Causation between Smoking and DNA methylation

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

5					
4 5	Madhurbain Singh <sup>1,2,3*</sup> , Conor V. Dolan <sup>3,4</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> ,				
6	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				
7	Dongen <sup>3,4,11</sup> *, and Michael C. Neale <sup>1,2,3,11</sup> *				
8					
9	1. Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry,				
10	Virginia Commonwealth University, Richmond, VA, USA				
11	2. Department of Human and Molecular Genetics, Virginia Commonwealth University,				
12	Richmond, VA, USA				
13	3. Department of Biological Psychology, Vrije Universiteit (VU) Amsterdam, Amsterdam,				
14	The Netherlands				
15	4. Amsterdam Public Health Research Institute, Amsterdam, The Netherlands				
16 17	<ol> <li>Department of Psychiatry and Behavioral Sciences, Texas A&amp;M University, College Station, TX, USA</li> </ol>				
18	6 Division of Cancer Enidemiology and Cancer Institute National				
10	Institutes of Health Department Health and Human Services, Bethesda, MD, USA				
20	7 UCL Cancer Institute University College London, London, UK				
20	<ol> <li>MC Integrative Epidemiology Unit, University of Pristel, Pristel, UK</li> </ol>				
21	0 Department of Psychiatry and Behavioral Sciences, SUNV Downstate Health Sciences				
22	9. Department of 1 sychiatry and Denavioral Sciences, SON 1 Downstate freatm Sciences				
23	10 Institute for Genomics in Health SUNY Downstate Health Sciences University				
2 <del>1</del> 25	Brooklyn NY USA				
26	Diookiyii, ivi, com				
27	11 These authors jointly supervised this work				
28	12 Current address: Department of Complex Trait Genetics Center for Neurogenomics and				
29	Cognitive Research Vrije Universiteit (VII) Amsterdam Amsterdam The Netherlands				
30					
31	*Corresponding authors:				
32	Madhurbain Singh. Email: singhm18@vcu.edu. Address: Virginia Institute for Psychiatric and				
33	Behavioral Genetics, 800 E. Leigh St., Suite 100, Richmond, VA 23298, USA				
34	Jenny van Dongen. Email: j.van.dongen@vu.nl. Address: Department of Biological Psychology,				
35	Vrije Universiteit Amsterdam, van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands				
36	Michael C. Neale. Email: michael.neale@vcuhealth.org. Address: Virginia Institute for				
37	Psychiatric and Behavioral Genetics, 800 E. Leigh St., Suite 100, Richmond, VA 23298, USA				
38					
39	Running head: Causation between smoking and DNA methylation				
	- •				

Causation between Smoking and DNA methylation

#### 40 Table of Contents

41	Funding	2
42	Acknowledgments	3
43	Conflicts of Interest	3
44	ORCID	3
45	Data Availability	3
46	Code Availability	4
47	Abstract	5
48	Introduction	6
49	Results	
50	cis-mQTLs identified for two-thirds of smoking-associated CpG sites	10
51	Exemplar: Putative causality between current smoking and AHRR DNAm	13
52	Evidence of more widespread effects of current smoking on DNAm than vice versa	17
53	Suggestive Evidence of Bidirectional Effects at Four CpG Sites	23
54	DNAm loci potentially influenced by smoking are enriched for biological processes relevant to	
55	smoking's adverse health outcomes	25
56	CpG sites with consistent effects on current smoking show enrichment for brain-related gene reg	ulatory
57	elements	25
58	Attenuated effects of former smoking on DNAm	29
59	Discussion	
60	Methods	
61	Study Sample	34
62	Peripheral Blood DNA Methylation and Cell Counts	34
63	Cigarette Smoking	35
64	Instrumental Variables	35
65	MR-DoC Models	37
66	Functional Enrichment Analyses	39
67	eFORGE (experimentally derived Functional element Overlap analysis of ReGions from EWAS)	39
68	References	41
69		

# 70 Funding

71 We acknowledge funding from the U.S. National Institute on Drug Abuse grant R01DA049867,

the Netherlands Organization for Scientific Research (NWO): Biobanking and Biomolecular

73 Research Infrastructure (BBMRI-NL, NWO 184.033.111) and the BBRMI-NL-financed BIOS

74 Consortium (NWO 184.021.007), NWO Large Scale infrastructures X-Omics (184.034.019),

75 Genotype/phenotype database for behavior genetic and genetic epidemiological studies (ZonMw

76 Middelgroot 911-09-032); Netherlands Twin Registry Repository: researching the interplay

- between genome and environment (NWO-Groot 480-15-001/674); the Avera Institute, Sioux
- Falls (USA), and the U.S. National Institutes of Health (NIH R01HD042157-01A1,
- 79 R01MH081802, R01MH125938, and Grand Opportunity grants 1RC2 MH089951 and 1RC2
- 80 MH089995). DML is supported by the NIH K01MH131847. DIB acknowledges the Royal

Causation between Smoking and DNA methylation

- 81 Netherlands Academy of Science Professor Award (PAH/6635). JLM and GH are supported by
- 82 the UK Medical Research Council (MRC) Integrative Epidemiology Unit at the University of
- 83 Bristol (MC\_UU\_00011/1, MC\_UU\_00011/5).

### 84 Acknowledgments

- 85 NTR warmly thanks all participants. Epigenetic data were generated at the Human Genomics
- 86 Facility (HuGe-F) at ErasmusMC Rotterdam (http://www.glimdna.org/) as part of the Biobank-
- 87 based Integrative Omics Study Consortium. We thank Dr. Scott Vrieze (University of
- 88 Minnesota) for providing the leave-one-out GWAS summary statistics from the GWAS &
- 89 Sequencing Consortium of Alcohol and Nicotine Use (GSCAN).

# 90 Conflicts of Interest

91 Nothing to declare.

# 92 **ORCID**

- 93 Madhurbain Singh: 0000-0002-9396-2860
- 94 Conor V. Dolan: 0000-0002-2496-8492
- 95 Dana M. Lapato: 0000-0001-8169-9754
- 96 Jouke-Jan Hottenga: 0000-0002-5668-2368
- 97 René Pool: 0000-0001-5579-0933
- 98 Brad Verhulst: 0000-0001-5369-9757
- 99 Dorret I. Boomsma: 0000-0002-7099-7972
- 100 Charles E. Breeze: 0000-0002-5294-915X
- 101 Eco J. C. de Geus: 0000-0001-6022-2666
- 102 Gibran Hemani: 0000-0003-0920-1055
- 103 Josine L. Min: 0000-0003-4456-9824
- 104 Roseann E. Peterson: 0000-0001-6402-849X
- 105 Hermine H. M. Maes: 0000-0001-7489-2214
- 106 Jenny van Dongen: 0000-0003-2063-8741
- 107 Michael C. Neale: 0000-0003-4887-659X

# 108 Data Availability

- 109 Data from the Netherlands Twin Register (NTR) may be accessed for research purposes by
- 110 submitting a data-sharing request. Further information about NTR data access is available at
- 111 <u>https://ntr-data-request.psy.vu.nl/</u>.
- 112
- 113 Results of all MR-DoC models fitted in this study are available as Supplementary Data.

Causation between Smoking and DNA methylation

# 114 Code Availability

- 115 The code used in the analyses for this study is available at: <u>https://github.com/singh-</u>
- 116 <u>madhur/MRDOC\_Smoking\_DNAm\_NTR</u>.

Causation between Smoking and DNA methylation

#### 118

### Abstract

- 119 Epigenome-wide association studies (EWAS) aim to identify differentially methylated loci
- 120 associated with complex traits and disorders. EWAS of cigarette smoking shows some of the
- 121 most widespread DNA methylation (DNAm) associations in blood. However, traditional EWAS
- 122 cannot differentiate between causation and confounding, leading to ambiguity in etiological
- 123 interpretations. Here, we apply an integrated approach combining Mendelian Randomization and
- 124 twin-based Direction-of-Causation analyses (MR-DoC) to examine causality underlying
- smoking-associated blood DNAm changes in the Netherlands Twin Register (N=2577).
- 126 Evidence across models suggests that current smoking's causal effects on DNAm likely drive
- 127 many of the previous EWAS findings, implicating functional pathways relevant to several
- 128 adverse health outcomes of smoking, including hemopoiesis, cell- and neuro-development, and
- 129 immune regulation. Additionally, we find evidence of potential reverse causal influences at some
- 130 DNAm sites, with 17 of these sites enriched for gene regulatory functional elements in the brain.
- 131 The top three sites with evidence of DNAm's effects on smoking annotate to genes involved in G
- 132 protein-coupled receptor signaling (*GNG7*, *RGS3*) and innate immune response (*SLC15A4*),
- 133 elucidating potential biological risk factors for smoking. This study highlights the utility of
- 134 integrating genotypic and DNAm measures in twin cohorts to clarify the causal relationships
- 135 between health behaviors and blood DNAm.

Causation between Smoking and DNA methylation

#### 137

## Introduction

138 Epigenome-wide association studies (EWASs) are valuable for identifying variation in DNA

- 139 methylation (DNAm) associated with complex human traits and diseases<sup>1</sup>. By far, the most
- 140 successful EWASs have been the studies of cigarette smoking. A large-scale EWAS meta-
- 141 analysis of smoking (N = 15,907 individuals) compared current versus never smoking to reveal
- 142 significant DNAm differences at 18,760 CpG (Cytosine-phosphate-Guanine) sites in peripheral
- 143 blood cells<sup>2</sup>. DNAm differences between former- and never-smoking individuals were
- 144 diminished but remained statistically significant at 2,568 sites<sup>2</sup>. Genes annotated to the
- 145 differentially methylated CpG sites have been implicated in genome-wide association studies
- 146 (GWAS) of several smoking-associated traits, including cancers, lung functions, cardiovascular
- 147 disorders, inflammatory disorders, and schizophrenia, indicating DNAm's potential role in the
- 148 adverse health effects of smoking<sup>2</sup>.
- 149

150 As cross-sectional EWAS in unrelated individuals cannot differentiate between causation and

151 confounding<sup>3</sup>, the widespread associations between cigarette smoking and DNAm<sup>2</sup> may originate

152 from a combination of different etiological mechanisms. These associations are typically

153 interpreted as the causal *effects* of smoking exposure on DNAm. However, some smoking-

associated CpG sites may have reverse or bidirectional causal links with smoking, such that

155 DNAm may reciprocally affect the development and maintenance of smoking behaviors<sup>4</sup>.

156 Moreover, associations between smoking and DNAm could be attributable to potential

157 confounders, such as schizophrenia<sup>5</sup>, alcohol consumption<sup>6</sup>, cannabis use<sup>7</sup>, and body mass

- 158 index<sup>8</sup>.
- 159

160 An alternative approach to causal inference in observational studies is Mendelian Randomization 161 (MR) analysis, using genetic variants as instrumental variables (IVs) to estimate causal effects under specific assumptions<sup>3,9</sup> (see **Methods**). Previous MR analyses have identified potential 162 163 effects of lifetime (current or former) smoking liability on blood DNAm at only 11 CpG sites<sup>10</sup>, along with potential reverse effects of blood DNAm at 9 sites<sup>11</sup>. The causal inference in MR is 164 165 based on the assumption that the genetic variants associated with the exposure influence the outcome exclusively through the exposure. In other words, the genetic variants used as IVs for 166 smoking may show vertical pleiotropy, but not horizontal pleiotropy, with DNAm. To minimize 167 potential violations of these assumptions, MR analyses require carefully selected single-168 169 nucleotide polymorphisms (SNPs), including using genetic colocalization to filter out SNPs 170 showing horizontal pleiotropy due to linkage disequilibrium (LD). Since individual SNPs usually 171 have minuscule effect sizes on complex traits, traditional MR approaches using a few selected 172 SNPs may have limited power to detect causality and may be subject to weak-instrument bias<sup>12</sup>. 173 174 Recent methodological developments<sup>13,14</sup> integrate the principles of MR with the twin-based

175 *Direction of Causation* model (hence called *MR-DoC*) from biometrical studies of mono- and

176 dizygotic twins<sup>15</sup>. Causal inference in twin data leverages the cross-twin cross-trait correlations

Causation between Smoking and DNA methylation

- to estimate the direction and magnitude of potential causal effects between traits<sup>18</sup>. Further, the
- 178 MR-DOC approaches, i.e., the unidirectional MR- $DoC1^{13}$  and the bidirectional MR- $DoC2^{14}$ , help
- account for some of the horizontal pleiotropic associations of the genetic IV with the outcome,
- 180 unmediated by the exposure trait. Consequently, MR-DoC models allow using polygenic risk
- 181 scores (PRS) as potential IVs, increasing the statistical power to estimate causal effects while
- 182 curtailing weak-instrument bias relative to traditional MR methods that use SNPs as IVs.
- 183 Incorporating MR with family data also helps to resolve additional assumptions of standard MR,
- 184 such as random mating and no dyadic effects  $^{13,16}$ .
- 185
- 186 The present study used MR-DoC models to examine bidirectional causal effects between
- 187 cigarette smoking and peripheral blood DNAm in European ancestry adult twins from the
- 188 Netherlands Twin Register (NTR)<sup>17</sup> (**Figure 1**). The target sample included 2,577 individuals
- 189 from 1,459 twin pairs with both genotypic and DNAm data, as well as their self-reported
- smoking status at the time of blood draw (comprising 528 currently, 549 formerly, and 1,492
- 191 never-smoking individuals). Across 16,940 smoking-related CpGs previously identified<sup>2</sup>, we
- 192 fitted separate models for current (versus never) and former (versus never) smoking. We
- 193 obtained a set of three causal estimates in each direction (*Smoking*  $\rightarrow$  *DNAm*, and *DNAm*  $\rightarrow$
- 194 *Smoking*): the estimates from bidirectional MR-DoC2 and two different model specifications of
- unidirectional MR-DoC1 (Figure 1). We triangulated evidence across the three models based on
- 196 the statistical significance and consistency of the causal estimates. The results indicated more
- 197 widespread putative causal influences of smoking on DNAm than *vice versa*. Follow-up
- 198 enrichment analyses highlighted biological processes and tissues relevant to the CpG sites with
- 199 potential effects in either direction of causation.



Causation between Smoking and DNA methylation

202 Figure 1. Study Design.

- 203 Overview of the data and MR-DoC models used to examine the causality between cigarette smoking and blood DNA methylation (DNAm) in the
- 204 Netherlands Twin Register. The models were fitted separately for current (versus never) and former (versus never) smoking. Applying the five MR-
- 205 DoC models shown in the path diagrams, we obtained a set of three causal estimates in each direction of causation: Smoking  $(Smk) \rightarrow DNAm$  (the
- blue paths labeled  $g_1$  and  $DNAm \rightarrow Smoking$  (the red paths labeled  $g_2$ ).
- 207 Note. For better readability, the path diagrams show only the within-individual part of the models fitted to data from twin pairs. The
- 208 squares/rectangles indicate observed variables, the circles indicate latent (unobserved) variables, the single-headed arrows indicate regression
- 209 *paths, and the double-headed curved arrows indicate (co-)variances.*

Causation between Smoking and DNA methylation

210

# Results

#### 211 cis-mQTLs identified for two-thirds of smoking-associated CpG sites

212 We used a weighted sum of relevant DNAm-increasing alleles at *cis*-methylation quantitative 213 trait loci (henceforth called *mOTL allelic score*) as the IV for DNAm. Of the 18,760 CpG sites 214 associated with current smoking in a previous independent EWAS meta-analysis<sup>2</sup>, 16,940 215 autosomal sites passed the QC metrics in NTR (hereafter called the "smoking-associated CpGs") 216 and were analyzed in the unidirectional MR-DoC1 models for Current Smoking  $\rightarrow$  DNAm (Figure 2). Of these sites, 13,275 had mQTL summary statistics from the Genetics of DNA 217 Methylation Consortium (GoDMC; excluding NTR)<sup>18</sup>. A subset of 12,940 sites had summary 218 219 statistics for *cis*-mQTLs, i.e., SNPs within 1Mb of the CpG. We used only *cis*-mQTLs to derive 220 the IVs for DNAm, given that SNPs located close to the CpG are more likely to be associated 221 with smoking *via* DNAm. To further guard against potential horizontal pleiotropy with smoking, 222 we relied on the consistency of the causal estimates in MR-DoC models accommodating 223 horizontal pleiotropy. To reduce the risk of weak-instrument bias in the estimated effects of 224 DNAm on smoking, we restricted the MR-DoC1 models for  $DNAm \rightarrow Current Smoking$  and the 225 bidirectional MR-DoC2 models to 11,124 (65.7%) smoking-associated CpGs having an mQTL 226 allelic score with F-statistic >10, the criterion for the "relevance" assumption of a valid  $IV^{19}$  (see Methods). The included mOTL allelic scores had an incremental R<sup>2</sup> for the respective CpG site 227 228 ranging from 0.43% to 76.95% (mean 9.04%, S.D. = 10.94%). Applying similar inclusion 229 criteria, we identified 2,330 autosomal, post-OC CpG sites previously associated with former 230 smoking<sup>2</sup> (hereafter called the "former-smoking-associated CpGs"), which were analyzed in the MR-DoC1 models for Former Smoking  $\rightarrow$  DNAm. A subset of 1,782 (76.5%) former-smoking-231 232 associated CpGs had mQTL allelic scores with F-statistic >10 and were examined in the MR-233 DoC1 models for  $DNAm \rightarrow Former Smoking$  and the bidirectional models. 234 We used a PRS of lifetime regular-smoking initiation<sup>20</sup> as the IV for smoking status, which had 235

an incremental liability-scale  $R^2$  of 5.07% (F-statistic = 73.2) for current versus never smoking

- and 2.02% (F-statistic = 28.8) for former versus never smoking in the target NTR dataset.
- 238

Causation between Smoking and DNA methylation



240 Figure 2. Selection of CpG sites tested in each MR-DoC model.

- 241 Previous independent EWAS meta-analysis of cigarette smoking<sup>2</sup> examined DNA methylation
- 242 (DNAm) at CpG sites from the Illumina HumanMethylation450 BeadChip array<sup>21</sup>, which was

- 243 also used to measure DNAm in the NTR biobank. In the unidirectional MR-DoC1 models for
- 244 Smoking  $\rightarrow$  DNAm, we included autosomal CpG sites associated with smoking in the EWAS
- 245 meta-analysis that also passed the QC metrics in NTR. The MR-DoC1 models for DNAm  $\rightarrow$
- 246 Smoking and the bidirectional MR-DoC2 models were restricted to a subset of these sites having
- 247 cis-mQTL summary statistics from the  $GoDMC^{18}$  and a resulting mQTL allelic score with F-
- 248 statistic > 10.
- 249

Causation between Smoking and DNA methylation

### 250 Exemplar: Putative causality between current smoking and AHRR DNAm

251 To illustrate the three MR-DoC models, we first present the results for two CpG sites

- 252 (cg23916896 and cg05575921) in the Aryl-Hydrocarbon Receptor Repressor (AHRR) gene,
- which are among the most well-established DNAm signatures of cigarette smoking<sup>2</sup>.
- 254
- 255 One of the two MR-DoC1 model specifications allowed us to estimate and account for potential 256 unbalanced horizontal pleiotropy from the mQTL allelic score to smoking in  $DNAm \rightarrow Smoking$ 257 models and from the smoking PRS to DNAm in *Smoking*  $\rightarrow$  *DNAm* models. However, to 258 estimate this pleiotropic association, the model requires fixing the confounding due to unique 259 environmental factors to a specific value (here, zero)<sup>13</sup>. In the second specification of MR-DoC1, we freely estimated and controlled for potential unique environmental confounding (labeled "rE" 260 261 in Figure 1), while instead assuming that the IV had no horizontal pleiotropy. In MR-DoC2 262 models, we estimated bidirectional causal effects by including both the smoking PRS and the 263 mQTL allelic score, while allowing the two IVs to covary with each other<sup>14</sup>. Covariance between 264 the PRS and the mQTL allelic score may arise from many possible sources, including shared 265 pleiotropic SNPs, LD between the constituent SNPs, and population structure. Therefore, MR-
- 266 DoC2 may help reduce potential biases in the causal estimates by accounting for these sources of
- 267 covariance between smoking PRS and mQTL allelic score. Across all models, causal
- relationships with the binary smoking variable are estimated on the latent liability scale<sup>22</sup>. So,
- even where smoking is the "exposure" variable, the causal estimate is interpreted as the effect ofthe underlying smoking *liability* rather than smoking *exposure*.
- 271

272 For probe cg23916896 (Figure 3A), the mOTL allelic score had an incremental R<sup>2</sup> of 8.03% (F-273 statistic = 156.4). The estimated effects indicated that higher liability for current smoking likely 274 causes hypomethylation of cg23916896, with consistently negative causal estimates: -0.82 (95% 275 confidence interval: -1.20, -0.44) in MR-DoC1 with horizontal pleiotropy, -0.43 (-0.62, -0.24) in 276 MR-DoC1 with unique environmental confounding, and -0.38 (-0.55, -0.21) in the bidirectional 277 MR-DoC2 model. These estimates remained statistically significant after FDR correction in all 278 three models. The estimated reverse effect of cg23916896 methylation on the liability for current 279 smoking also had consistently negative estimates in all models: -0.24 (-0.37, -0.12) in MR-DoC1 280 with horizontal pleiotropy, -0.32 (-0.61, -0.04) in MR-DoC1 with unique environmental 281 confounding, and -0.32 (-0.61, -0.04) in MR-DoC2. That is, hypomethylation of cg23916896 putatively increases the liability for current smoking. These estimates were statistically 282 283 significant at false discovery rate (FDR) <0.05 in MR-DoC1 with horizontal pleiotropy, but only 284 nominally significant (p <0.05) in the other two models. Taken together, these results provide 285 robust evidence for current smoking's causal effects on cg23916896 methylation, with suggestive evidence for the reverse effect of cg23916896 methylation on smoking. Previous MR 286 287 studies of lifetime smoking and DNAm have not examined this CpG site, as these studies focused on a few selected sites<sup>10,11</sup>. Our analyses indicate that cg23916896 potentially has a 288

Causation between Smoking and DNA methylation

- 289 bidirectional causal relationship with cigarette smoking, such that the smoking-induced
- 290 hypomethylation at this locus may reciprocally increase the liability for smoking.
- 291
- In comparison, probe cg05575921 (one of the CpGs most robustly associated with cigarette
- smoking) had an mQTL allelic score with a relatively modest incremental R<sup>2</sup> of 1.74% (F-
- statistic = 31.6). Similar to cg23916896, the effect of current smoking liability on cg05575921
- 295 methylation had consistently negative, robust estimates, with FDR <0.05 in all three models
- 296 (**Figure 3B**), which also aligns with the previously reported negative, albeit non-significant,
- 297 effect of *lifetime* smoking<sup>10</sup>. The reverse effect of cg05575921 methylation on smoking liability
- was estimated to be -1.29 (-1.62, -0.96) in MR-DoC1 with horizontal pleiotropy, -0.41 (-1.03,
- 0.21) in MR-DoC1 with unique environmental confounding, and -0.37 (-1.00, 0.26) in MR-
- 300 DoC2. Although the point estimates were negative in all three models, they were not statistically
- 301 significant in the latter two models. Notably, the point estimates for cg05575921 are comparable
- 302 to those for cg23916896 but have larger standard errors, likely due to the former's weaker IV
- 303 (mQTL allelic score).

Causation between Smoking and DNA methylation



Figure 3. Illustrative MR-DoC models of causality between current smoking and blood DNAm
at (A) cg23916896 and (B) cg05575921 in the AHRR gene.

- 308 We fitted five MR-DoC models at each CpG: (1) Smoking  $\rightarrow$  DNAm MR-DoC1 with horizontal
- 309 pleiotropy, (2) Smoking  $\rightarrow$  DNAm MR-DoC1 with unique environmental confounding, (3) DNAm
- 310  $\rightarrow$  Smoking MR-DoC1 with horizontal pleiotropy, (4) DNAm  $\rightarrow$  Smoking MR-DoC1 with unique
- 311 environmental confounding, and (5) bidirectional MR-DoC2. Thus, for each CpG, three causal
- 312 *estimates were obtained in either direction of causation.*
- 313 In the path diagrams, squares/rectangles indicate observed variables, circles indicate latent
- 314 (unobserved variables), single-headed arrows indicate regression paths, and double-headed
- 315 curved arrows indicate (co-)variance. The residual variance of smoking status liability is
- 316 partitioned into additive genetic (Asmk) and unique environmental (Esmk) components. Likewise,
- 317 the residual variance of DNAm is partitioned into ADNAm and EDNAm. The correlation between
- 318 Asmk and ADNAm represents the confounding between smoking and DNAm due to latent
- 319 (unobserved) additive genetic factors, while the correlation between E<sub>Smk</sub> and E<sub>DNAm</sub> represents
- 320 confounding due to latent unique environmental factors. Each model included age and sex as
- 321 covariates of smoking status (not shown). DNAm β-values were residualized for standard
- 322 biological and technical covariates used in EWAS (see Methods). The smoking PRS and the
- 323 *mQTL* allelic scores were residualized for standard GWAS covariates, including genetic
- 324 principal components and genotyping platform. In the path diagrams, the residualized PRS and
- 325 *mQTL* allelic scores are regressed on respective latent factors, representing the underlying
- 326 *"true" standardized scores (mean = zero; variance = one). The coefficient of the path from the*
- 327 *latent score to the observed score estimates the standard deviation of the observed score.*
- 328 Note. The paths are labeled by the point estimate and its S.E. in parentheses. For better
- 329 readability, the path diagrams show only the within-individual part of the models fitted to data
- *from twin pairs.* 330
- 331

Causation between Smoking and DNA methylation

#### 332 Evidence of more widespread effects of current smoking on DNAm than vice

#### 333 *versa*

334 To evaluate whether there was evidence of widespread, small causal effects of current smoking 335 on DNAm, we examined the Bayesian genomic inflation factor<sup>23</sup> ( $\lambda$ ) using p-values of the causal 336 estimates. Across the 16,940 smoking-associated CpG sites, MR-DoC1 with horizontal 337 pleiotropy had  $\lambda = 1.44$ , while MR-DoC1 with unique environmental confounding showed  $\lambda =$ 338 1.20. For comparison, fitting similar models epigenome-wide at 411,169 autosomal, post-OC 339 CpGs showed much less inflation ( $\lambda = 0.98$  and  $\lambda = 1.09$ , respectively), suggesting enrichment of 340 low p-values among the smoking-associated CpGs. The epigenome-wide inflation is in line with 341 that for cigarettes per day ( $\lambda > 1.1$ ) previously reported using two-sample MR<sup>18</sup>. Corresponding 342 QQ plots showed a deviation of the causal estimate p-values from the null hypothesis across a 343 broad range of smoking-associated CpG sites (Supplementary Figures S1, S2). Across the 344 11,124 CpG sites with bidirectional MR-DoC2 models, the estimated reverse effects of DNAm 345 on current smoking showed little inflation ( $\lambda = 1.01$ ) compared to the effects of current smoking 346 on DNAm in the same model ( $\lambda = 1.20$ ; Supplementary Figures S3, S4). These findings 347 suggest that the causal influences of current smoking on DNAm likely contribute, at least partly, 348 to the previously reported EWAS hits. For the reverse effects of DNAm on current smoking, the 349 absence of  $\lambda$  inflation does not preclude potential localized small effects at several CpG sites.

- 350 Furthermore, despite the inflation of the test statistics, our sample size might be insufficient to
- 351 obtain significant estimates of relatively small effects in either direction of causation.
- 352

353 There also was considerable variability in the number of CpG sites with statistically significant causal estimates across models. The estimated *Current Smoking*  $\rightarrow$  *DNAm* effects had FDR 354 355 <0.05 at 1,368 CpGs in MR-DoC1 with horizontal pleiotropy, 334 CpGs in MR-DoC1 with 356 unique environmental confounding, and 275 CpGs in MR-DoC2 (Figure 4; top panel). The 357 relatively higher number of statistically significant causal estimates in MR-DoC1 with horizontal 358 pleiotropy may partly be due to its higher power compared to the other models<sup>24</sup>. Looking at the 359 intersection of significant estimates across models, 259 CpG sites showed FDR <0.05 in at least 360 two models, while 64 sites showed FDR < 0.05 in all three models. These 64 sites also showed 361 consistency in the direction of effect across all three models (Supplementary Figure S5, Table 362 S1). Thus, we considered these 64 CpG sites to exhibit robust evidence for the causal effects of 363 current smoking liability on DNAm, including hypomethylation of 59 sites and hypermethylation 364 of the other five (Figure 4; bottom panel). These CpGs are annotated to some of the top genes 365 implicated in prior EWAS of smoking<sup>2</sup>, including hypomethylation of CpGs in/near AHRR, 366 ALPPL2 (alkaline phosphatase placental-like 2), CNTNAP2 (contactin-associated protein 2), and 367 PARD3 (par-3 family cell polarity regulator) and hypermethylation of CpGs in MYO1G (myosin 368 1G). Only one of these 64 CpG sites lies within the major histocompatibility complex (MHC) 369 region (chr6:28477797-33448354): cg06126421 located near gene HLA-DRB5. Due to its 370 complex LD structure, the causal estimates of the sites in the MHC region should be interpreted 371 with caution.

Causation between Smoking and DNA methylation

#### 372

#### Putative Causal Effects of Current Smoking on DNA Methylation in MR-DoC Models



Causation between Smoking and DNA methylation

# Figure 4. Putative Causal Effects of Current Smoking on Blood DNA Methylation in MR DoC Models

- 377 The top panel shows an UpSet plot of the intersection of CpG sites with statistically significant
- 378 (FDR < 0.05) estimates of Current Smoking  $\rightarrow$  DNAm in the three MR-DoC models. The matrix
- 379 consists of the models along the three rows and their intersections along the columns. The
- 380 horizontal bars on the left represent the number of CpGs with significant (FDR < 0.05) causal
- 381 estimates in each model. The vertical bars represent the number of CpGs belonging to the
- 382 *respective intersection in the matrix.*
- 383 The bottom panel shows a Miami plot of the Current Smoking  $\rightarrow$  DNAm causal estimates across
- 384 16,940 smoking-associated CpGs. The X-axis shows the genomic positions of the CpG sites
- aligned to Genome Reference Consortium Human Build 37 (GRCh37). The Y-axis shows the Z-
- 386 statistic of the estimated effect of the liability for current (versus never) smoking on (residualized
- and standardized) DNA methylation  $\beta$ -values in the MR-DoC1 model with unique environmental
- 388 confounding (rE). The solid points indicate the 64 sites with significant causal estimates (FDR
- 389 <0.05) in all three models (i.e., the blue vertical bar in the UpSet plot). The CpG sites with
- 390 causal estimates significant after Bonferroni correction in more than one model are labeled by
- 391 *their respective nearest gene.*
- 392 Note. The data underlying these plots are in **Supplementary Table S1**.
- 393

- 394 On applying a more conservative Bonferroni correction for multiple testing, 14 sites had
- 395 significant *Current Smoking*  $\rightarrow$  *DNAm* causal estimates in more than one model, while only four
- 396 CpGs had significant estimates in all three models (**Supplementary Figure S6**). Thus, these four
- 397 CpGs showed the most robust evidence for the effects of current smoking on DNAm, comprising
- three sites with hypomethylation (cg05951221 and cg01940273 near *ALPPL2*, and cg06126421
- 399 near *HLA-DRB5*) and one with hypermethylation (cg12803068 in *MYO1G*).
- 400
- 401 The estimated  $DNAm \rightarrow Current Smoking$  effects were significant (FDR <0.05) at 1,081 CpGs
- 402 in MR-DoC1 with horizontal pleiotropy, 51 CpGs in MR-DoC1 with unique environmental
- 403 confounding, and 54 CpGs in MR-DoC2 (**Figure 5**; right panel). Further, 44 CpGs showed
- 404 FDR <0.05 in at least two models, but only three CpGs had FDR <0.05 in all three models. The
- 405 three CpGs also had consistent, positive estimates across models, suggesting that
- 406 hypermethylation of CpG sites in GNG7 (G-Protein Subunit Gamma 7), RGS3 (Regulator of G-
- 407 Protein Signaling 3), and *SLC15A4* (Solute Carrier Family 15 Member 4) genes may increase the
- 408 liability for current smoking (**Figure 5; left panel**). None of these sites has been previously
- 409 reported to have effects on smoking liability<sup>11</sup>. Applying the more conservative Bonferroni
- 410 correction, nine CpGs had significant  $DNAm \rightarrow Current Smoking$  causal estimates in more than
- 411 one model, but none showed Bonferroni-corrected significant effects in all three models
- 412 (Supplementary Figure S7).

#### 413

Putative Causal Effects of DNA Methylation on Current Smoking in MR-DoC Models



#### 414

#### 415 Figure 5. Putative causal effects of blood DNA methylation on current-smoking liability in MR-DoC models

416 The left panel shows the estimates and Wald-type 95% confidence intervals of the causal effects of (residualized and standardized)

- 417 DNA methylation β-values on the liability for current (versus never) smoking in each of the three MR-DoC models: bidirectional MR-
- 418 *DoC2*, *MR-DoC1* with horizontal pleiotropic path, and *MR-DoC1* with unique environmental confounding (*rE*). The text labels
- 419 *indicate the gene to which the CpG is annotated.*

- 420 The right panel shows an UpSet plot of the intersection of CpG sites with statistically significant (FDR < 0.05) estimates of DNAm  $\rightarrow$
- 421 Current Smoking in each of the three MR-DoC models. The matrix consists of the models along the three rows and their intersections
- 422 along the columns. The horizontal bars on the left represent the number of CpGs with significant (FDR < 0.05) causal estimates in
- 423 each model. The vertical bars represent the number of CpGs belonging to the respective intersection in the matrix.
- 424 Note. The data underlying these plots are in **Supplementary Table S3**.

Causation between Smoking and DNA methylation

#### 426 Suggestive Evidence of Bidirectional Effects at Four CpG Sites

- 427 The 64 CpG sites with robust evidence of current smoking's effects on DNAm do not overlap
- 428 with the three sites with robust evidence of reverse effects. However, further examining the
- 429 causal estimates revealed that three of the 64 sites also had consistently negative, nominally
- 430 significant (p <0.05) estimates of  $DNAm \rightarrow Current Smoking$  effects in all models (Figure 6).
- 431 The three CpGs (cg23916896, cg11902777, cg01899089) are all located in the *AHRR* gene,
- 432 suggesting potential bidirectional effects between current smoking and AHRR DNAm. That is,
- 433 current smoking putatively causes hypomethylation of CpGs in *AHRR*, which, in turn, may
- 434 further increase smoking liability as a feedback effect. Among the CpGs with robust evidence of
- 435 DNAm's effects on current smoking, cg13078421 (GNG7) also showed consistently positive,
- 436 nominally significant estimates of current smoking's effects on DNAm. Thus, GNG7
- 437 hypermethylation putatively increases smoking liability, with a potential reverse effect of current
- 438 smoking on *GNG7* methylation. Additionally, 15 CpGs had consistent, nominally significant
- 439 bidirectional causal estimates in all three models, though the estimates were not significant after
- 440 FDR correction in either direction (Supplementary Figure S8).
- 441

Suggestive Bidirectional Causal Effects between Current Smoking and Blood DNA Methylation



442

#### 443 Figure 6. Potential bidirectional effects between current smoking and blood DNA methylation

- 444 Estimates and Wald-type 95% confidence intervals of bidirectional causal effects between the
- 445 liability for current (versus never) smoking and (residualized and standardized) DNA
- 446 methylation  $\beta$ -values in the three MR-DoC models: bidirectional MR-DoC2, MR-DoC1 with
- 447 horizontal pleiotropic path, and MR-DoC1 with unique environmental confounding (rE). The Y-

- 448 axis labels indicate the CpG probe IDs and the respective genes in which the CpGs are located.
- 449 Three of the four CpGs are in the AHRR gene and show robust evidence of the causal effects of
- 450 *current smoking on DNAm, along with weaker evidence of the reverse effects of DNAm on*
- 451 smoking. On the other hand, the fourth CpG is located in the GNG7 gene and shows robust
- 452 evidence of the causal effects of DNAm on current smoking, with weaker evidence of the reverse
- 453 *effects of smoking on DNAm.*
- 454 Note. The data underlying these plots are in **Supplementary Tables S1-S4**.
- 455

Causation between Smoking and DNA methylation

# 456 DNAm loci potentially influenced by smoking are enriched for biological 457 processes relevant to smoking's adverse health outcomes

- 458 For follow-up gene-set annotation and functional enrichment analyses<sup>25</sup>, we identified 525 CpG
- 459 sites (outside the MHC region) with potential effects of current smoking on DNAm based on
- 460 consistent, nominally significant estimates in all three models (**Supplementary Table S1**). The
- 461 genes mapped by these CpGs showed extensive significant enrichment (FDR <0.05) for ontology
- 462 clusters, including hemopoiesis, cell morphogenesis, inflammatory response, regulation of cell
- 463 differentiation, and regulation of nervous system development, underscoring DNAm's potential
- 464 role in the adverse health sequelae of smoking (**Supplementary Figures S9-S11**; **Tables S5-S6**).
- 465
- 466 Next, we performed *eFORGE 2.0* (experimentally derived Functional element Overlap analysis
- 467 of ReGions from EWAS)<sup>26,27</sup> analyses to explore the tissue-specific functional relevance of these
- 468 CpG sites. These sites were significantly enriched (FDR <0.05) for overlap with a wide range of
- 469 gene regulatory elements, including chromatin states, histone marks, and DNase-I hotspots, in
- 470 most of the tissue/cell types in reference datasets. These findings suggest that the functional
- 471 consequences of the effects of smoking on DNAm are likely widespread across the body rather
- 472 than specific to a few tissue types (Supplementary Figures S12-S14; Tables S7-S9).

# 473 CpG sites with consistent effects on current smoking show enrichment for

### 474 brain-related gene regulatory elements

- 475 For potential  $DNAm \rightarrow Current Smoking$  effects, we identified 64 CpGs with consistent,
- 476 nominally significant estimates in all three models (**Supplementary Figure S15**). In the gene-set
- 477 enrichment analyses (**Supplementary Figures S16-S17**; **Tables S10-S11**), the genes mapped by
- 478 these CpGs did not show significant functional enrichment (FDR <0.05), likely due to too few
- 479 loci implicated in this direction of causation. However, in the eFORGE analyses, which use
- 480 precise chromatin-based information for each CpG, these CpG sites showed significant
- 481 enrichment (FDR <0.05) for overlap with enhancers in the brain (fetal brain), blood (primary B
- 482 cells, hematopoietic stem cells), lung, and mesodermal embryonic stem cells (Supplementary
- 483 **Figures S18-S20**; **Tables S12-S14**). This set of CpGs also showed significant enrichment for
- 484 histone marks in multiple tissues/cell types (including the brain, blood, and lung), but the overlap
- 485 with DNase-I hotspots was not significantly enriched. The tissues/cell types predicted to be
- 486 relevant for DNAm's effects on smoking liability may be prioritized for follow-up tissue-/cell
- 487 type-specific studies.
- 488
- 489 To further gauge the tissue-specificity of the eFORGE enrichment, we performed iterative
- 490 follow-up analyses with the CpGs overlapping with tissue/cell types of interest (see Methods
- 491 and **Supplementary Figures S21-S23; Tables S15-S17**). These analyses elucidated a subset of
- 492 17 CpGs with significant and highly specific enrichment for enhancers and histone marks
- 493 (H3K4me1 and H3K4me4) in the brain (**Figure 7**), along with weaker enrichment for H3K4me1

- 494 in the adrenal gland and thymus. Ten of the 17 sites also overlapped with DNase-I hotspots in the
- 495 brain, though the enrichment was not statistically significant (FDR = 0.08) (Supplementary
- 496 Figure S24, Table S20). The causal estimates and the nearest gene of these 17 CpGs are shown
- 497 in Supplementary Figure S25. Four of these CpGs also had consistent estimates of the reverse
- 498 effects of current smoking on DNAm (identified by the column "g1\_nominal" in
- 499 **Supplementary Table S4**): cg25612391 (*SLC25A42*), cg05424060 (*GNAI1*), cg10590964 (near
- 500 *KIAA2012*), and cg05877788 (*TP53I13*). Furthermore, prior pre-clinical and clinical studies have
- 501 implicated 14 of the 17 mapped genes, including three with potential bidirectional effects, in
- 502 behavioral or neurological traits, such as alcohol dependence (*OSBPL5*)<sup>28</sup>, cocaine use
- 503  $(SLCO5A1)^{29}$ , anxiety  $(CCDC92)^{30}$ , depression  $(GNAI1)^{31}$ , encephalomyopathy and brain stress
- response (*SLC25A42*)<sup>32,33</sup>, and dementia or Alzheimer's disease pathology (*SIAH3*, *SRM*,
- 505 TP53I13)<sup>34–36</sup>.
- 506
- 507 Similar follow-up analyses with other subsets of CpGs (e.g., probe sets enriched for enhancers in
- 508 the lung or cord blood primary B cells) showed enrichment across several tissue/cell types,
- 509 suggesting non-specificity of the enrichment seen in these tissues (Supplementary Figures S26-
- 510 S31; Tables S21-S26). The enrichment for specific blood cell types (here, B cells) may be partly
- 511 confounded by residual cell-composition effects in whole blood analyses<sup>26</sup>. The 18 CpGs
- 512 overlapping with enhancers in primary B cells mapped to 16 genes, of which five have been
- 513 previously associated with (any) blood cell counts but only one with lymphocyte count in
- 514 GWAS<sup>37</sup>. Thus, the sites driving the enrichment for B cells had little overlap with the known
- 515 lymphocyte-count GWAS associations. For comparison, the 64 CpGs with potential  $DNAm \rightarrow$
- 516 *Current Smoking* effects annotated to 51 genes, of which 16 are known to be associated with
- 517 (any) blood cell counts and only two with lymphocyte count.



eFORGE Analyses of the CpG Sites with Potential Effects of DNA Methylation on Current Smoking

518

# 519 Figure 7. Among the CpG sites with potential effects of blood DNA methylation on current smoking liability, iterative eFORGE

- 520 analyses elucidated sites enriched for overlap with brain-related chromatin states and histone marks.
- 521 The first iteration of eFORGE examined the 64 CpG sites with potential effects of blood DNA methylation on current smoking liability
- 522 (Supplementary Figure S15), revealing 21 CpGs enriched for overlap with enhancers in the brain (Supplementary Figure S18/Table
- 523 S12). In follow-up analyses restricted to these 21 CpGs (eFORGE iteration 2), all 21 probes were also enriched for the brain

- 524 H3K4me1 marks, while 17 of these probes overlapped with H3K4me3 marks in the brain (Supplementary Figure S22/Table S16). This
- 525 iteration also showed significant enrichment (FDR q < 0.01) for histone marks in other tissues, including small and large intestines,
- 526 adrenal gland, and thymus. So, to identify a subset of these CpGs with potentially more specific enrichment for brain-related
- 527 functional elements, we restricted further analyses to the 17 sites overlapping with the brain H3K4me3 marks (eFORGE iteration 3).
- 528 As seen in this figure, these 17 sites showed highly specific enrichment for enhancers and histone marks in the brain (Supplementary
- 529 *Tables S18-S19*). Ten of these sites also overlapped with DNase-I hotspots in the brain (Supplementary Table S20).

Causation between Smoking and DNA methylation

#### 530 Attenuated effects of former smoking on DNAm

- 531 MR-DoC analyses estimating the causal effects between former smoking and DNAm showed
- 532 attenuated inflation factor ( $\lambda$ ) in all models, compared to the  $\lambda$  values in similar models fitted to
- 533 current smoking. For instance, the MR-DoC2 models fitted across the 11,124 smoking-
- associated CpGs had  $\lambda = 1.11$  for Former Smoking  $\rightarrow$  DNAm and  $\lambda = 0.99$  for DNAm  $\rightarrow$  Former
- 535 *Smoking*, compared to 1.20 and 1.01, respectively, for current smoking. Note that these  $\lambda$
- 536 calculations were not restricted to the former-smoking-associated CpGs to allow for a
- 537 comparison with current smoking.
- 538
- 539 Among the former-smoking-associated CpGs, only five sites showed robust evidence of causal
- 540 effects of former smoking on DNAm, with consistent, statistically significant (FDR <0.05)
- 541 causal estimates in all three models (**Supplementary Figure S32**). These CpGs include
- 542 cg05575921 in *AHRR*, cg05951221, cg01940273, and cg21566642 near *ALPPL2*, and
- 543 cg06126421 near *HLA-DRB5* gene (in the MHC region). The causal estimates at these sites are
- similar to those of the effects of current smoking on DNAm, with overlapping confidence
- 545 intervals (**Figure 8**). Thus, unlike most smoking-associated CpGs<sup>38</sup>, smoking's effects on DNAm
- 546 at these sites likely have limited reversibility, in line with the previously reported persistent
- 547 associations of these sites with former smoking 30 years after cessation<sup>2</sup>. For the reverse effects
- of DNAm on former smoking, no CpG showed consistent (at least nominally significant) causal
- 549 estimates across models (**Supplementary Figure S33**). Nevertheless, of the three CpGs with
- 550 robust evidence of DNAm's effects on current smoking, two were among the former-smoking-
- 551 associated CpGs and had overlapping confidence intervals of the estimated effects of DNAm on
- 552 *former* smoking and *current* smoking (**Supplementary Figure S34**).
- 553

Causation between Smoking and DNA methylation



#### CpGs with Putative Effects of Former Smoking on DNA Methylation

Sites with FDR < 0.05 in All Three Models

554

#### 555 Figure 8. Putative causal effects of former smoking on blood DNA methylation.

556 Estimates and Wald-type 95% confidence intervals of the causal effects of the liability for former

557 (versus never) smoking and (residualized and standardized) DNA methylation beta-values in

558 each of the three MR-DoC models: bidirectional MR-DoC2, MR-DoC1 with horizontal

559 pleiotropic path, and MR-DoC1 with unique environmental confounding (rE). The

560 corresponding estimates for current (versus never) smoking are also shown with dashed lines.

561 The text labels on the left indicate the CpG probe IDs and the genes mapped by the CpGs.

562 Note. The data underlying these plots are in Supplementary Tables S1 and S27, indicated by the

563 *column g1\_robust.* 

Causation between Smoking and DNA methylation

#### 565

### Discussion

566 Results from integrated MR and biometrical genetic (MR-DoC) modeling suggest that the causal 567 effects of cigarette smoking on blood DNAm likely underlie many of the associations seen in 568 EWAS. Compared to a handful of CpGs previously found to be causally linked with smoking in 569 standard MR studies, we found over 500 CpGs with consistent, nominally significant effects of 570 current smoking on DNAm. These CpGs show broad enrichment for tissue types and functional 571 pathways that implicate numerous well-established harmful health outcomes of smoking, 572 including cell- and neuro-development, carcinogenesis, and immune regulation. In the analyses 573 of former smoking, most of the estimated effects of smoking on DNAm were no longer 574 significant, consistent with the reversibility of smoking's effects at these loci. Additionally, 575 several CpG sites showed evidence of reverse and possibly bidirectional effects of DNAm on the 576 liability for current smoking, with a subset of these loci enriched for gene regulatory functional 577 elements in the brain. The detection of reverse or bidirectional causal effects of blood DNAm on 578 smoking highlights the potential utility of blood DNAm as a putative biomarker to monitor 579 addiction or interventions. 580 581 Previous analyses of smoking-discordant twin pairs in NTR, a subset of the current study sample, 582 found 13 CpG sites with significant DNAm differences between MZ twins discordant for current 583 smoking<sup>39</sup>, suggesting potential causality. In our MR-DoC analyses, eight of the 13 CpGs 584 showed robust evidence of causal effects of current smoking on DNAm, while none showed 585 reverse effects. Taken together, the findings from the two studies further triangulate the evidence 586 for smoking's effects on DNAm at these sites. Prior summary-statistics-based MR studies have 587 examined causality between *lifetime* (current or former) cigarette smoking and blood DNAm. The MR analyses in GoDMC<sup>18</sup> did not find evidence of causal effects of lifetime smoking on 588 589 DNAm, nor vice versa. Another study<sup>10</sup> applied a single MR method and found nominally 590 significant effects of lifetime smoking on DNAm at 11 CpG sites from the Illumina 591 MethylationEPIC array<sup>40</sup>, of which two (cg14580211, cg15212295) overlap with Illumina 450k 592 array data used in the current study. In our MR-DoC analyses, only cg14580211 showed 593 replication in the form of consistent negative causal estimates of current smoking on DNAm. 594 The novel and more extensive causal effects found in our analyses may partly be attributable to 595 the study design's ability to estimate the causal influences of *current* smoking specifically, as 596 most smoking-associated DNAm changes exhibit substantial reversibility upon smoking cessation<sup>2,21</sup>. Furthermore, the nine CpGs with previously reported reverse effects of DNAm on 597 lifetime smoking behavior (a composite index of initiation, heaviness, and cessation)<sup>11</sup> in a single 598

599 MR model showed inconsistent estimates in the three MR-DoC models. Interestingly, two of

600 these CpGs (cg09099830 and cg24033122; both located in gene *ITGAL*) instead showed

601 consistent, nomically significant effects of current smoking on DNAm, underscoring the need for

602 further replication of both prior and current findings.

Causation between Smoking and DNA methylation

604 Of the three CpG sites with robust evidence of DNAm's effects on current smoking liability, two 605 are located in genes GNG7 and RGS3 that are integral to G protein-coupled receptor (GPCR) 606 signaling, adding to the growing literature on GPCR signaling pathways' potential role in behavioral and neuropsychiatric outcomes<sup>41</sup>. Specifically, differential expressions of both 607 GNG7<sup>42</sup> and RGS3<sup>43</sup> have been associated with addiction-related phenotypes in model 608 609 organisms. The third CpG annotates to SLC15A4, which encodes a lysosomal peptide/histidine transporter involved in antigen presentation and innate immune response<sup>44</sup>, including in mast 610 611 cells<sup>45</sup>. Thus, DNAm variation at this locus may plausibly reflect individual differences in 612 immunological tolerance of cigarette smoke and, consequently, maintenance of smoking 613 behavior. Interestingly, these CpGs were significantly associated with neither cannabis use<sup>7</sup> nor 614 alcohol consumption<sup>6</sup> in recent large-scale EWASs. Notably, though, both these studies reported 615 DNAm associations conditional on cigarette smoking, making them unsuitable for gauging whether the CpGs with putative effects on smoking liability are also associated with other 616 617 substances. This raises the question of whether cigarette smoking should always be used as a 618 covariate in EWAS. If so, it may be prudent to report supplementary EWAS results without 619 smoking as a covariate, as some CpGs may have a reverse or bidirectional causal relationship with smoking. Note that the EWAS of cannabis use<sup>7</sup> did perform such preliminary analyses but 620

- only reported the results conditional on cigarette smoking.
- 622

623 Several factors need to be considered when interpreting the above results. Although we found

relatively few sites with putative effects of whole blood DNAm on smoking liability or with

625 suggestive bidirectional effects, the situation might differ in specific blood cell types or other

- tissues relevant to smoking, like the brain. The results may also vary in other peripheral tissues,
- 627 like buccal cells<sup>46</sup>. Moreover, the highly variable predictive strength of mQTL allelic scores
- across CpG sites (incremental- $R^2$  range: 0.43-76.95%; median 4.61%) likely also influenced the
- 629 power to detect true causal effects of blood DNAm on smoking liability<sup>24</sup>. When considering
- 630 similar model applications across different health traits, this impact on power is relevant to both 631 directions of causation, as the IV of other traits may not be as strong as the smoking PRS.
- 632 Additionally, the current study analyzed CpGs from the Illumina 450k array, which covers a
- 633 small fraction of genome-wide potential methylation sites. Further, many of the measured
- 634 smoking-associated CpGs lacked a "relevant" mQTL allelic score with F-statistic >10
- 635 (Supplementary Figure S35) and so are yet to be tested for  $DNAm \rightarrow Smoking$  causal effects.
- 636 Newer low-cost sequencing technology<sup>47</sup> may help uncover more such causal relationships in the
- 637 future.
- 638

639 Like all MR studies, the current results depend on the validity of the IV assumptions<sup>19</sup>, which

- 640 cannot always be empirically tested. Here, we relied on the statistical significance and
- 641 consistency of the causal estimates across different specifications of MR-DoC models to account
- 642 for potential assumption violations, particularly horizontal pleiotropy. Yet, residual bias due to
- 643 violations of the assumptions underlying MR<sup>19</sup> and biometrical twin modeling<sup>48</sup> cannot be ruled

- out with certainty. Moreover, current MR-DoC models estimated linear causal effects. However,
- since DNAm is constrained within certain biologically plausible values, the impact of smoking
- on DNAm may depend on prior DNAm. To examine such non-linear causal relationships, MR-
- 647 DoC with interaction or quadratic effects would be a valuable area of further model
- 648 development, with numerous potential applications. Finally, we examined causality using only
- binary smoking-status variables, as the sub-samples restricted to current or former smoking were
- too small to fit MR-DoC models to smoking quantity (e.g., cigarettes per day) or time since
- quitting. Further research with larger samples is needed to examine such dose-response causal
- c52 relationships.
- 653
- The current study included participants of European ancestry only. Although prior EWASs show
- highly concordant associations across ancestries<sup>2,7</sup>, examining the generalizability of causal
- 656 estimates in non-European populations is a critical subject of further research. As MR-DoC
- models provide causal inference specific to the target dataset, rather than the discovery GWAS
- 658 samples, future research may apply this study design to subpopulations of interest, e.g., stratified
- by sex or age (such as children or elderly populations), provided the results from population-
- 660 wide GWAS generalize adequately. Future applications of MR-DoC analyses to DNAm data
- 661 may also extend the current work to other health traits and disorders that show robust
- associations with DNAm and have strong genetic IVs. Recent developments in cost-effective
- 663 population-scale DNAm microarray technology<sup>49</sup> can help increase the sample sizes of twin
- 664 cohorts with DNAm data, enabling wider application of similar causal analyses.
- 665
- In conclusion, the inability to establish causality is one of the key limitations of EWAS based on
- 667 surrogate tissues such as blood. Here, we demonstrate an application of the MR-DoC design to
- examine causality between cigarette smoking and blood DNAm. The results suggest that many
- of the EWAS associations are likely driven by the causal effects of current smoking on DNAm,
- though we also find evidence of reverse and potentially bidirectional causal relationships at some
- sites. Our study highlights the value of integrating DNAm, phenotypic information, and genetic
- data in twin studies to uncover causal relationships of peripheral blood DNAm with human traits.
- This study design might be valuable for detecting causal epigenetic biomarkers of (mental)
- 674 health in general.
- 675

Causation between Smoking and DNA methylation

#### 676

# Methods

#### 677 Study Sample

The Netherlands Twin Register (NTR) is a community-based twin registry with longitudinal dataon health, behaviors, and lifestyle factors, combined with biological samples, including DNA

from blood and buccal samples. In the current analyses, we analyzed data from 2,577 individuals

participating in the NTR longitudinal surveys<sup>17</sup> and the NTR biobank project<sup>50</sup>. The study

participants comprised 1,730 (67%) female and 847 (33%) male individuals of European genetic

ancestry, including 706 monozygotic (MZ) twin pairs, 161 MZ individuals without their co-twin,
412 dizygotic (DZ) twin pairs, and 180 DZ individuals without their co-twin. The participants

had both genotypic and epigenome-wide DNAm data and were aged between 18 and 79 years

686 (mean 35.2; S.D. 11.7 years) at the time of blood sample collection.

687

688 Previous studies have described the NTR cohort in greater detail<sup>39,51</sup>. NTR genotypic sample and

689 variant quality control (QC), imputation, genetic principal component analysis (PCA), and

ancestry-outlier pruning have been described previously<sup>52</sup>. Details specific to the participants

691 included in the present study are included in the **Supplementary Methods**. Since GoDMC<sup>18</sup>

692 summary statistics are available for European ancestry only, the current study sample excluded

693 109 participants identified as European-ancestry outliers in PCA to avoid bias due to ancestry

694 mismatch. The NTR is approved by the Central Ethics Committee on Research Involving Human

695 Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board

696 certified by the U.S. Office of Human Research Protections (IRB number IRB00002991 under

697 Federal-wide Assurance- FWA00017598; IRB/institute codes, NTR 98-222, 2003-180, 2008-

698 244). All participants provided written informed consent before data collection.

#### 699 Peripheral Blood DNA Methylation and Cell Counts

700 Epigenome-wide DNAm in peripheral whole blood was measured with the Infinium

701 HumanMethylation450 BeadChip Kit (i.e., the Illumina 450k microarray), following the

702 manufacturer's protocol<sup>21</sup>. QC and normalization of the DNAm data were performed using a

703 custom pipeline developed by the BIOS (Biobank-based Integrative Omics Study) Consortium,

as previously described<sup>51</sup>. Briefly, sample QC was done using MethylAid<sup>53</sup>, followed by probe

705 OC with DNAmArray<sup>54</sup>. The latter removed the probes with a raw signal intensity of zero, bead

number <3, or a detection p-value >0.01, as well as the ambiguously mapped probes. Next,

samples and probes with >5% missingness were removed. The resulting DNAm data were

normalized using the Functional normalization<sup>55</sup> algorithm implemented in DNAmArray<sup>54</sup>, with

the first four PCs (with eigenvalue >1) derived from control probes. Finally, the probes

containing a SNP within the CpG site (at C or G nucleotide) were removed regardless of the

711 minor allele frequency. These SNPs were previously identified using DNA sequencing data from

the Dutch population<sup>56</sup>. For the current analyses, only autosomal probes were included, yielding

Causation between Smoking and DNA methylation

- 411,169 CpG sites that passed all QC metrics, of which 16,940 sites were reported as associated
- vith current smoking (FDR <0.05) in an independent EWAS<sup>2</sup>. Differential white blood cell
- counts were also measured in the blood samples to estimate the proportions of neutrophils,
- 716 lymphocytes, monocytes, eosinophils, and basophils<sup>51</sup>.
- 717
- 718 Using linear regression models, the normalized  $\beta$ -values of DNAm at each CpG were corrected
- 719 for commonly used EWAS covariates<sup>57</sup>, including age at blood draw, sex (genotypically inferred
- biological sex, matched with self-reported gender), measured white blood cell percentages
- 721 (neutrophils, monocytes, and eosinophils) at blood draw, MH450k array row, and bisulfite
- sample plate (dummy variables). The residuals from these regression models were standardized
- (mean = 0, S.D. = 1) and used in MR-DoC models. As in the previous work in this dataset<sup>39</sup>, we
- did not include lymphocyte percentage as a regression covariate to prevent multicollinearity with
- neutrophil percentage, while basophil percentage was not included because it had little variation
- 726 between individuals.

# 727 Cigarette Smoking

- 728 Self-reported cigarette smoking status was recorded through an interview during the home visit
- for blood sample collection in 2004-2008 and 2010-2011. Participants were asked, "Do you
- smoke?" with one of three possible answers: "No, I never smoked" (N = 1,492), "No, but I did in
- the past" (N = 549), and "Yes" (N = 528). See **Supplementary Methods** for the original (N = 549), (N = 549), (N = 528).
- wording in Dutch. Those endorsing current smoking were asked how many years they had been
- smoking and how many cigarettes or rolling tobacco they smoked per day. Those endorsing
  former smoking were asked how many years ago they guit smoking, how many years they had
- rother shoking were asked now many years ago mey duit shoking, now many years mey nad smoked before quitting, and the maximum number of cigarettes or rolling tobacco they used to
- sinoked before quitting, and the maximum number of eigarctics of forming tobacco they used to smoke per day. The responses were checked for consistency with the information from the NTR
- 737 longitudinal surveys filled out closest to blood sampling. As previously described<sup>58</sup>, potential
- 738 misclassification of smoking status through self-reports was evaluated based on plasma cotinine
- the sample. Of the 591 individuals with self-reported never smoking and measured plasma
- cotinine, only five (0.8%) had cotinine levels indicative of smoking ( $\geq 15$  ng/ml), thus suggesting
- 742 low misclassification of smoking status. The number of individuals endorsing current or former
- smoking was too small to evaluate a dose-response relationship of the causal effects in MR-DoC
- models restricted to currently or formerly smoking individuals. Likewise, the sample with former
- smoking was too small to examine the effect of "time since quitting smoking" on DNAm.

# 746 Instrumental Variables

- 747 *mQTL allelic scores.* We identified 12,940 smoking-associated CpGs with *cis*-mQTL summary
- statistics available from GoDMC<sup>18</sup> (excluding NTR), using GoDMC's definition of "*cis*" interval
- 749 (within 1Mb of the CpG). In GoDMC, the contributing cohorts performed genome-wide mQTL
- analyses, testing the associations of ~480,000 CpG sites with ~12 million SNPs. However,

Causation between Smoking and DNA methylation

before the meta-analysis, the cohort-level results were filtered to retain the SNP-CpG pairs with

- $p < 1 \times 10^{-5}$  within the cohort. Thus, since the summary statistics were already partly
- thresholded, we computed the mQTL allelic scores by applying clumping and thresholding in
- 754 *PLINK1.9*<sup>59</sup>. Linkage disequilibrium (LD)-based clumping was performed using --clump-pl
- 755 1 --clump-kb 250, with two levels of LD  $r^2$  (0.5 and 0.1) specified for --clump-r2, thus
- yielding two sets of LD-clumped *cis*-SNPs. Using either set of SNPs, we computed the allelic
- score with --score at a threshold of 0.05 (applied with --q-score-range). If none of the
- SNPs had p <0.05, no threshold was applied for score calculation. An additional allelic score was
- calculated using the top *cis*-mQTL (with the minimum association p-value) for each CpG. Thus,
   for every CpG, three scores were calculated (two LD-clumped mQTL allelic scores, plus the top-
- 761 mQTL), though these scores were not necessarily distinct; for example, if a CpG had only one
- 762 *cis*-SNP, all three criteria yielded the same score. Likewise, for some CpGs, the two LD-
- clumping cut-offs resulted in the same set of SNPs and, hence, identical mQTL allelic scores.
- 764

To assess the strength of an mQTL allelic score, we first estimated its incremental  $R^2$  by fitting generalized estimating equations (GEE), controlling for the standard EWAS covariates (as above), genotyping platform, and the first ten genetic PCs. For each CpG, the effective GEE sample size ( $N_{Eff}$ ) was computed using the following formulae:

- 769
- 770  $N_{Eff}^{MZ} = \frac{2 * N_{MZ}}{1 + r_{MZ}}$
- 771

772 
$$N_{Eff}^{DZ} = \frac{2 * N_{DZ}}{1 + r_{DZ}}$$

- 773 774
- $N_{Eff} = N_{Eff}^{MZ} + N_{Eff}^{DZ} + N_{Ind}$
- 775

where,  $N_{Eff}^{MZ}$  and  $N_{Eff}^{DZ}$  are the estimated effective sample sizes of MZ and DZ twins,  $N_{MZ}$  and  $N_{DZ}$  are the numbers of complete MZ and DZ twin pairs, while  $r_{MZ}$  and  $r_{DZ}$  are the twin phenotypic (DNAm) correlations in MZ and DZ twin pairs, respectively.  $N_{Ind}$  is the number of individuals without the co-twin. The estimated effective sample size was then used to transform the incremental R<sup>2</sup> value into an F-statistic as: 781

782

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

783

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

Causation between Smoking and DNA methylation

787 **PRS of Regular Smoking Initiation.** We used European-ancestry GWAS summary statistics for 788 smoking initiation (i.e., initiation of regular smoking) from the GWAS & Sequencing 789 Consortium of Alcohol and Nicotine use (GSCAN; excluding NTR)<sup>20</sup> to compute the PRS of smoking in NTR using LDpred v0.9<sup>60</sup>. Of note, the phenotypic definition in the GWAS (smoking 790 791 initiation = current/former versus never smoking) was different from the smoking phenotypes 792 (current versus never and former versus never smoking) in the MR-DoC models. However, in 793 these causal models, the strength of the IV, the extent of horizontal pleiotropy with DNAm, and 794 the estimated causal effects on DNAm are specific to the smoking phenotype used in the models. 795 As a result, this approach allowed us to assess the causal relationships of DNAm with current 796 and former smoking separately. See Supplementary Methods for a detailed description of PRS 797 calculation and estimation of incremental R<sup>2</sup>. Using linear regression models, we residualized the 798 PRS of smoking and all mQTL allelic scores for the genotyping platform and the first ten genetic

- PCs. The residuals were scaled to have a mean of zero and a variance of one before being
- 800 included as IVs in MR-DoC models.

#### 801 MR-DoC Models

802 Causal inference in the twin *Direction-of-Causation* models uses the differences in cross-twin

- 803 cross-trait correlations under different directions of causation to identify the model that fits the
- data best<sup>15</sup>. On the other hand, MR analyses rely on three assumptions of a valid  $IV^{3,19}$ , that the
- 805 IV is (1) associated with the exposure ("relevance"), (2) not correlated with any omitted
- 806 confounding variables ("exchangeability"), and (3) independent of the outcome, given the 807 exposure ("exclusion restriction"). Here, we used the criterion of F-statistic >10 to define the
- "relevance" of the IV. Further, germline genetic variants are often assumed to satisfy the
- 809 "exchangeability" assumption due to Mendel's laws of random segregation and independent
- 810 assortment. The "exclusion restriction" assumption for a genetic IV implies no horizontal
- 811 pleiotropy with the outcome. As described above, we relied on different specifications of MR-
- 812 DoC models to account for potential horizontal pleiotropy. MR-DoC1 model allowed estimating
- and controlling for horizontal pleiotropy from the IV to the outcome, though it required us to fix
- the unique environmental confounding at a specific value (here, zero)<sup>13</sup>. MR-DoC2 model
- 815 leverages the covariance between two polygenic or multiallelic IVs, beyond the bidirectional
- 816 causal effects, to partly accommodate horizontal pleiotropy $^{14}$ .
- 817
- 818 We used the *OpenMx* (version 2.21.8)<sup>61</sup> package in R (version 4.3.2) to fit the MR-DoC models, 819 using the code provided in the original publications<sup>13,14</sup>. Binary smoking status was examined 820 under the liability threshold model<sup>62</sup>, assuming a latent liability distribution with its mean fixed 821 at zero and variance fixed at one, while the threshold was freely estimated.
- 822
- 823 Before fitting the MR-DoC models, we examined univariate ACE twin models of smoking status
- to estimate the additive genetic (A), shared environmental (C), and unique environmental (E)
- 825 variance components of the latent liability scale, with age and sex as covariates. Maximum-

Causation between Smoking and DNA methylation

876	likelihood tetracharic correlation estimates for current versus never smoking were: $r_{\rm c} = 0.025$
820	(S $E = 0.021$ ) in MZ pairs and $r_{max} = 0.533$ (S $E = 0.083$ ) in DZ pairs. Likewise, former
878	$(5.20.021)$ in WZ pairs and $r_{DZ} = 0.555 (5.20.005)$ in DZ pairs. Electric, former versus never smoking had $r_{DZ} = 0.822 (5.20.038)$ and $r_{DZ} = 0.474 (5.20.096)$ . Based
820	on likelihood-ratio tests (LRT) on AE twin model was the most parsimonious model for both
820	on incernious ratio tests (ECT), an AE twin model was the most parametrized noder for both our current versus power (AE versus ACE LPT $n = 0.417$ ) and former versus power smoking (AE
03U 021	current versus never (AE versus ACE LR1 $p = 0.417$ ) and former versus never smoking (AE
021	versus ACE LRT $p = 0.550$ (Supplementary Table 551). The estimated variance components of summary versus power smalling lightlifty were $A = 0.027$ (maximum likelihood 05% confidence
032 022	of current versus never smoking hability were $A = 0.927$ (maximum-likelihood 95% confidence intervals 0.870, 0.050) and $E = 0.072$ (0.041, 0.121). The corresponding estimates of former
833	interval: $0.879, 0.959$ and $E = 0.073 (0.041, 0.121)$ . The corresponding estimates of former
834	versus never smoking were $A = 0.827 (0.745, 0.888)$ and $E = 0.173 (0.112, 0.255)$ .
835	
836	Prior twin analyses of DNAm at CpG sites in NTR <sup>31</sup> showed that, of the 411,169 autosomal post-
837	QC CpG sites, the AE twin model was the best fitting model at all but 426 sites, with significant
838	(after multiple-testing correction of LRT p-values) C variance at 185 sites and significant non-
839	additive genetic (D) variance at 241 sites. Of the smoking-associated CpGs <sup>2</sup> , only two CpGs had
840	significant estimates of C, while only seven CpGs had significant estimates of D. Thus, in MR-
841	DoC models, we specified an AE variance decomposition of DNAm at all smoking-associated
842	CpGs. Note that, in the results presented above, none of the CpG sites with consistent, nominally
843	significant estimates of causal effects in either direction (525 sites with <i>current smoking</i> $\rightarrow$
844	DNAm; 64 sites with $DNAm \rightarrow current \ smoking$ ) have significant C or D estimates per the
845	previous univariate twin analyses <sup>51</sup> . Moreover, since smoking status liability also has an AE
846	variance decomposition, including a C or D variance component of DNAm in the model would
847	not change the possible sources of covariance between smoking status and DNAm in the model.
848	
849	We fitted five sets of MR-DoC models with current versus never smoking and similar sets with
850	former versus never smoking (Figure 1): (1) Smoking $\rightarrow$ DNAm MR-DoC1 with horizontal
851	pleiotropy, (2) Smoking $\rightarrow$ DNAm MR-DoC1 with unique environmental confounding, (3)
852	$DNAm \rightarrow Smoking$ MR-DoC1 with horizontal pleiotropy, (4) $DNAm \rightarrow Smoking$ MR-DoC1
853	with unique environmental confounding, and (5) bidirectional MR-DoC2. Each model included
854	age and sex as covariates of smoking status. In each model, the residual variance of smoking
855	status liability is decomposed into $a_s^2$ (A) and $e_s^2$ (E), while that of DNAm is decomposed into
856	$a_D^2$ (A) and $e_D^2$ (E). The correlation between the latent A factors of smoking and DNAm (rA)
857	represents the confounding due to additive genetic factors. The correlation between the latent E
858	factors (rE) represents the confounding due to unique environmental factors. Across all models,
859	the causal path from smoking to DNAm is labeled $g_1$ , while that from DNAm to smoking is
860	labeled g <sub>2</sub> . The residualized PRS and mQTL allelic scores are regressed on respective latent
861	factors, representing the underlying "true" standardized scores with mean fixed at zero and

variance fixed at one. The coefficient of the path from the latent score to the observed score

863 estimates the standard deviation of the observed score ( $SD_{PRS}$  and  $SD_{mQTL}$ , respectively).

Causation between Smoking and DNA methylation

- 865 Thus, for each CpG site included in the analyses, three causal estimates were obtained in either
- 866 direction (*Smoking*  $\rightarrow$  *DNAm*, or *DNAm*  $\rightarrow$  *Smoking*) from (1) MR-DoC1 with horizontal
- 867 pleiotropy, (2) MR-DoC1 with unique environmental confounding, and (3) MR-DoC2. For each
- 868 set of causal estimates across CpG sites, we calculated the Bayesian inflation factor ( $\lambda$ ) using the
- 869 R package  $bacon^{23}$ , made QQ plots using the R package  $GWASTools^{63}$ , and then applied
- 870 Benjamini-Hochberg FDR correction<sup>64</sup> to the p-values using the R package *qvalue*<sup>65</sup>. For
- 871 Bonferroni multiple-testing correction, the significance level was defined as  $\alpha = 0.05/16940 =$
- 872  $2.95 \times 10^{-6}$  for *Current Smoking*  $\rightarrow$  *DNAm* MR-DoC1 models and  $\alpha = 0.05/11124 =$
- 873  $4.49 \times 10^{-6}$  for *DNAm*  $\rightarrow$  *Current Smoking* MR-DoC1 and bidirectional current-smoking MR-
- 874 DoC2 models.

#### 875 Functional Enrichment Analyses

876 We used Metascape<sup>25</sup> (v3.5.20240101; <u>https://metascape.org/gp/index.html#/main/step1</u>, with

- the default settings for "Express" analyses) to perform gene-set annotation and functional
- 878 enrichment analyses of the CpGs with potential causal effects in either direction. The input list of
- gene IDs was selected based on proximity to the CpGs with consistent and nominally significant
- 880 (p <0.05) estimates in all three models; i.e., 64 CpGs with potential  $DNAm \rightarrow Current Smoking$
- effects ("Nearest Gene" in **Supplementary Table S3**) and 525 CpGs with potential *Current*
- 882  $Smoking \rightarrow DNAm$  effects ("Nearest Gene" in **Supplementary Table S1**). None of the sites with
- potential  $DNAm \rightarrow Current Smoking$  effects are located in the MHC region. For *Current*
- 884  $Smoking \rightarrow DNAm$  effects, 21 additional sites in the MHC region showed consistent, nominally 885 significant estimates. There was no significant relationship between a CpG site having consistent
- causal estimates and its being located in the MHC region (Fisher's exact test p-value = 0.5455).
- However, out of an abundance of caution, the sites located in this region were not included in the
- enrichment analyses to avoid sites with potentially unreliable results due to its complex LD
- 889 structure.
- 890
- 891 As described in the Metascape manuscript<sup>25</sup>, the program performed integrated enrichment
- analyses against multiple reference ontology knowledgebases, including GO processes<sup>66</sup>, KEGG
- pathways<sup>67</sup>, canonical pathways<sup>68</sup>, and Reactome gene sets<sup>69</sup>. The significant terms with a
- hypergeometric p-value <0.01 and >1.5-fold enrichment were clustered into a hierarchical tree
- based on Kappa-statistical similarities among their gene memberships. The tree was then cast
- 896 into clusters based on a threshold of 0.3 kappa score to obtain enriched, non-redundant ontology
- terms.

# 898 eFORGE (experimentally derived Functional element Overlap analysis of 899 ReGions from EWAS)

- 900 We performed  $eFORGE 2.0^{26,27,70}$  analyses of the selected CpG probe IDs with consistent and
- 901 nominally significant (p <0.05) estimates in either direction (from Supplementary Tables S1,

Causation between Smoking and DNA methylation

902 **S3**). Using the web-based tool (https://eforge.altiusinstitute.org/), we examined the overlap 903 between the implicated CpGs and multiple comprehensive reference sets of genomic and 904 epigenomic features that regulate gene expression in different tissues and cell types. The 905 platform was set as "Illumina 450k array", with default analysis options: proximity = 1kb 906 window, background repetitions = 1000, and significance thresholds of FDR <0.01 (strict) and 907 FDR <0.05 (marginal). Three sets of analyses were performed for each list of probe IDs, 908 selecting the reference data from "Consolidated Roadmap Epigenomics - Chromatin - All 15-909 state marks", "Consolidated Roadmap Epigenomics - DHS", and "Consolidated Roadmap 910 Epigenomics - All H3 marks".

911

912 The eFORGE results include the specific probe IDs overlapping between the input set and the

- 913 reference sample. We performed iterative follow-up analyses for the CpGs with potential *DNAm*
- 914  $\rightarrow$  *Current Smoking* effects, based on the overlapping probe IDs to examine the specificity of
- significant (FDR <0.01) enrichment in tissues of interest. Analyses restricted to the 21 CpGs
- 916 overlapping with enhancers in the fetal brain (Supplementary Figure S18, Table S12) showed
- 917 significant enrichment only for enhancers in the fetal brain samples, suggesting high specificity
- 918 (Supplementary Figure S21). The histone mark analyses also showed enrichment in the fetal
- 919 brain (though not specific to the brain), wherein all 21 CpGs overlapped with H3K4me1, while a
- subset of 17 CpGs overlapped with H3K4me3 (**Supplementary Figure S22**). Finally, we
- 921 performed analyses restricted to these 17 CpGs.
- 922

923 We performed similar follow-up analyses with probe IDs showing overlap with enhancers in the

924 lung (potentially etiologically relevant tissue) and the primary B-cells in cord blood (the tissue

925 type with the most significant enrichment) (from **Supplementary Figure S18, Table S12**). We

also examined the overlap between the CpGs with potential  $DNAm \rightarrow Current Smoking$  effects

927 and the genes implicated in the GWAS of blood cell counts<sup>37</sup> to probe the potential impact of the

928 cell-count GWAS associations on the causal inference and cell-type enrichment. Similar overlap

929 was examined for the subset of CpGs overlapping with enhancers in cord blood primary B cells.

Causation between Smoking and DNA methylation

931		References
932 933	1.	Wei, S. et al. Ten Years of EWAS. Adv. Sci. 8, 2100727 (2021).
934	2.	Joehanes, R. et al. Epigenetic Signatures of Cigarette Smoking. Circ. Cardiovasc. Genet. 9,
935		436–447 (2016).
936	3.	Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian
937		randomization: using genes as instruments for making causal inferences in epidemiology.
938		Stat Med 27, 1133–63 (2008).
939	4.	Zillich, L. et al. Epigenetic Signatures of Smoking in Five Brain Regions. J. Pers. Med. 12,
940		566 (2022).
941	5.	Hannon, E. et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for
942		co-localization of genetic associations and differential DNA methylation. Genome Biol. 17,
943		176 (2016).
944	6.	Lohoff, F. W. <i>et al.</i> Epigenome-wide association study of alcohol consumption in $N = 8161$
945		individuals and relevance to alcohol use disorder pathophysiology: identification of the
946		cystine/glutamate transporter SLC7A11 as a top target. Mol. Psychiatry 27, 1754–1764
947		(2022).
948	7.	Fang, F. et al. Trans-ancestry epigenome-wide association meta-analysis of DNA
949		methylation with lifetime cannabis use. Mol. Psychiatry (2023) doi:10.1038/s41380-023-
950		02310-w.
951	8.	Dhana, K. et al. An Epigenome-Wide Association Study of Obesity-Related Traits. Am. J.
952		<i>Epidemiol.</i> <b>187</b> , 1662–1669 (2018).

953	9. Davey Smith, G. & Ebrahim, S. Mendelian randomization: Can genetic epidemiology
954	contribute to understanding environmental determinants of disease? Int J Epidemiol 32, 1-22
955	(2003).
956	10. Sun, YQ. et al. Assessing the role of genome-wide DNA methylation between smoking and
957	risk of lung cancer using repeated measurements: the HUNT study. Int. J. Epidemiol. 50,
958	1482–1497 (2021).
959	11. Jamieson, E. et al. Smoking, DNA Methylation, and Lung Function: A Mendelian
960	Randomization Analysis to Investigate Causal Pathways. Am. J. Hum. Genet. 106, 315–326
961	(2020).
962	12. Burgess, S., Thompson, S. G. & CRP CHD Genetics Collaboration. Avoiding bias from
963	weak instruments in Mendelian randomization studies. Int J Epidemiol 40, 755–64 (2011).
964	13. Minică, C. C., Dolan, C. V., Boomsma, D. I., De Geus, E. & Neale, M. C. Extending
965	Causality Tests with Genetic Instruments: An Integration of Mendelian Randomization with
966	the Classical Twin Design. Behav. Genet. 48, 337–349 (2018).
967	14. Castro-de-Araujo, L. F. S. et al. MR-DoC2: Bidirectional Causal Modeling with
968	Instrumental Variables and Data from Relatives. Behav. Genet. 53, 63–73 (2023).
969	15. Heath, A. C. et al. Testing hypotheses about direction of causation using cross-sectional
970	family data. Behav. Genet. 23, 29–50 (1993).
971	16. Minică, C. C., Boomsma, D. I., Dolan, C. V., De Geus, E. & Neale, M. C. Empirical
972	comparisons of multiple Mendelian randomization approaches in the presence of assortative
973	mating. Int. J. Epidemiol. 49, 1185–1193 (2020).
974	17. Ligthart, L. et al. The Netherlands Twin Register: Longitudinal Research Based on Twin and
975	Twin-Family Designs. Twin Res. Hum. Genet. 22, 623-636 (2019).
	42

- 976 18. Min, J. L. et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA
- 977 methylation. *Nat. Genet.* **53**, 1311–1321 (2021).
- 978 19. Richmond, R. C. & Davey Smith, G. Mendelian Randomization: Concepts and Scope. Cold
- 979 Spring Harb Perspect Med (2021) doi:10.1101/cshperspect.a040501.
- 980 20. Saunders, G. R. B. et al. Genetic diversity fuels gene discovery for tobacco and alcohol use.
- 981 *Nature* **612**, 720–724 (2022).
- 982 21. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution.
  983 *New Genomic Technol. Appl.* 98, 288–295 (2011).
- 984 22. Falconer, D. S. Introduction To Quantitative Genetics. (Oliver and Boyd, Edinburgh, 1960).
- 985 23. van Iterson, M., van Zwet, E. W., Heijmans, B. T., & the BIOS Consortium. Controlling bias
- and inflation in epigenome- and transcriptome-wide association studies using the empirical
  null distribution. *Genome Biol.* 18, 19 (2017).
- 988 24. Castro-de-Araujo, L. F. et al. Power, measurement error, and pleiotropy robustness in twin-
- design extensions to Mendelian Randomization. *Research square* rs.3.rs-3411642 Preprint at
- 990 https://doi.org/10.21203/rs.3.rs-3411642/v1 (2023).
- 25. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systemslevel datasets. *Nat. Commun.* 10, (2019).
- 993 26. Breeze, C. E. et al. eFORGE: A Tool for Identifying Cell Type-Specific Signal in
- 994 Epigenomic Data. Cell Rep. 17, 2137–2150 (2016).
- 27. Breeze, C. E. *et al.* eFORGE v2.0: updated analysis of cell type-specific signal in
  epigenomic data. *Bioinformatics* 35, 4767–4769 (2019).
- 28. Edenberg, H. J. et al. Genome-Wide Association Study of Alcohol Dependence Implicates a
- 998 Region on Chromosome 11. *Alcohol. Clin. Exp. Res.* **34**, 840–852 (2010).

- 999 29. Khan, A. H. et al. Genetic pathways regulating the longitudinal acquisition of cocaine self-
- administration in a panel of inbred and recombinant inbred mice. Cell Rep. 42, 112856
- 1001 (2023).
- 1002 30. Jin, X., Dong, S., Yang, Y., Bao, G. & Ma, H. Nominating novel proteins for anxiety via
- 1003 integrating human brain proteomes and genome-wide association study. J. Affect. Disord.
- **358**, 129–137 (2024).
- 31. Sarkar, A. *et al.* Hippocampal HDAC4 Contributes to Postnatal Fluoxetine-Evoked
  Depression-Like Behavior. *Neuropsychopharmacology* **39**, 2221–2232 (2014).
- 1007 32. Aldosary, M. et al. SLC25A42-associated mitochondrial encephalomyopathy: Report of
- additional founder cases and functional characterization of a novel deletion. *JIMD Rep.* 60,
  75–87 (2021).
- 1010 33. Stankiewicz, A. M., Jaszczyk, A., Goscik, J. & Juszczak, G. R. Stress and the brain
- 1011 transcriptome: Identifying commonalities and clusters in standardized data from published
- 1012 experiments. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **119**, 110558 (2022).
- 1013 34. Cochran, J. N. et al. Genetic associations with age at dementia onset in the PSEN1 E280A
- 1014 Colombian kindred. *Alzheimers Dement.* **19**, 3835–3847 (2023).
- 1015 35. Mahajan, U. V. et al. Dysregulation of multiple metabolic networks related to brain
- 1016 transmethylation and polyamine pathways in Alzheimer disease: A targeted metabolomic
- and transcriptomic study. *PLOS Med.* **17**, e1003012 (2020).
- 1018 36. Blanco-Luquin, I. et al. Early epigenetic changes of Alzheimer's disease in the human
- 1019 hippocampus. *Epigenetics* **15**, 1083–1092 (2020).
- 1020 37. Vuckovic, D. et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell
- 1021 **182**, 1214-1231.e11 (2020).

- 1022 38. Dugué, P.-A. et al. Smoking and blood DNA methylation: an epigenome-wide association
- study and assessment of reversibility. *Epigenetics* **15**, 358–368 (2020).
- 1024 39. van Dongen, J. et al. Effects of smoking on genome-wide DNA methylation profiles: A
- study of discordant and concordant monozygotic twin pairs. *eLife* **12**, e83286 (2023).
- 1026 40. Pidsley, R. et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray
- 1027 for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
- 41. Wong, T.-S. *et al.* G protein-coupled receptors in neurodegenerative diseases and psychiatric
  disorders. *Signal Transduct. Target. Ther.* 8, 177 (2023).
- 1030 42. Stankiewicz, A. M. et al. Novel candidate genes for alcoholism transcriptomic analysis of
- 1031 prefrontal medial cortex, hippocampus and nucleus accumbens of Warsaw alcohol-preferring
- and non-preferring rats. *Pharmacol. Biochem. Behav.* **139**, 27–38 (2015).
- 1033 43. Burchett, S. A., Bannon, M. J. & Granneman, J. G. RGS mRNA Expression in Rat Striatum.
- 1034 J. Neurochem. 72, 1529–1533 (1999).
- 1035 44. Chen, X. et al. Structural basis for recruitment of TASL by SLC15A4 in human
- 1036 endolysosomal TLR signaling. *Nat. Commun.* **14**, 6627 (2023).
- 1037 45. Kobayashi, T. et al. Lysosome biogenesis regulated by the amino-acid transporter SLC15A4
- 1038 is critical for functional integrity of mast cells. *Int. Immunol.* **29**, 551–566 (2017).
- 1039 46. Teschendorff, A. E. et al. Correlation of Smoking-Associated DNA Methylation Changes in
- 1040 Buccal Cells With DNA Methylation Changes in Epithelial Cancer. JAMA Oncol. 1, 476–
- 1041 485 (2015).
- 1042 47. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. Nat.
- 1043 *Methods* **14**, 407–410 (2017).

- 1044 48. Evans, D. M., Gillespie, N. A. & Martin, N. G. Biometrical genetics. Biol. Psychol. 61, 33–
- 1045 51 (2002).
- 1046 49. Illumina, Inc. Infinium Methylation Screening Array-48 Kit. (2024).
- 1047 50. Willemsen, G. et al. The Netherlands Twin Register Biobank: A Resource for Genetic
- 1048 Epidemiological Studies. *Twin Res. Hum. Genet.* **13**, 231–245 (2010).
- 1049 51. van Dongen, J. et al. Genetic and environmental influences interact with age and sex in
- shaping the human methylome. *Nat. Commun.* **7**, 11115 (2016).
- 1051 52. Singh, M. et al. Using Instrumental Variables to Measure Causation over Time in Cross-
- 1052 Lagged Panel Models. *Multivar. Behav. Res.* **59**, 342–370 (2024).
- 1053 53. van Iterson, M. *et al.* MethylAid: visual and interactive quality control of large Illumina 450k
  1054 datasets. *Bioinformatics* 30, 3435–3437 (2014).
- 1055 54. Sinke, L., van Iterson, M., Cats, D., Slieker, R. & Heijmans, B. DNAmArray: Streamlined
- 1056 workflow for the quality control, normalization, and analysis of Illumina methylation array
- 1057 data. Zenodo https://doi.org/10.5281/zenodo.3355292 (2019).
- 1058 55. Fortin, J.-P. et al. Functional normalization of 450k methylation array data improves
- 1059 replication in large cancer studies. *Genome Biol.* **15**, 503 (2014).
- 1060 56. Francioli, L. C. et al. Whole-genome sequence variation, population structure and
- 1061 demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
- 1062 57. van Rooij, J. et al. Evaluation of commonly used analysis strategies for epigenome- and
- 1063 transcriptome-wide association studies through replication of large-scale population studies.
- 1064 *Genome Biol.* **20**, 235 (2019).
- 1065 58. van Dongen, J. et al. DNA methylation signatures of educational attainment. Npj Sci. Learn.
- 1066 **3**, 7 (2018).

- 1067 59. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer
- 1068 datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).
- 1069 60. Vilhjálmsson, J. et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic
- 1070 Risk Scores. Am. J. Hum. Genet. 97, 576–592 (2015).
- 1071 61. Neale, M. C. et al. OpenMx 2.0: Extended Structural Equation and Statistical Modeling.
- 1072 *Psychometrika* **81**, 535–549 (2016).
- 1073 62. Verhulst, B. & Neale, M. C. Best Practices for Binary and Ordinal Data Analyses. *Behav*.
- 1074 *Genet.* **51**, 204–214 (2021).
- 1075 63. Gogarten, S. M. et al. GWASTools: an R/Bioconductor package for quality control and
- analysis of genome-wide association studies. *Bioinformatics* **28**, 3329–3331 (2012).
- 1077 64. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
- 1078 Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300
- 1079 (1995).
- 1080 65. Storey, J., Bass, A., Dabney, A. & Robinson, D. qvalue: Q-value estimation for false
  1081 discovery rate control. doi:10.18129/B9.bioc.qvalue. (2023).
- 1082 66. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–
  1083 29 (2000).
- 1084 67. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids*1085 *Res.* 28, 27–30 (2000).
- 1086 68. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for
- 1087 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550
- 1088 (2005).

Causation between Smoking and DNA methylation

- 1089 69. Fabregat, A. et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 46, D649–
- 1090 D655 (2018).
- 1091 70. Breeze, C. E. Cell Type-Specific Signal Analysis in Epigenome-Wide Association Studies.
- 1092 in Epigenome-Wide Association Studies: Methods and Protocols (ed. Guan, W.) 57–71
- 1093 (Springer US, New York, NY, 2022). doi:10.1007/978-1-0716-1994-0\_5.