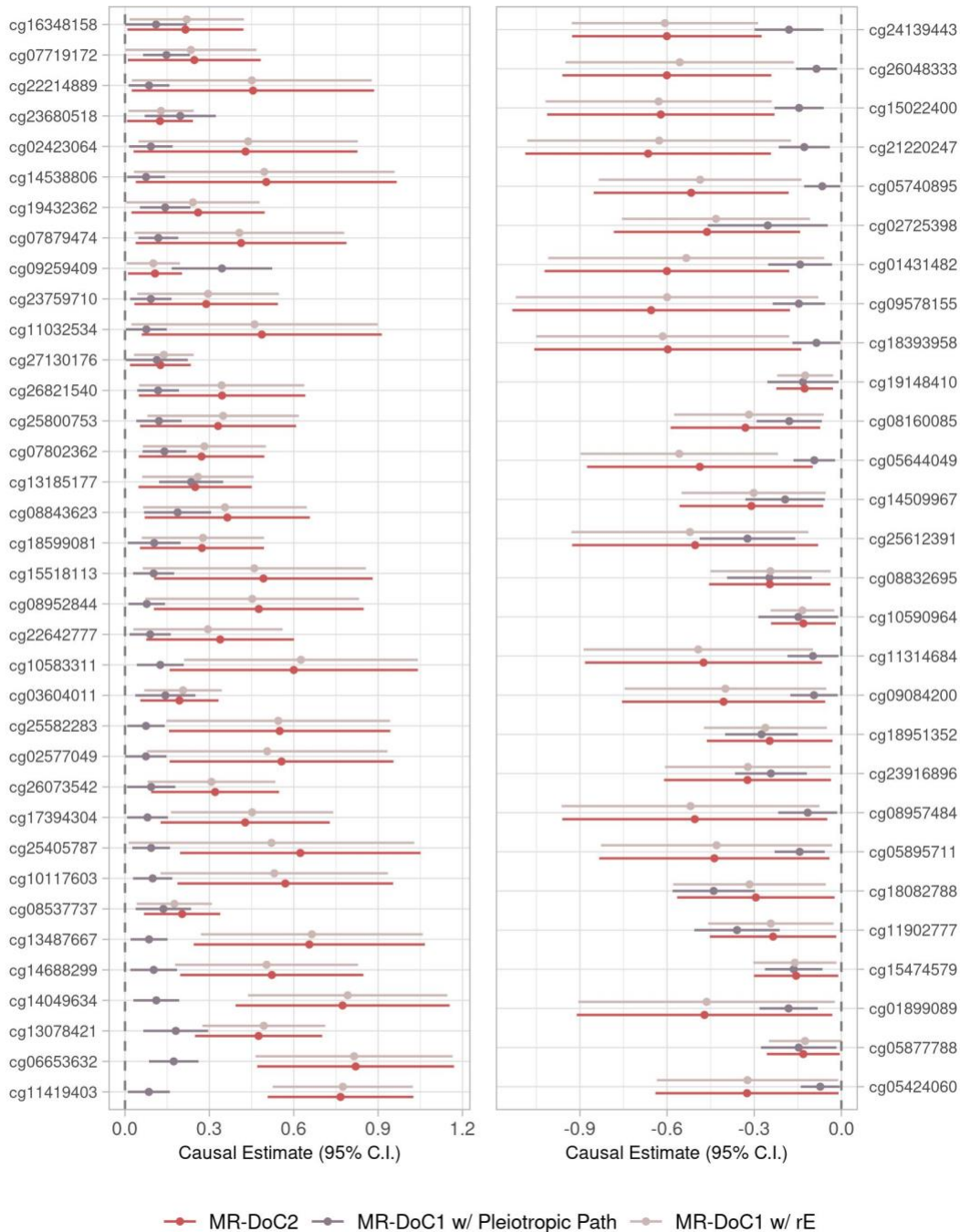
248

249 *Note.* Please refer to **Supplementary Table S9** for the corresponding data.

250

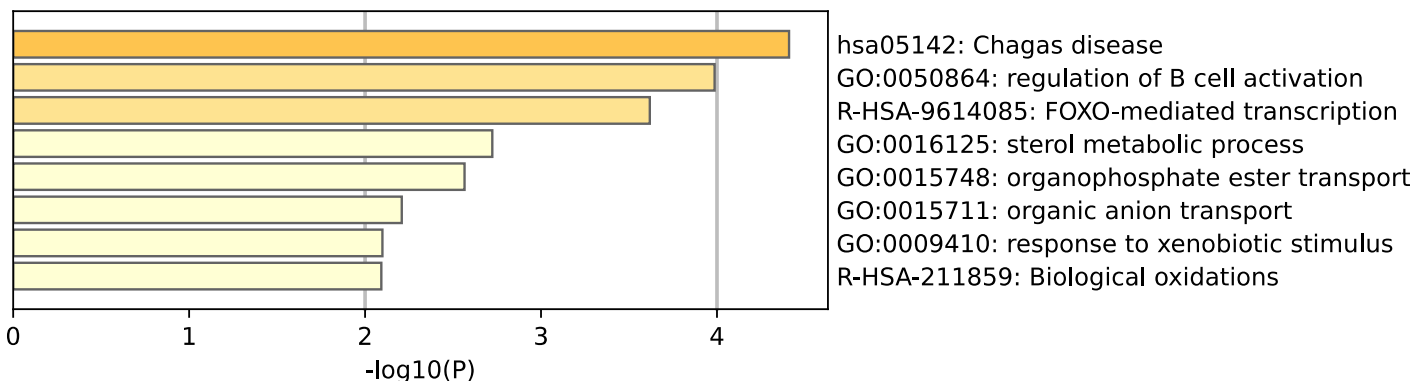
251 **Figure S15**
 252 *64 CpGs with potential DNAm → Current Smoking effects, based on consistent, nominally significant*
 253 *estimates across models*

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



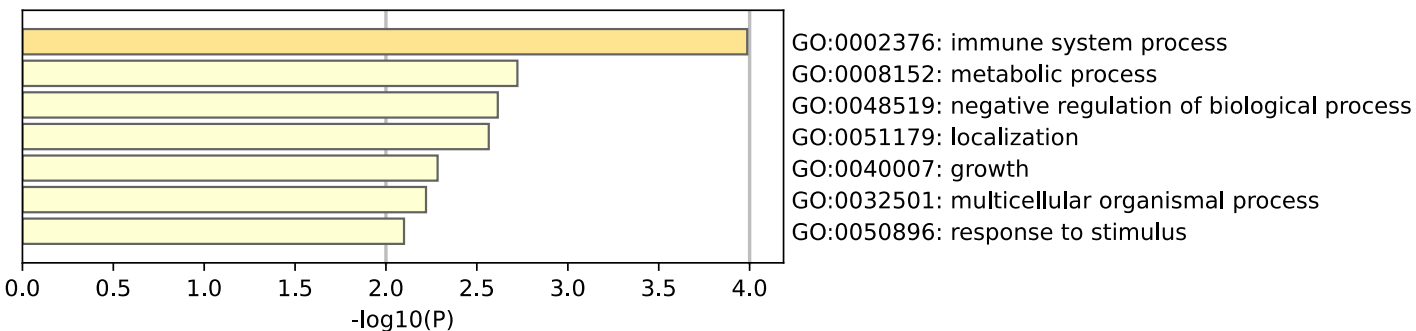
254
 255 *Note.* These CpGs were used for the follow-up enrichment analyses with eFORGE¹¹ and Metascape¹².
 256 None of these sites are in the MHC region. Please refer to **Supplementary Table S3** for the
 257 corresponding data.

258 **Figure S16**
 259 *Top Ontology Clusters in Metascape’s Gene Annotation and Functional Enrichment Analyses of the 64*
 260 *CpGs with Potential DNAm → Current Smoking effects*
 261



262
 263 *Note.* The “NearestGene” IDs from *Supplementary Table S3* were used as the input data for Metascape¹².
 264 None of the ontology terms were significant after multiple-testing correction. Please refer to
 265 **Supplementary Tables S10 and S11** for all annotation and enrichment results.
 266

267 **Figure S17**
 268 *Enrichment Results for Gene-Ontology (GO) Processes in Metascape’s Gene Annotation and Functional*
 269 *Enrichment Analyses of the 64 CpGs with Potential DNAm → Current Smoking effects*
 270



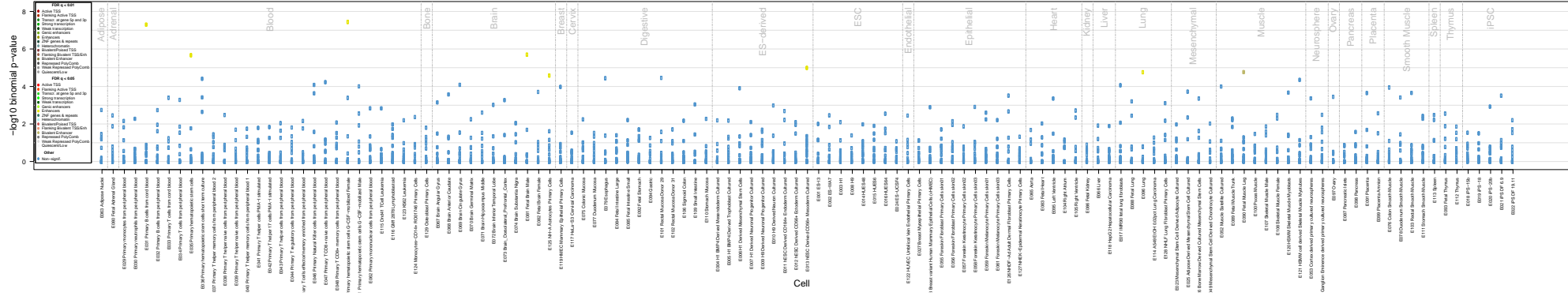
271
 272 *Note.* The “NearestGene” IDs from *Supplementary Table S3* were used as the input data for Metascape¹².
 273 None of the ontology terms were significant after multiple-testing correction. Please refer to
 274 **Supplementary Table S11** for all enrichment results.
 275
 276

277
278

Figure S18

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

DMPs analyzed across samples for *erc2*-chromatin15state-all Sites_with_Consistent_Effects_of_DNAm_on_Smoking_in_All_3_Models



279
280
281

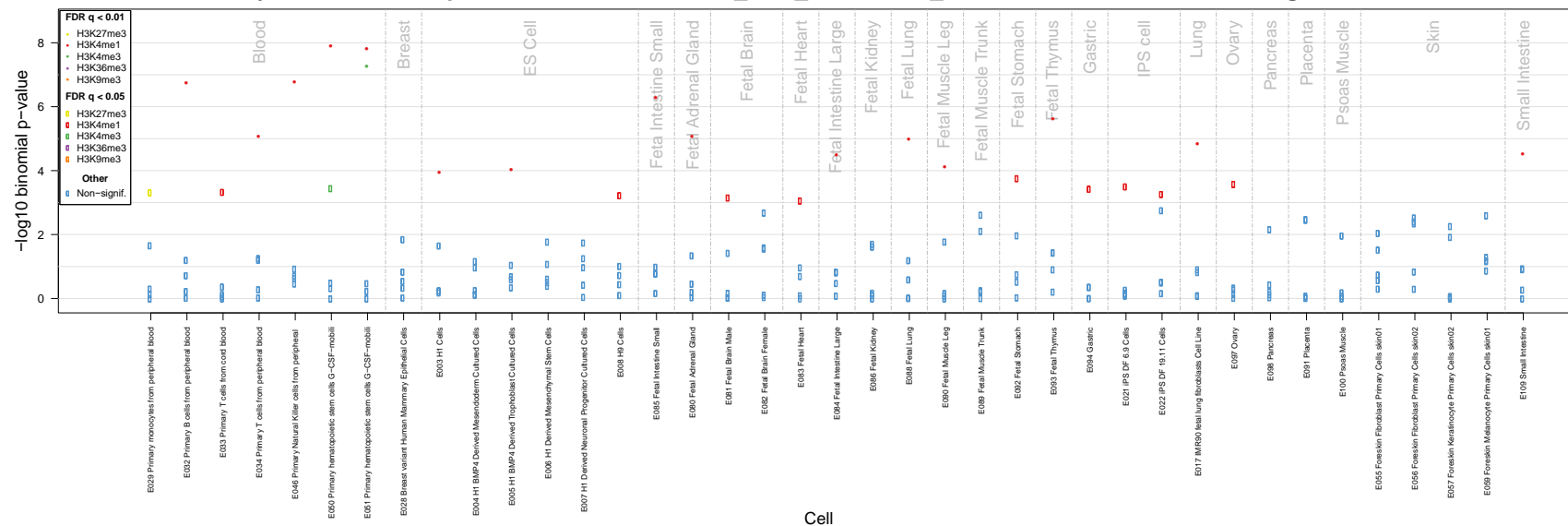
Note. Please refer to **Supplementary Table S12** for the corresponding data.

282
283

Figure S19

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



284
285
286
287

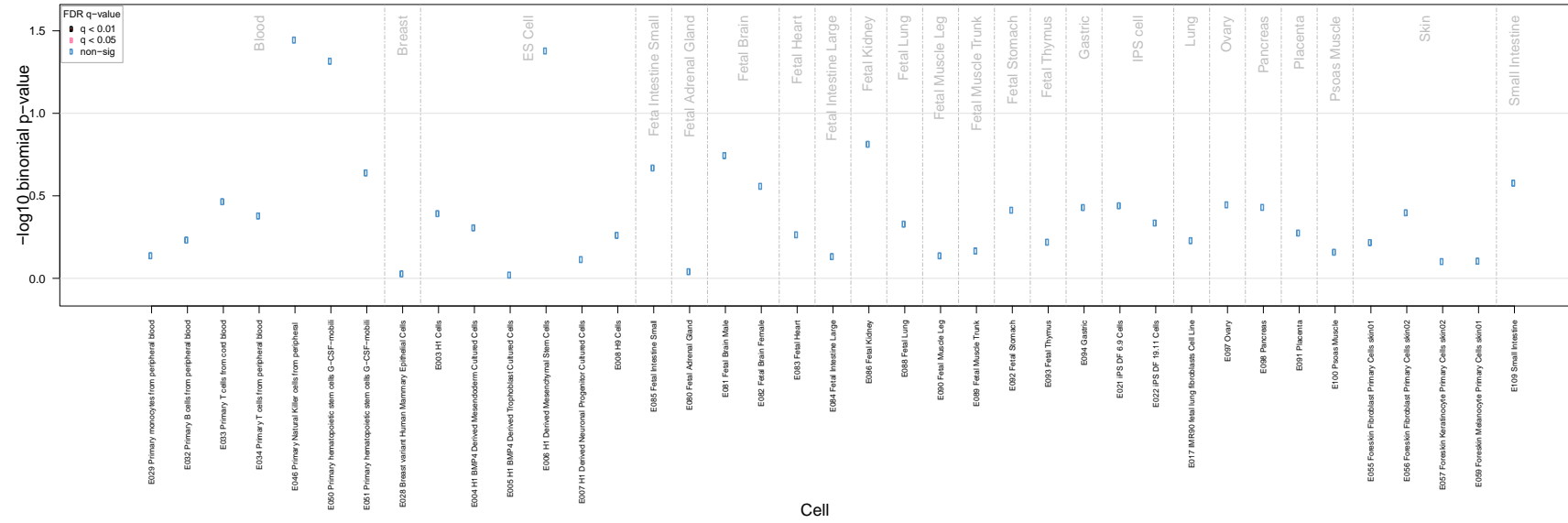
Note. Please refer to **Supplementary Table S13** for the corresponding data.

288
289

Figure S20

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



290
291
292
293
294

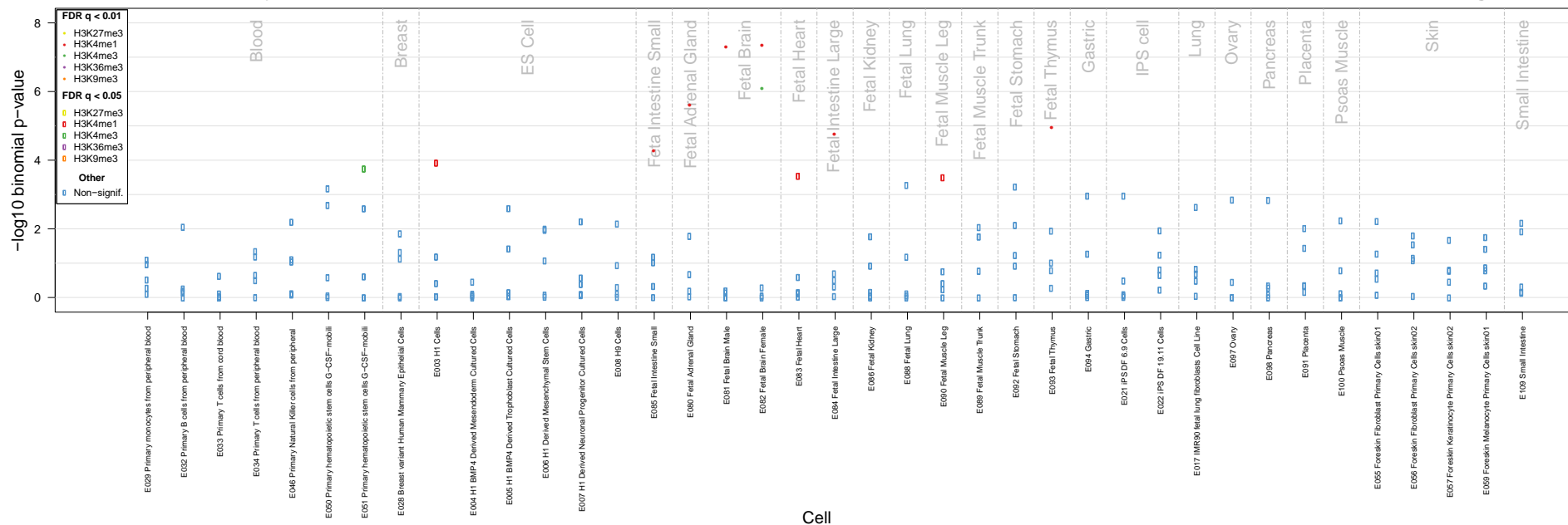
Note. Please refer to **Supplementary Table S14** for the corresponding data.

303
304
305

Figure S22

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 S10/Table S5

DMPs analyzed across samples for erc2-H3-all CpGs_Enriched_for_Fetal_Brain_Enhancers_and_Effects_on_Smoking



306
307
308
309
310

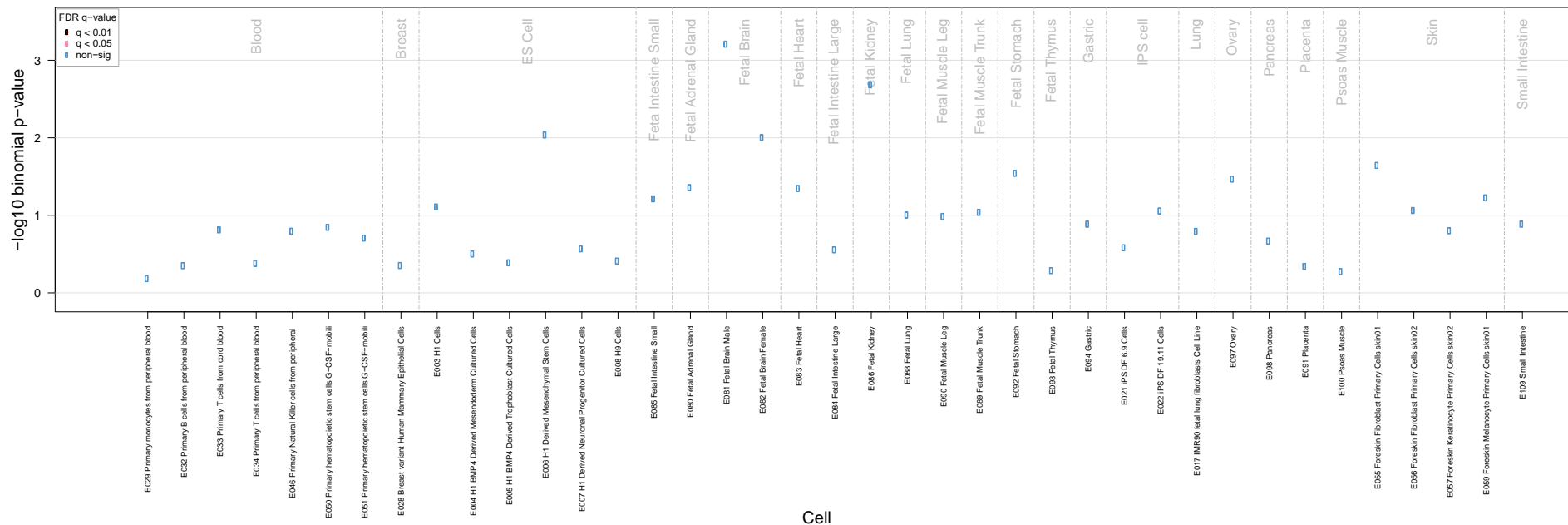
Note. Please refer to **Supplementary Table S16** for the corresponding data.

311
312
313

Figure S23

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 S10/Table S5

DMPs analyzed across samples for erc2-DHS CpGs_Enriched_for_Fetal_Brain_Enhancers_and_Effects_on_Smoking



314
315
316
317
318

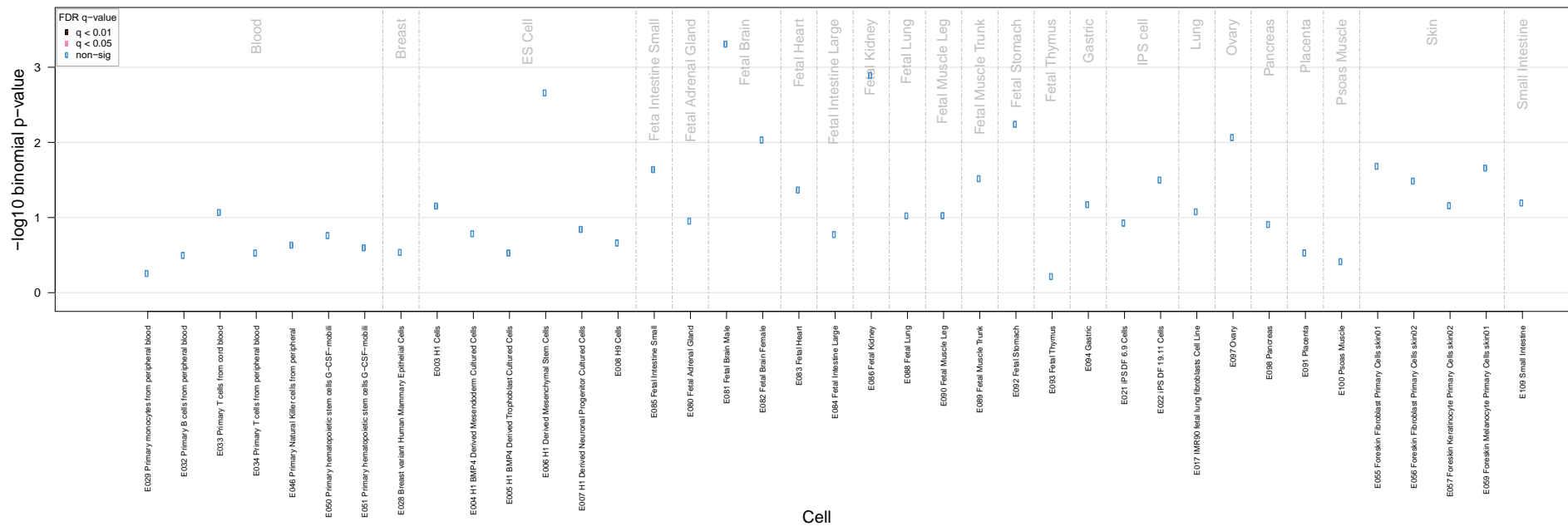
Note. Please refer to **Supplementary Table S17** for the corresponding data.

319
320
321

Figure S24

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 S14/Table S9

DMPs analyzed across samples for erc2–DHS CpGs_Enriched_for_Fetal_Brain_H3K4me3_and_Effects_on_Smoking



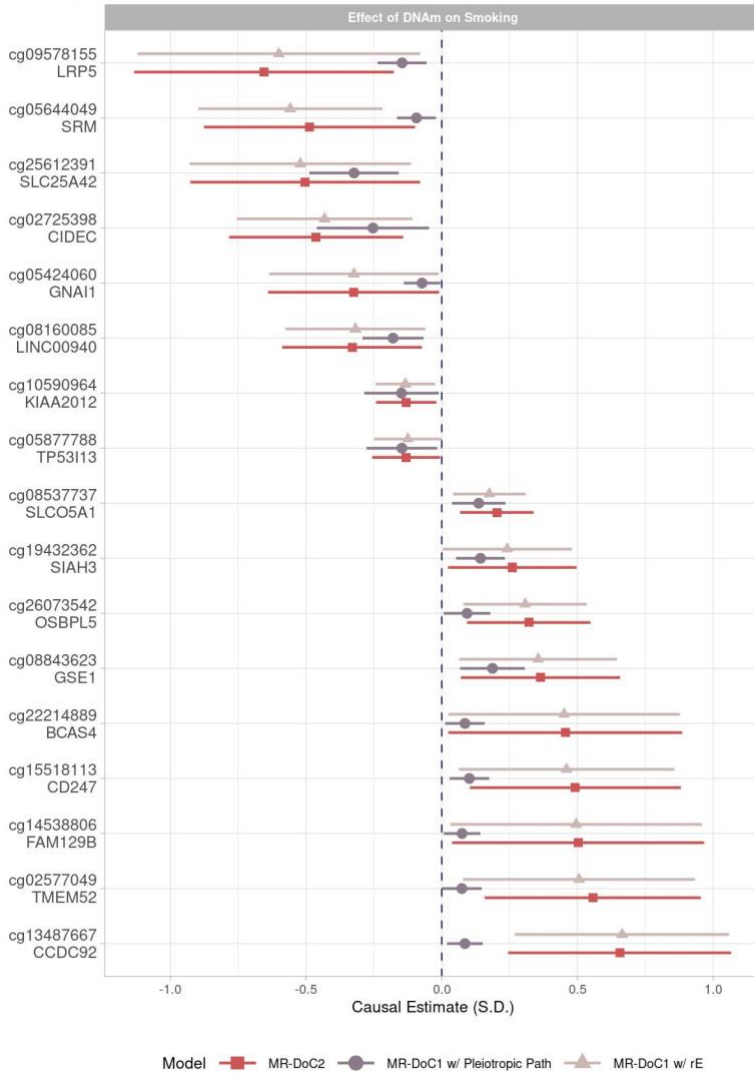
322
323
324

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

Figure S25

Estimated DNAm → Current Smoking effects at the 17 CpGs showing highly specific enrichment for overlap with gene-regulatory elements in the brain in Figure 7

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

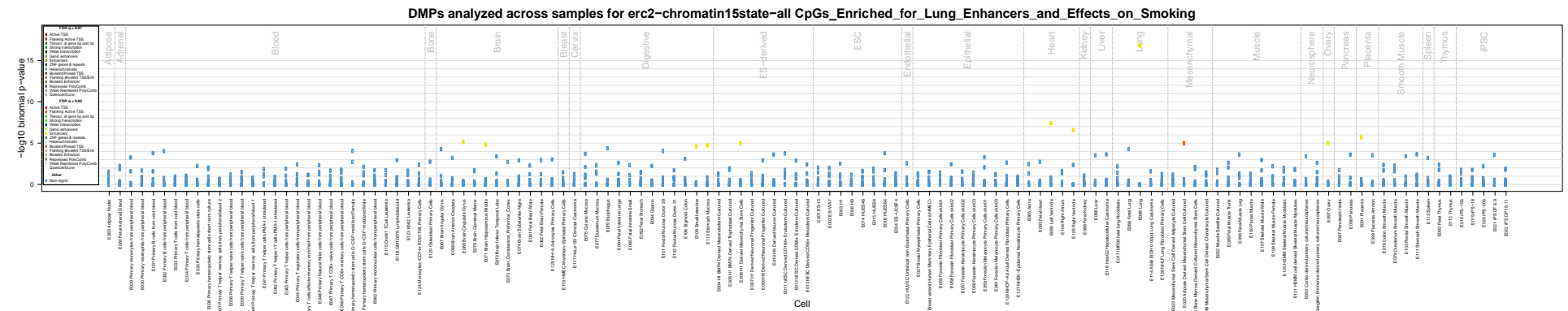


Note. The Y-axis shows the probe ID and the “Nearest Gene”. For the corresponding data, please refer to **Supplementary Table S3**.

333
334
335

Figure S26

*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 S10/Table S5



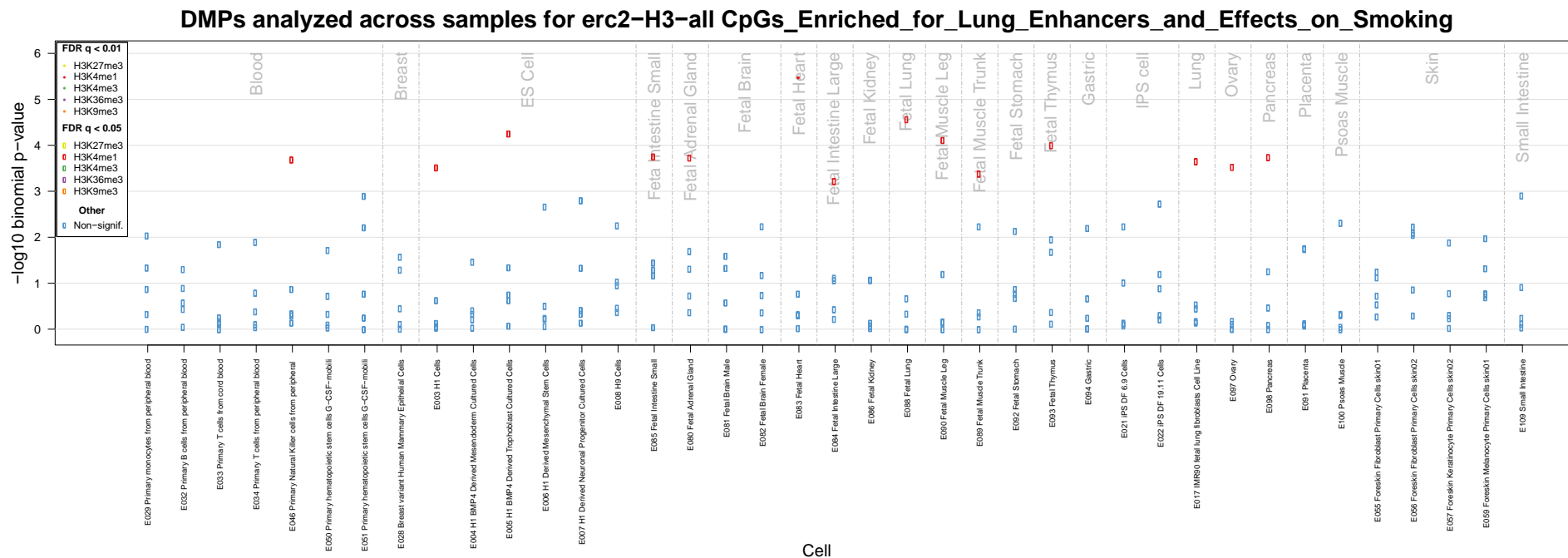
336
337
338
339
340

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

341
342
343
344

Figure S27

eFORGE analyses of overlap between histone-mark modifications and the 18 CpGs underlying the enriched overlap with Enhancers in the “Lung” sample in Figure S10/Table S5



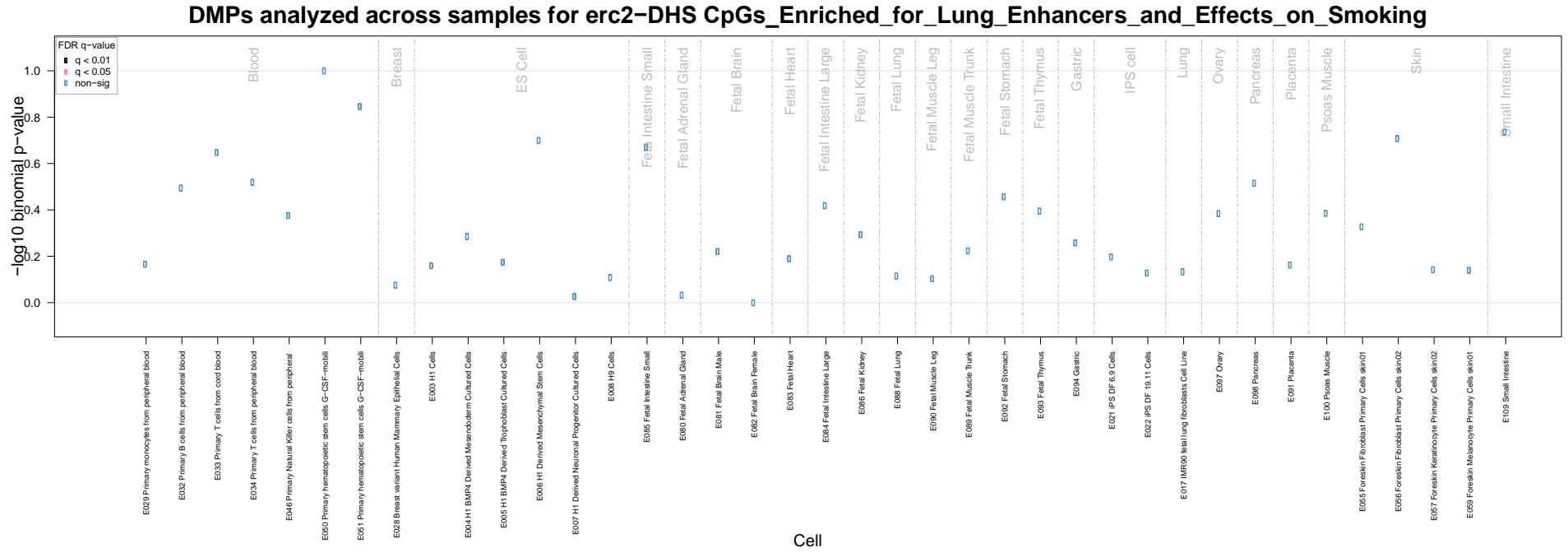
345
346
347

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

348
349
350
351

Figure S28

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 S10/Table S5



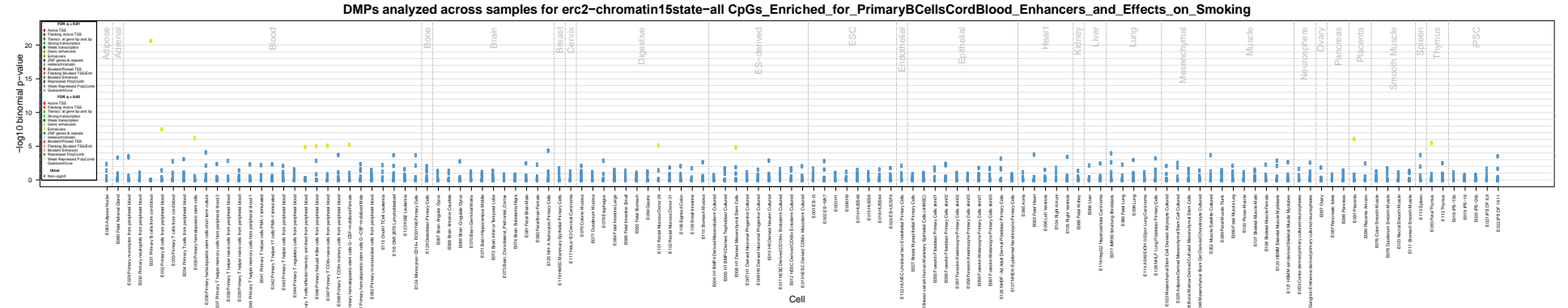
352
353
354
355
356
357

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

358
359
360

Figure S29

*e*FORGE 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 S10/Table S5



361
362
363
364
365

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

366

367

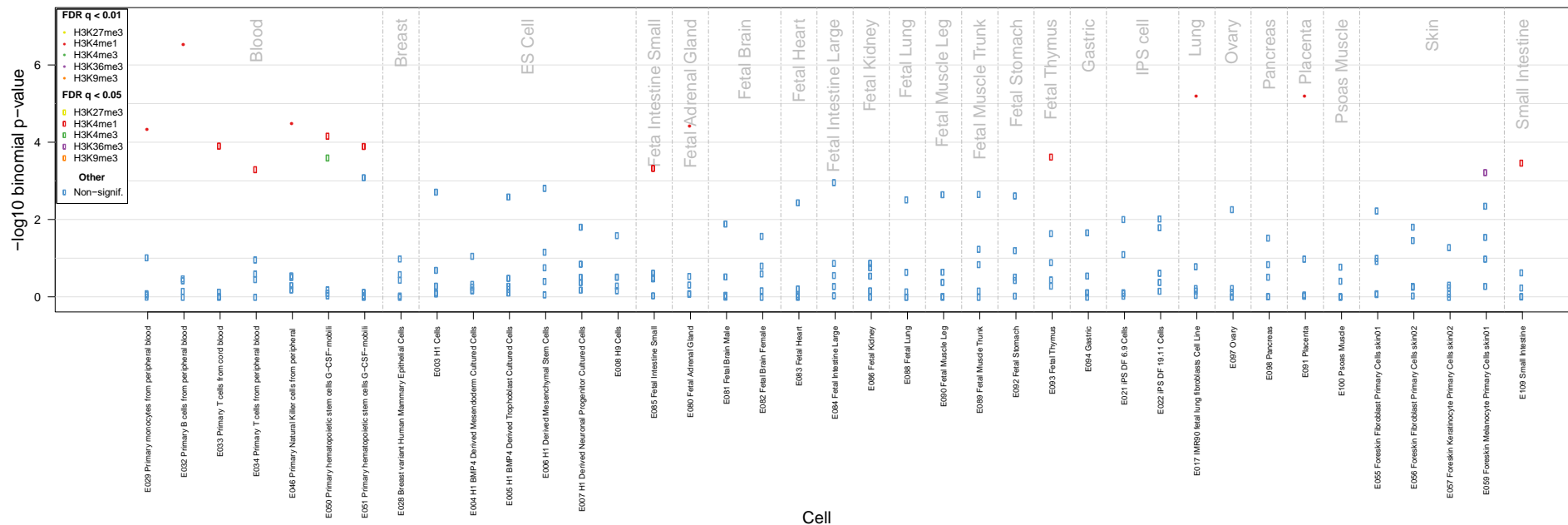
Figure S30

368

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 S10/Table S5

369

DMPs analyzed across samples for erc2-H3-all CpGs_Enriched_for_PrimaryBCellsCordBlood_Enhancers_and_Effects_on_Smoking



370

371

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

372

373

374

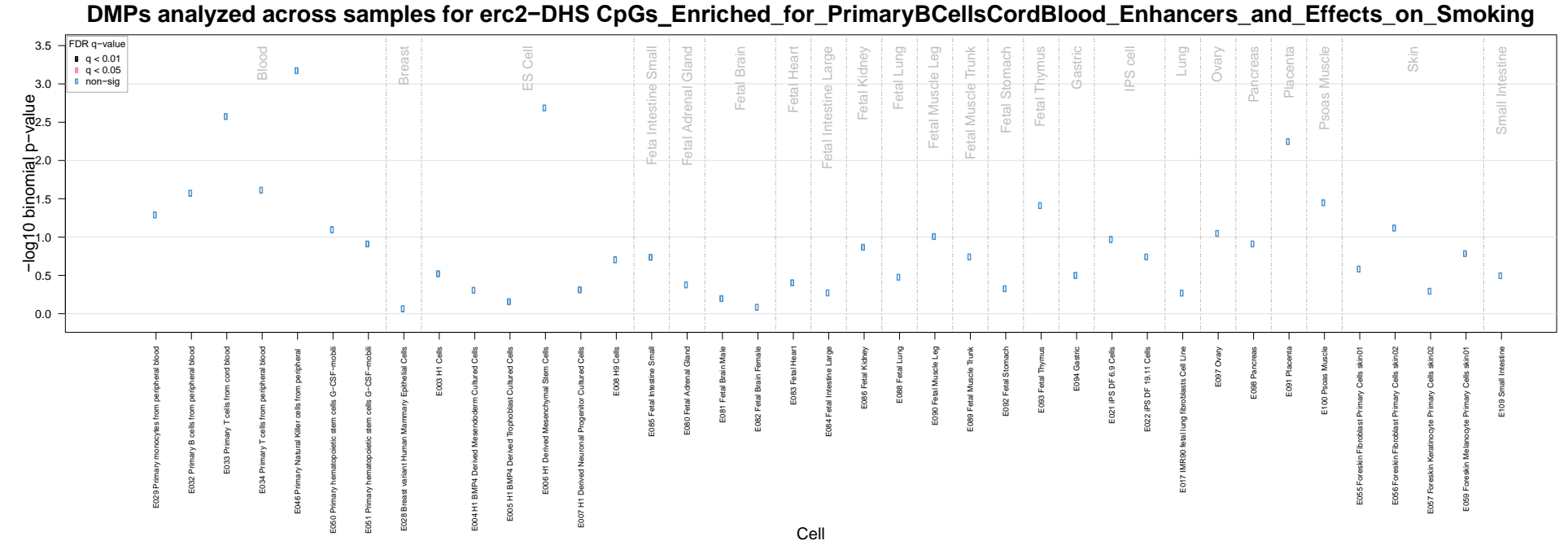
375

376

377
378
379
380

Figure S31

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 S10/Table S5

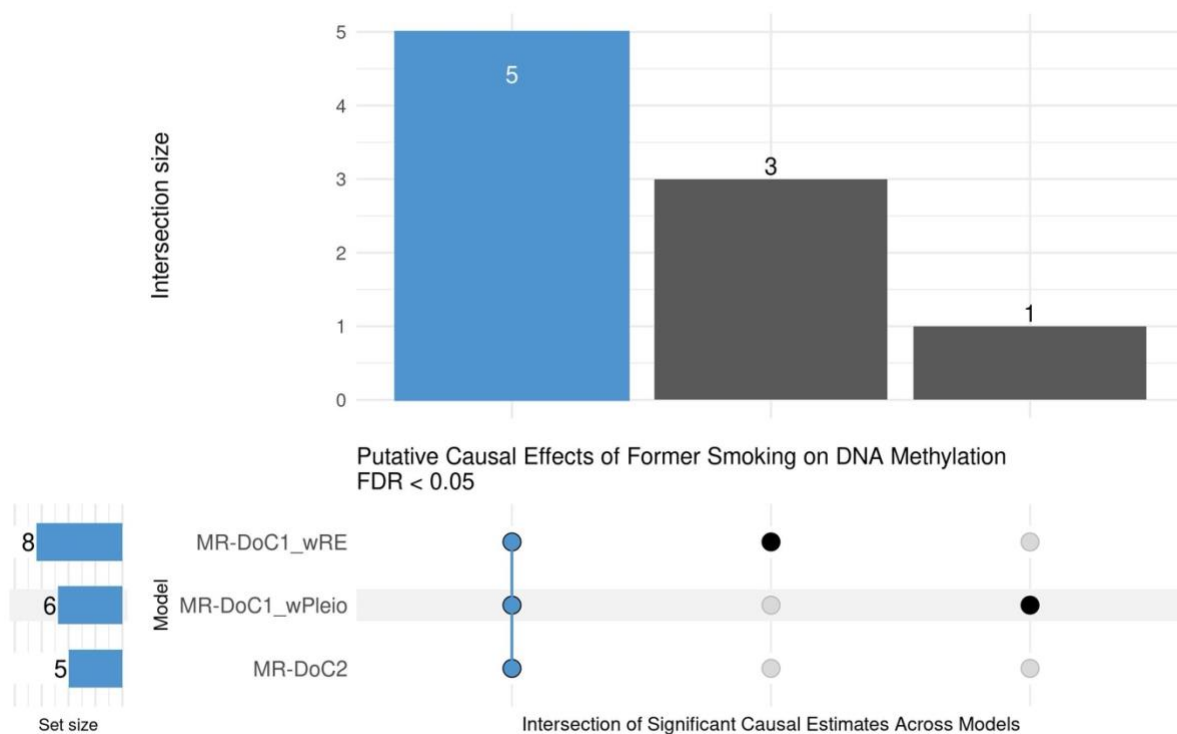


381
382
383
384
385

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

386 **Figure S32**

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



389
390

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

392
393
394

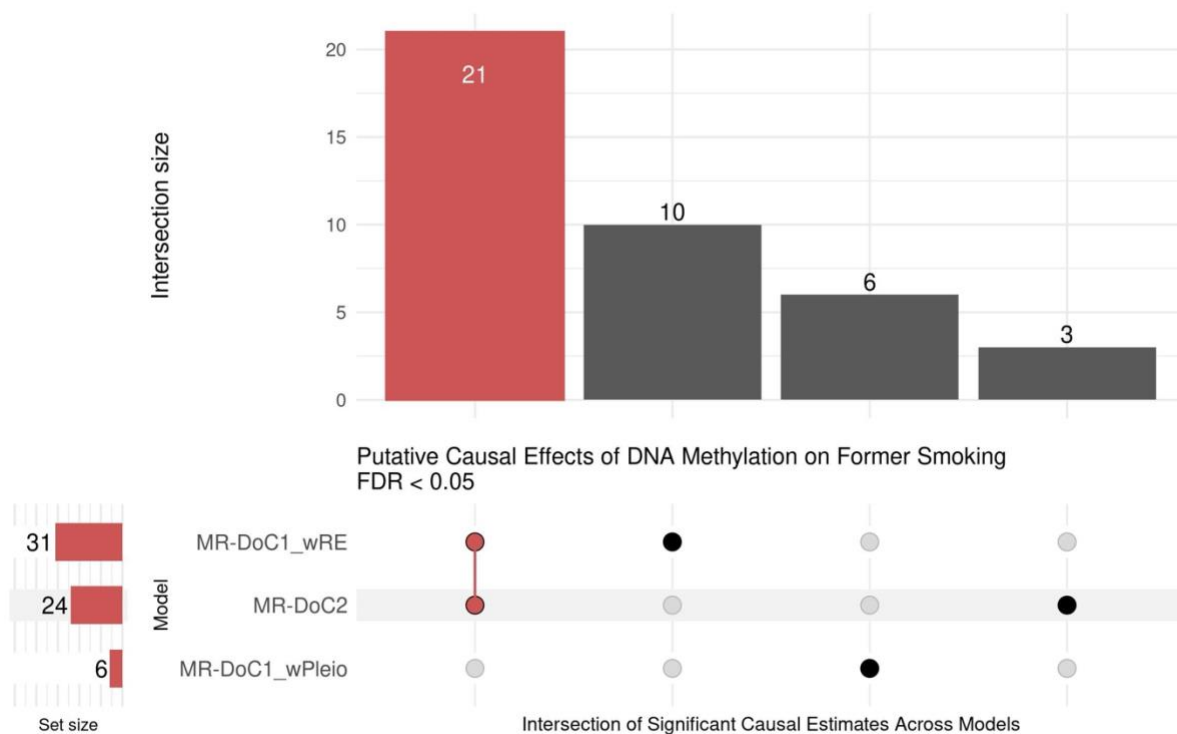
395

Figure S33

396

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

397



398

399

400

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

401

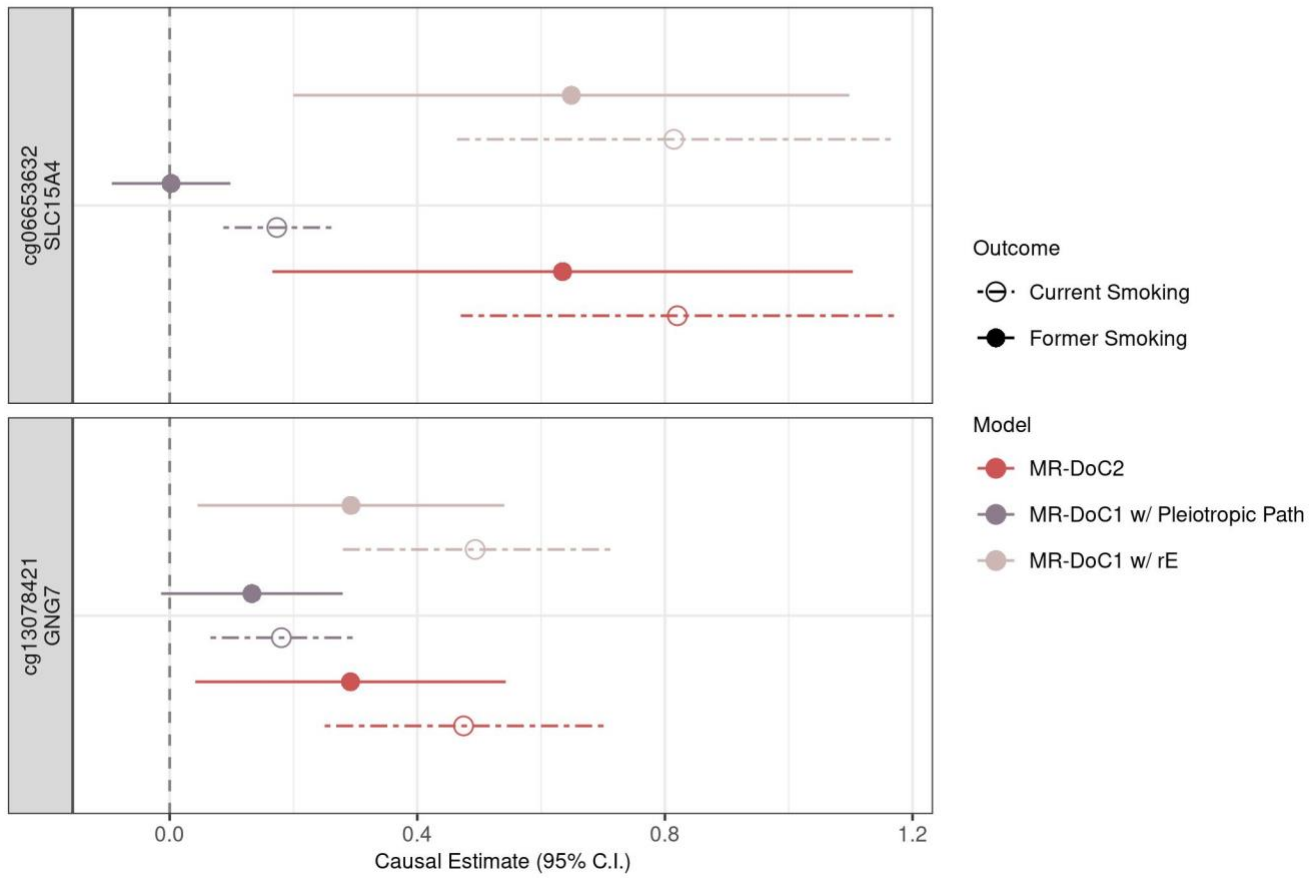
402

403
404
405
406

Figure S34

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



407
408
409
410
411

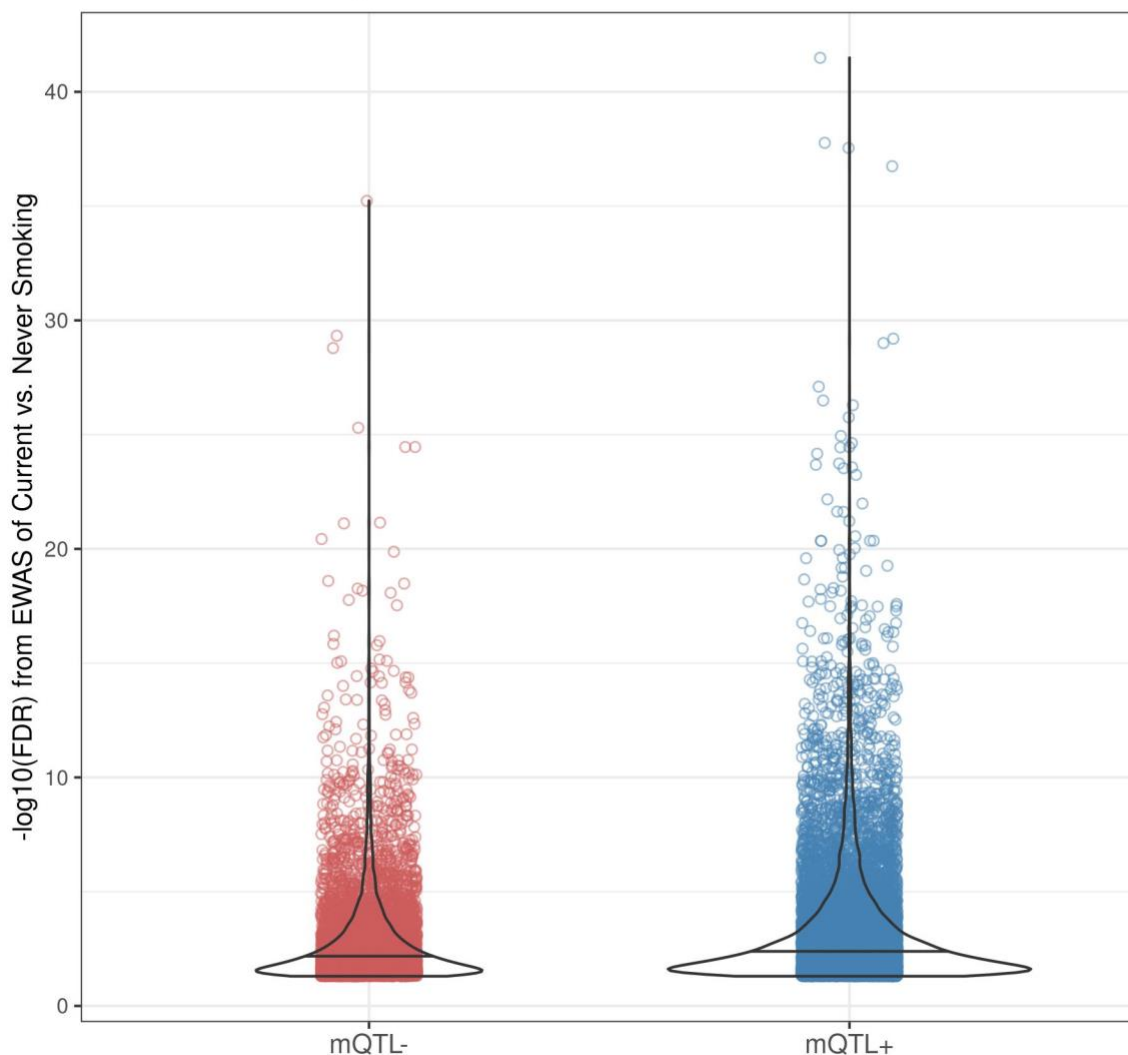
Note. Please refer to **Supplementary Tables S3 and S29** for the corresponding data.

412
413
414
415

Figure S35

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



416
417
418
419
420
421

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 $-\log_{10}(\text{FDR})$ values of the association results from the previous EWAS meta-analysis of current vs. never smoking¹³. The “mQTL-” CpGs were not tested for DNAm \rightarrow Smoking causal effects in the current study.

422 **References**

- 423 1. Ligthart, L. *et al.* The Netherlands Twin Register: Longitudinal Research Based on Twin and Twin-
424 Family Designs. *Twin Research and Human Genetics* **22**, 623–636 (2019).
- 425 2. Minică, C. C., Dolan, C. V., Boomsma, D. I., De Geus, E. & Neale, M. C. Extending Causality Tests
426 with Genetic Instruments: An Integration of Mendelian Randomization with the Classical Twin
427 Design. *Behavior Genetics* **48**, 337–349 (2018).
- 428 3. Castro-de-Araujo, L. F. S. *et al.* MR-DoC2: Bidirectional Causal Modeling with Instrumental
429 Variables and Data from Relatives. *Behavior Genetics* **53**, 63–73 (2023).
- 430 4. Singh, M. *et al.* Using Instrumental Variables to Measure Causation over Time in Cross-Lagged Panel
431 Models. *Multivariate Behavioral Research* **59**, 342–370 (2024).
- 432 5. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- 433 6. Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation.
434 *Nature Genetics* **48**, 1279–1283 (2016).
- 435 7. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history
436 of the Dutch population. *Nature Genetics* **46**, 818–825 (2014).
- 437 8. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association
438 studies. *Nature Genetics* **38**, 904–909 (2006).
- 439 9. Saunders, G. R. B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature*
440 **612**, 720–724 (2022).
- 441 10. Vilhjálmsón, J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk
442 Scores. *The American Journal of Human Genetics* **97**, 576–592 (2015).
- 443 11. Breeze, C. E. *et al.* eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic
444 Data. *Cell Reports* **17**, 2137–2150 (2016).
- 445 12. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level
446 datasets. *Nature Communications* **10**, (2019).
- 447 13. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circulation. Cardiovascular*
448 *genetics* **9**, 436–447 (2016).

449
450