Appraising the causal relevance of DNA methylation for risk of lung cancer

Thomas Battram[1,2λ*], Rebecca C Richmond[1,2λ], Laura Baglietto[3†], Philip C Haycock[1,2†], Vittorio Perduca[4], Stig E Bojesen[5,6,7], Tom R Gaunt[1,2], Gibran Hemani[1,2], Florence Guida[8], Robert Carreras-Torres[8], Rayjean Hung[9], Christopher I Amos[10], Joshua R Freeman[11], Torkjel M Sandanger[12], Therese H Nøst[13], Børge Nordestgaard[5,6,7], Andrew E Teschendorff[14,15,16], Silvia Polidoro[17], Paolo Vineis[17,18], Gianluca Severi[19,20,21,22], Allison M Hodge[21,22], Graham G Giles[21,22], Kjell Grankvist[23], Mikael B Johansson[24], Mattias Johansson[8], George Davey Smith[1,2$], Caroline L Relton[1,2$]

1. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

2. Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

3. Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy

4. Laboratoire de Mathématiques Appliquées – MAP5 (UMR CNRS 8145), Université Paris Descartes

5. Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark

6. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

7. The Copenhagen City Heart Study, Frederiksberg Hospital, Copenhagen University Hospital, Copenhagen, Denmark

8. Genetic Epidemiology Division, International Agency for Research on Cancer, Lyon, France

9. Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada

10. Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, USA

11. Department of Biostatistics and Epidemiology, University of Massachusetts, Massachusetts, USA

12. Department of Community Medicine, UiT- The Arctic University of Norway, Tromso, Norway

13. Department of Public Health and Nursing, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

14. Department of Women's Cancer, Institute for Women's Health, University College London, London, UK

15. UCL Cancer Institute, University College London, London, UK

16. Chinese Academy of Sciences (CAS) Key Laboratory of Computational Biology, CAS–Max Planck Gesellschaft (MPG) Partner Institute for Computational Biology, Shanghai, China

17. Italian Institute for Genomic Medicine, Torino, Italy

18. Department of Epidemiology and Biostatistics, the School of Public Health, Imperial College London, London, UK

19. Centre de Recherche en Epidémiologie et Santé des Populations – CESP (UMR INSERM 1018), Université Paris-Saclay, Université Paris-Sud, Paris, France

20. Gustave Roussy, Villejuif, France

21. Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, Australia

22. Centre for Epidemiology and Biostatistics, Melbourne School of Population & Global Health, The University of Melbourne, Australia

23. Department of Biobank Research, Umeå University, Sweden

24. Department of Radiation Sciences, Umeå University, Sweden

[λ]joint first authors

[†]joint second authors

[$]joint last authors

*Corresponding author:

Email address: tb13101@bristol.ac.uk

Telephone: 01173314043

Address: Oakfield House, Bristol, BS8 2BN

Word count: 3065

Abstract

Background

DNA methylation changes in peripheral blood have recently been identified in relation to lung cancer risk. Some of these changes have been suggested to mediate part of the effect of smoking on lung cancer. However, limitations with conventional mediation analyses mean that the causal nature of these methylation changes has yet to be fully elucidated.

Methods

We first performed a meta-analysis of four epigenome-wide association studies (EWAS) of lung cancer (918 cases, 918 controls). Next, we conducted a two-sample Mendelian randomization analysis, using genetic instruments for methylation at CpG sites identified in the EWAS meta-analysis, and 29,863 cases and 55,586 controls from the TRICL-ILCCO lung cancer consortium, to appraise the possible causal role of methylation at these sites on lung cancer.

Results

16 CpG sites were identified from the EWAS meta-analysis (FDR < 0.05), 14 of which we could identify genetic instruments for. Mendelian randomization provided little evidence that DNA methylation in peripheral blood at the 14 CpG sites play a causal role in lung cancer development (FDR>0.05), including for cg05575921-*AHRR* where methylation is strongly associated with both smoke exposure and lung cancer risk.

Conclusions

The results contrast with previous observational and mediation analysis, which have made strong claims regarding the causal role of DNA methylation. Thus, previous suggestions of a mediating role of methylation at sites identified in peripheral blood, such as cg05575921-*AHRR*, could be unfounded. However, this study does not preclude the possibility that differential DNA methylation at other sites is causally involved in lung cancer development, especially within lung tissue.

Keywords: Lung cancer, DNA methylation, Mendelian randomization, ALSPAC, ARIES

Key Messages

- DNA methylation is a modifiable biomarker, giving it the potential to be targeted for intervention in many diseases, including lung cancer which is the most common cause of cancer-related death.
- This Mendelian randomization study attempted to evaluate whether there was a causal relationship, and thus potential for intervention, between DNA methylation measured in peripheral blood and lung cancer by assessing whether genetically altered DNA methylation levels impart differential lung cancer risks.
- Differential methylation at 14 CpG sites identified in epigenome-wide association analysis of lung cancer were assessed. Despite >99% power to detect the observational effect sizes, our Mendelian randomisation analysis gave little evidence that any of sites were causally linked to lung cancer.

- This is in stark contrast to previous analyses that suggested two CpG sites within the *AHRR* and *F2RL3* locus, which were also observed in this analysis, mediate >30% of the effect of smoking on lung cancer.
- Overall findings suggest there is little or no role of differential methylation at the CpG sites identified within the blood in the development of lung cancer. Thus, targeting these sites for prevention of lung cancer is unlikely to yield effective treatments.

Background

Lung cancer is the most common cause of cancer-related death worldwide (1). Several DNA methylation changes have been recently identified in relation to lung cancer risk (2-4). Given the plasticity of epigenetic markers, any DNA methylation changes that are causally linked to lung cancer are potentially appealing targets for intervention (5, 6). However, these epigenetic markers are sensitive to reverse causation, being affected by cancer processes (6), and are also prone to confounding, for example by socio-economic and lifestyle factors (7, 8).

One CpG site, cg05575921 within the aryl hydrocarbon receptor repressor (*AHRR*) gene, has been consistently replicated in relation to both smoking (9) and lung cancer (2, 3, 10) and functional evidence suggests that this region could be causally involved in lung cancer (11). However, the observed association between methylation and lung cancer might simply reflect separate effects of smoking on lung cancer and DNA methylation, i.e. the association may be a result of confounding (12), including residual confounding after adjustment for self-reported smoking behaviour (13, 14). Furthermore, recent epigenome-wide association studies (EWAS) for lung cancer have revealed additional CpG sites which may be causally implicated in development of the disease (2, 3).

Mendelian randomization (MR) uses genetic variants associated with modifiable factors as instruments to infer causality between the modifiable factor and outcome, overcoming most unmeasured or residual confounding and reverse causation (15, 16). In

order to infer causality, three core assumptions of MR should be met: 1) The instrument is associated with the exposure, 2) The instrument is not associated with any confounders, 3) The instrument is associated with the outcome only through the exposure. MR may be adapted to the setting of DNA methylation (17-19) with the use of single nucleotide polymorphisms (SNPs) that correlate with methylation of CpG sites, known as methylation quantitative trait loci (mQTLs) (20).

In this study, we performed a meta-analysis of four lung cancer EWAS (918 case-control pairs) from prospective cohort studies to identify CpG sites associated with lung cancer risk and applied MR to investigate whether the observed DNA methylation changes at these sites are causally linked to lung cancer.

142 ## Methods

143 ### EWAS Meta-analysis

144 We conducted a meta-analysis of four lung cancer case-control EWAS that assessed
145 DNA methylation using the Illumina Infinium® HumanMethylation450 BeadChip. All EWAS
146 are nested within prospective cohorts that measured DNA methylation in peripheral blood
147 samples before diagnosis: EPIC-Italy (185 case-control pairs), Melbourne Collaborative
148 Cohort Study (MCCS) (367 case-control pairs), Norwegian Women and Cancer (NOWAC)
149 (132 case-control pairs) and the Northern Sweden Health and Disease Study (NSHDS) (234
150 case-control pairs). Study populations, laboratory methods, data pre-processing and quality
151 control methods have been described in detail elsewhere (3) and outlined in the
152 **Supplementary Methods**.

153

154 To quantify the association between the methylation level at each CpG and the risk
155 of lung cancer we fitted conditional logistic regression models for beta values of
156 methylation (which ranges from 0 (no cytosines methylated) to 1 (all cytosines methylated))
157 on lung cancer status for the four studies. The cases and controls in each study were
158 matched, details of this are in the **Supplementary Methods**. Surrogate variables were
159 computed in the four studies using the SVA R package (21) and the proportion of CD8+ and
160 CD4+ T cells, B cells, monocytes, natural killer cells and granulocytes within whole blood
161 were derived from DNA methylation (22). The following EWAS models were included in the
162 meta-analysis: Model 1 – unadjusted; Model 2 – adjusted for 10 surrogate variables (SVs);
163 Model 3 – adjusted for 10 SVs and derived cell proportions. EWAS stratified by smoking
164 status was also conducted (never (N=304), former (N=648) and current smoking (N=857)).

For Model 1 and Model 2, the case-control studies not matched on smoking status (EPIC-Italy and NOWAC) were adjusted for smoking.

We performed an inverse-variance weighted fixed effects meta-analysis of the EWAS (918 case-control pairs) using the METAL software (http://csg.sph.umich.edu/abecasis/metal/). Direction of effect, effect estimates and the I^2 statistic were used to assess heterogeneity across the studies in addition to effect estimates across smoking strata (never, former and current). All sites identified at a false discovery rate (FDR)<0.05 in Model 2 and 3 were also present in the sites identified in Model 1. The effect size differences between models for all sites identified in Model 1 were assessed by a Kruskal-Wallis test and a post-hoc Dunn's test. There was little evidence for a difference (P > 0.1), so to maximize inclusion into the MR analyses we took forward the sites identified in the unadjusted model (Model 1).

Mendelian randomization

Two-sample MR was used to establish potential causal effect of differential methylation on lung cancer risk (23, 24). In the first sample, we identified mQTL-methylation effect estimates (β_{GP}) for each CpG site of interest in an mQTL database from the Accessible Resource for Integrated Epigenomic Studies (ARIES) (http://www.mqtldb.org). Details on the methylation pre-processing, genotyping and quality control (QC) pipelines are outlined in the **Supplementary Methods**. In the second sample, we used summary data from a GWAS meta-analysis of lung cancer risk conducted by the Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium (TRICL-ILCCO) (29,863 cases, 55,586 controls) to obtain mQTL-lung cancer estimates (β_{GD}) (25).

188 For each independent mQTL (r^2<0.01, we calculated the log odds ratio (OR) per SD
189 unit increase in methylation by the formula β_{GD}/β_{GP} (Wald ratio). Standard errors were
190 approximated by the delta method (26). Where multiple independent mQTLs were available
191 for one CpG site, these were combined in a fixed effects meta-analysis after weighting each
192 ratio estimate by the inverse variance of their associations with the outcome. Heterogeneity
193 in Wald ratios across mQTLs was estimated using Cochran's Q test, which can be used to
194 indicate horizontal pleiotropy (27). Differences between the observational and MR
195 estimates were assessed using a Z-test for difference.

196

197 If there was evidence for an mQTL-CpG site association in ARIES in at least one time-
198 point, we assessed whether the mQTL replicated across time points in ARIES (FDR < 0.05,
199 same direction of effect). Further, we re-analysed this association using linear regression of
200 methylation on each genotyped SNP available in an independent cohort (NSHDS), using
201 rvtests (28) (**Supplementary Methods**). Replicated mQTLs were included where possible to
202 reduce the effect of winner's curse using effect estimates from ARIES. We assessed the
203 instrument strength of the mQTLs by investigating the variance explained in methylation by
204 each mQTL (r^2) as well as the F-statistic in ARIES (**Supplementary Table 1**). The power to
205 detect the observational effect estimates in the two-sample MR analysis was assessed *a*
206 *priori*, based on an alpha of 0.05, sample size of 29,863 cases and 55,586 controls (from
207 TRICL-ILCCO) and calculated variance explained (r^2).

208

209 MR analyses were also performed to investigate the impact of methylation on lung
210 cancer subtypes in TRICL-ILCCO: adenocarcinoma (11,245 cases, 54,619 controls), small cell

211 carcinoma (2791 cases, 20,580 controls), and squamous cell carcinoma (7704 cases, 54,763
212 controls). We also assessed the association in never smokers (2303 cases, 6995 controls)
213 and ever smokers (23,848 cases, 16,605 controls) (25). Differences between the smoking
214 subgroups were assessed using a Z-test for difference.

215

216 We next investigated the extent to which the mQTLs at cancer-related CpGs were
217 associated with four smoking behaviour traits which could confound the methylation-lung
218 cancer association: number of cigarettes per day, smoking cessation rate, smoking initiation
219 and age of smoking initiation using GWAS data from the Tobacco and Genetics (TAG)
220 consortium (N=74,053) (29).

221

222 Supplementary analyses

223 *Assessing the potential causal effect of AHRR methylation: one sample MR*

224 Given previous findings implicating methylation at *AHRR* in relation to lung cancer (2,
225 3), we performed a one-sample MR analysis (30) of *AHRR* methylation on lung cancer
226 incidence using individual-level data from the Copenhagen City Heart Study (CCHS) (357
227 incident cases, 8401 remaining free of lung cancer). Details of the phenotypic, methylation
228 and genetic data, as well as the linked lung cancer data, are outlined in the **Supplementary**
229 **Methods**.

230

231 An allele score of mQTLs located with 1Mb of cg05575921-*AHRR* was created and its
232 association with *AHRR* methylation tested (**Supplementary Methods**). We investigated
233 associations between the allele score and several potential confounding factors (sex, alcohol
234 consumption, smoking status, occupational exposure to dust and/or welding fumes, passive
235 smoking). We next performed MR analyses using two-stage Cox regression, with adjustment
236 for age and sex, and further stratified by smoking status.

237

238 *Tumour and adjacent normal methylation patterns*

239 DNA methylation data from lung cancer tissue and matched normal adjacent tissue
240 (N=40 squamous cell carcinoma and N=29 adenocarcinoma), profiled as part of The Cancer
241 Genome Atlas (TCGA), were used to assess tissue-specific DNA methylation changes across
242 sites identified in the meta-analysis of EWAS, as outlined previously (31),

243

244 *mQTL association with gene expression*

245 For the genes annotated to CpG sites identified in the lung cancer EWAS, we
246 examined gene expression in whole blood and lung tissue using data from the gene-tissue
247 expression (GTEx) consortium (32).

248

249 Analyses were conducted in Stata (version 14) and R (version 3.2.2). For the two-
250 sample MR analysis we used the MR-Base R package TwoSampleMR (33). An adjusted P

251 value that limited the FDR was calculated using the Benjamini-Hochberg method (34). All
252 statistical tests were two-sided.
253

254 Results

255 A flowchart representing our study design along with a summary of our results at each step
256 is displayed in **Figure 1**.
257 (Figure 1 here)
258

259 EWAS meta-analysis

260 The basic meta-analysis adjusted for study-specific covariates identified 16 CpG sites
261 which were hypomethylated in relation to lung cancer (FDR<0.05, Model 1, **Figure 2**).
262 Adjusting for 10 surrogate variables (Model 2) and derived cell counts (Model 3) gave similar
263 results (**Table 1**). The direction of effect at the 16 sites did not vary between studies
264 (median I^2=38.6) (**Supplementary Table 2**), but there was evidence for heterogeneity of
265 effect estimates at some sites when stratifying individuals by smoking status (**Table 1**).
266 (Figure 2 here)
267 (Table 1 here)
268

269 Mendelian randomization

270 We identified 15 independent mQTLs (r^2<0.01) associated with methylation at 14 of
271 16 CpGs. Nine mQTLs replicated at FDR<0.05 in NSHDS (**Supplementary Table 3**). MR power

272 analyses indicated >99% power to detect ORs for lung cancer of the same magnitude as
273 those in the meta-analysis of EWAS.

274

275 There was little evidence for an effect of methylation at these 14 sites on lung cancer
276 (FDR>0.05, **Supplementary Table 4**). For nine of 14 CpG sites the point estimates from the
277 MR analysis were in the same direction as in the EWAS, but of a much smaller magnitude (Z-
278 test for difference, P<0.001) (**Figure 3**).

279 For 9 of out the 16 mQTL-CpG associations, there was strong replication across time
280 points (**Supplementary Table 5**) and 9 out of 16 mQTL-CpG associations replicated at
281 FDR<0.05 in an independent adult cohort (NSHDS). Using mQTL effect estimates from
282 NSHDS for the nine CpG sites that replicated (FDR<0.05), findings were consistent with
283 limited evidence for a causal effect of peripheral blood-derived DNA methylation on lung
284 cancer (**Supplementary Figure 1**).

285 (Figure 3 here)

286 There was little evidence of different effect estimates between ever and never
287 smokers at individual CpG sites (**Supplementary Figure 2,** Z-test for difference, P>0.5). There
288 was some evidence for a possible effect of methylation at cg21566642-*ALPPL2* and
289 cg23771366-*PRSS23* on squamous cell lung cancer (OR=0.85 [95% confidence interval
290 (CI)=0.75,0.97] and 0.91 [95% CI=0.84,1.00] per SD [14.4% and 5.8%] increase, respectively)
291 as well as methylation at cg23387569-*AGAP2,* cg16823042-*AGAP2*, and cg01901332-*ARRB1*
292 on lung adenocarcinoma (OR=0.86 [95% CI=0.77,0.96], 0.84 [95% CI=0.74,0.95], and 0.89
293 [95% CI=0.80,1.00] per SD [9.47%, 8.35%, and 8.91%] increase, respectively). However, none

294 of the results withstood multiple testing correction (FDR<0.05) (**Supplementary Figure 3**).
295 For those CpGs where multiple mQTLs were used as instruments (cg05575921-*AHRR* and
296 cg01901332-*ARRB1*), there was limited evidence for heterogeneity in MR effect estimates
297 (Q-test, P>0.05, **Supplementary Table 6**).

298 Single mQTLs for cg05575921-*AHRR,* cg27241845-*ALPPL2,* and cg26963277-*KCNQ1*
299 showed some evidence of association with smoking cessation (former vs. current smokers),
300 although these associations were not below the FDR<0.05 threshold (**Supplementary Figure**
301 **4**).

302 *Potential causal effect of* AHRR *methylation on lung cancer risk: one sample MR*

303 In the CCHS, a per (average methylation-increasing) allele change in a four-mQTL
304 allele score was associated with a 0.73% [95% CI=0.56,0.90] increase in methylation (P<1 x
305 10^{-10}) and explained 0.8% of the variance in cg05575921-*AHRR* methylation (F-
306 statistic=74.2). Confounding factors were not strongly associated with the genotypes in this
307 cohort (P≥0.11) (**Supplementary Table 7**). Results provided some evidence for an effect of
308 cg05575921 methylation on total lung cancer risk (HR=0.30 [95% CI=0.10,1.00] per SD
309 (9.2%) increase) (**Supplementary Table 8**). The effect estimate did not change substantively
310 when stratified by smoking status (**Supplementary Table 9**).

311

312 Given contrasting findings with the main MR analysis, where cg05575921-*AHRR*
313 methylation was not causally implicated in lung cancer, and the lower power in the one-
314 sample analysis to detect an effect of equivalent size to the observational results (power =
315 19% at alpha = 0.05), we performed further two-sample MR based on the four mQTLs using

data from the TRICL-ILCCO consortium. Results showed no strong evidence for a causal effect of DNA methylation on total lung cancer risk (OR=1.00 [95% CI=0.83,1.10] per SD increase) (**Supplementary Figure 5)**. There was also limited evidence for an effect of cg05575921-*AHRR* methylation when stratified by cancer subtype and smoking status (**Supplementary Figure 5**) and no strong evidence for heterogeneity of the mQTL effects (**Supplementary Table 9**). Conclusions were consistent when MR-Egger (27) was applied (**Supplementary Figure 5**) and when accounting for correlation structure between the mQTLs (**Supplementary Table 9**).

Tumour and adjacent normal lung tissue methylation patterns

For cg05575921-*AHRR,* there was no strong evidence for differential methylation between adenocarcinoma tissue and adjacent healthy tissue (P=0.963), and weak evidence for hypermethylation in squamous cell carcinoma tissue (P=0.035) (**Figure 4**, **Supplementary Table 10**). For the other CpG sites there was evidence for a difference in DNA methylation between tumour and healthy adjacent tissue at several sites in both adenocarcinoma and squamous cell carcinoma, with consistent differences for CpG sites in *ALPPL2* (cg2156642, cg05951221 and cg01940273), as well as cg23771366-*PRSS23*, cg26963277-*KCNQ1*, cg09935388-*GFI1*, cg0101332-*ARRB1*, cg08709672-*AVPR1B* and cg25305703-*CASC21*. However, hypermethylation in tumour tissue was found for the majority of these sites, which is opposite to what was observed in the EWAS analysis.

(Figure 4 here)

Gene expression associated with mQTLs in blood and lung tissue

Of the 10 genes annotated to the 14 CpG sites, eight genes were expressed sufficiently to be detected in lung (*AVPR1B* and *CASC21* were not) and seven in blood (*AVPR1B*, *CASC21* and *ALPPL2* were not). Of these, gene expression of *ARRB1* could not be investigated as the mQTLs in that region were not present in the GTEx data. rs3748971 and rs878481, mQTLs for cg21566642 and cg05951221 respectively, were associated with increased expression of *ALPPL2* (P=0.002 and P=0.0001). No other mQTLs were associated with expression of the annotated gene at a Bonferroni corrected P value threshold (P<0.05/19=0.0026) (**Supplementary Table 11**).

Discussion

In this study, we identified 16 CpG sites associated with lung cancer, of which 14 have been previously identified in relation to smoke exposure (9) and six were highlighted in a previous study as being associated with lung cancer (3). This previous study used the same data from the four cohorts investigated here, but in a discovery and replication, rather than meta-analysis framework. Overall, using MR we found limited evidence supporting a potential causal effect of methylation at the CpG sites identified in peripheral blood on lung cancer. These findings are in contrast to previous analyses suggesting that methylation at two CpG sites investigated (in *AHRR* and *F2RL3*) mediated > 30% of the effect of smoking on lung cancer risk (2). This previous study used methods which are sensitive to residual confounding and measurement error that may have biased results (12, 35). These limitations are largely overcome using MR (12). While there was some evidence for an effect of methylation at some of the other CpG sites on risk of subtypes of lung cancer,

these effects were not robust to multiple testing correction and were not validated in the analysis of tumour and adjacent normal lung tissue methylation nor in gene expression analysis.

A major strength of the study was the use of two-sample MR to integrate an extensive epigenetic resource and summary data from a large lung cancer GWAS to appraise causality of observational associations with >99% power. Evidence against the observational findings were also acquired through tissue-specific DNA methylation and gene expression analyses.

Limitations include potential "winner's curse" which may bias causal estimates in a two-sample MR analysis towards the null if the discovery sample for identifying genetic instruments is used as the first sample, as was done for our main MR analysis using data from ARIES (36). However, findings were similar when using replicated mQTLs in NSHDS, indicating the potential impact of this bias was minimal (**Supplementary Figure 1**). Another limitation relates to the potential issue of consistency and validity of the instruments across the two samples. For a minority of the mQTL-CpG associations (4 out of 16), there was limited replication across time points and in particular, 6 mQTLs were not strongly associated with DNA methylation in adults. Caution is therefore warranted when interpreting the null results for the two-sample MR estimates for these CpG sites, which could be the result of weak-instrument bias.

The lack of independent mQTLs for each CpG site did not allow us to properly appraise horizontal pleiotropy in our MR analyses. Where possible we only included cis-acting mQTLs to minimise pleiotropy and investigated heterogeneity where there were multiple independent mQTLs. Three mQTLs were nominally associated with smoking phenotypes, but not to the extent that this would bias our MR results substantially. Some of the mQTLs used influence multiple CpGs in the same region, suggesting genomic control of methylation at a regional rather than single CpG level. This was untested, but methods to detect differentially methylated regions (DMRs) and identify genetic variants which proxy for them may be fruitful in probing the effect of methylation across gene regions.

A further limitation relates to the inconsistency in effect estimates between the one- and two-sample MR analysis to appraise the causal role of *AHRR* methylation. While findings in CCHS were supportive of a causal effect of *AHRR* methylation on lung cancer (HR=0.30 [95% CI=0.10,1.00] per SD), in two-sample MR this site was not causally implicated (OR=1.01 [95% CI=0.993, 1.03]). We verified that this was not due to differences in the genetic instruments used, nor due to issues of weak instrument bias. Given the CCHS one-sample MR had little power (19% at alpha = 0.05) to detect a causal effect, we have more confidence in the results from the two-sample approach.

Peripheral blood may not be the ideal tissue to assess the association between DNA methylation and lung cancer. While a high degree of concordance in mQTLs has been observed across lung tissue, skin and peripheral blood DNA (37), we were unable to directly evaluate this here. A possible explanation for a lack of causal effect at *AHRR* is due to the limitation of tissue specificity as we found that the mQTLs used to instrument cg05575921

were not strongly related to expression of *AHRR* in lung tissue. However, findings from MR analysis were corroborated by the lack of evidence for differential methylation at *AHRR* between lung adenocarcinoma tissue and adjacent healthy tissue, and weak evidence for hypermethylation (opposite to the expected direction) in squamous cell lung cancer tissue. This result may be interesting in itself as smoking is hypothesized to influence squamous cell carcinoma more than adenocarcinoma. However, the result conflicts with that found in the MR analysis. Furthermore, another study investigating tumorous lung tissue (N=511) found only weak evidence for an association between smoking and cg05575921 *AHRR* methylation, that did not survive multiple testing correction (P=0.02) (38). However, our results do not fully exclude AHRR from involvement in the disease process. AHRR and AHR form a regulatory feedback loop, which means that the actual effect of differential methylation or differential expression of *AHR/AHRR* on pathway activity is complex (39). In addition, some of the CpG sites identified in the EWAS were found to be differentially methylated in the tumour and adjacent normal lung tissue comparison. While this could represent a false negative result of the MR analysis, it is of interest that differential methylation in the tissue comparison analysis was typically in the opposite direction to that observed in the EWAS. Furthermore, while this method can be used to minimize confounding, it does not fully eliminate the possibility of bias due to reverse causation (whereby cancer induces changes in DNA methylation) or intra-individual confounding e.g. by gene expression. Therefore, it doesn't give conclusive evidence that DNA methylation changes at these sites are not relevant to the development of lung cancer.

While DNA methylation in peripheral blood may be predictive of lung cancer risk, according to the present analysis it is unlikely to play a causal role in lung carcinogenesis at the CpG sites investigated. Findings from this study issue caution over the use of traditional mediation analyses to implicate intermediate biomarkers (such as DNA methylation) in pathways linking an exposure with disease, given the potential for residual confounding in this context (12). However, the findings of this study do not preclude the possibility that other DNA methylation changes are causally related to lung cancer (or other smoking-associated disease) (40).

References

1. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, C. M. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet] 2013 [Available from: http://globocan.iarc.fr.
2. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun. 2015;6:10192.
3. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. Int J Cancer. 2017;140(1):50-61.
4. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279-83.
5. Strathdee G, Brown R. Aberrant DNA methylation in cancer: potential clinical interventions. Expert Rev Mol Med. 2002;4(4):1-17.
6. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002;3(6):415-28.
7. Borghol N, Suderman M, McArdle W, Racine A, Hallett M, Pembrey M, et al. Associations with early-life socio-economic position in adult DNA methylation. Int J Epidemiol. 2012;41(1):62-74.
8. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics. 2014;6(1):4.
9. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet. 2016;9(5):436-47.
10. Bojesen SE, Timpson N, Relton C, Davey Smith G, Nordestgaard BG. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. Thorax. 2017.

11. Zudaire E, Cuesta N, Murty V, Woodson K, Adarns L, Gonzalez N, et al. The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers. J Clin Invest. 2008;118(2):640-50.
12. Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. Challenges and novel approaches for investigating molecular mediation. Hum Mol Genet. 2016;25(R2):R149-R56.
13. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol. 2007;166(6):646-55.
14. Munafo MR, Timofeeva MN, Morris RW, Prieto-Merino D, Sattar N, Brennan P, et al. Association Between Genetic Variants on Chromosome 15q25 Locus and Objective Measures of Tobacco Exposure. Jnci-J Natl Cancer I. 2012;104(10):740-8.
15. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23(R1):R89-98.
16. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology. 2003;32(1):1-22.
17. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. Int J Epidemiol. 2012;41(1):161-76.
18. Relton CL, Davey Smith G. Mendelian randomization: applications and limitations in epigenetic studies. Epigenomics. 2015;7(8):1239-43.
19. Richardson TG, Zheng J, Davey Smith G, Timpson NJ, Gaunt TR, Relton CL, et al. Mendelian Randomization Analysis Identifies CpG Sites as Putative Mediators for Genetic Influences on Cardiovascular Disease Risk. Am J Hum Genet. 2017;101(4):590-602.
20. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17:61.
21. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD. sva: Surrogate Variable Analysis. R package version 30. 2017.
22. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. Bmc Bioinformatics. 2012;13.
23. Inoue A, Solon G. Two-Sample Instrumental Variables Estimators. Rev Econ Stat. 2010;92(3):557-61.
24. Pierce BL, Burgess S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am J Epidemiol. 2013;178(7):1177-84.
25. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017.
26. Thomas DC, Lawlor DA, Thompson JR. Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by bautista et al. Ann Epidemiol. 2007;17(7):511-3.
27. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology. 2015;44(2):512-25.

28. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics. 2016;32(9):1423-6.
29. Tobacco, Genetics C. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010;42(5):441-7.
30. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr. 2016;103(4):965-78.
31. Teschendorff AE, Yang Z, Wong A, Pipinikas CP, Jiao Y, Jones A, et al. Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. JAMA Oncol. 2015;1(4):476-85.
32. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.
33. Hemani G, Zheng J, Wade KH, Elsworth B, Langdon R, Burgess S, et al. MR-Base: a platform for Mendelian randomization using summary data from genome-wide association studies. eLife. 2018;7:e34408.
34. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met. 1995;57(1):289-300.
35. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet. 2017;13(11):e1007081.
36. Burgess S, Thompson SG, Collaboration CCG. Avoiding bias from weak instruments in Mendelian randomization studies. International Journal of Epidemiology. 2011;40(3):755-64.
37. Shi J, Marconett CN, Duan J, Hyland PL, Li P, Wang Z, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun. 2014;5:3365.
38. Freeman JR, Chu S, Hsu T, Huang YT. Epigenome-wide association study of smoking and DNA methylation in non-small cell lung neoplasms. Oncotarget. 2016;7(43):69579-91.
39. Chen YT, Widschwendter M, Teschendorff AE. Systems-epigenomics inference of transcription factor activity implicates aryl-hydrocarbon-receptor inactivation as a key event in lung cancer development. Genome Biology. 2017;18.
40. Gao X, Zhang Y, Breitling LP, Brenner H. Tobacco smoking and methylation of genes related to lung cancer development. Oncotarget. 2016;7(37):59017-28.

Funding

This work was partly supported by a Wellcome Trust PhD studentship to TB [203746]; and by Cancer Research UK [C18281/A19169, C57854/A22171 and C52724/A20138]. This work was also supported by the UK Medical Research Council [MC_UU_12013/1 and MC_UU_12013/2], which funds a Unit at the University of Bristol where TB, RCR, PCH, TRG, GDS and CLR work. Funding to pay the Open Access publication charges for this article was provided by the University of Bristol RCUK.

The UK Medical Research Council and Wellcome [Grant ref: 102215/2/13/2] and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and TB, RCR and CLR will serve as guarantors for the contents of this paper.

Methylation data in the ALSPAC cohort were generated as part of the UK BBSRC funded [BB/I025751/1 and BB/I025263/1] Accessible Resource for Integrated Epigenomic Studies (ARIES, http://www.ariesepigenomics.org.uk).

Competing interests statement

The authors declare no competing interests.

Acknowledgements

For the contributions of ALSPAC data to our study: we are extremely grateful to all the families who took part, the midwives for their help in recruiting them, and the whole

ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

Figure legends

Figure 1. Study design with results summary. ARIES = Accessible Resource for Integrated Epigenomic Studies, TRICL-ILLCO = Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium, MR = Mendelian randomization, CCHS = Copenhagen City Heart Study, TCGA = The Cancer Genome Atlas. * = 2 000 individuals with samples at multiple timepoints.

Figure 2. Observational associations of DNA methylation and lung cancer: A fixed effects meta-analysis of lung cancer EWAS weighted on the inverse variance was performed to establish the observational association between differential DNA methylation and lung cancer. **a)** Manhattan plot**,** all points above the solid line are at $P < 1x10^{-7}$ and all points above the dashed line (and triangular points) are at FDR < 0.05. In total 16 CpG sites are associated with lung cancer (FDR < 0.05). **b)** Quantile-quantile plot of the EWAS results (same data as the **Figure 2a** Manhattan plot).

Figure 3. Mendelian randomization (MR) vs. observational analysis. Two-sample MR was carried out with methylation at 14/16 CpG sites identified in the EWAS meta-analysis as the exposure and lung cancer as the outcome. cg01901332 and cg05575921 had 2 instruments so the estimate was calculated using the inverse variance weighted method, for the rest the MR estimate was calculated using a Wald ratio. Only 14 of 16 sites could be instrumented using mQTLs from mqtldb.org. * = instrumental variable not replicated in independent dataset (NSHDS). The

sites for which instrumental variables have not been replicated are cg01901332, cg21566642, cg05575921 and cg08709672. OR = odds ratio per SD increase in DNA methylation.

Figure 4. Differential DNA methylation in lung cancer tissue: A comparison of methylation at each of the 16 CpG sites identified in our meta-analysis was made between lung cancer tissue and adjacent healthy lung tissue for patients with a) lung adenocarcinoma and b) squamous cell lung cancer. Data from The Cancer Genome Atlas was used for this analysis.

Tables

Table 1. Meta-analyses of EWAS of lung cancer using four separate cohorts: 16 CpG sites associated with lung cancer at false discovery rate < 0.05.

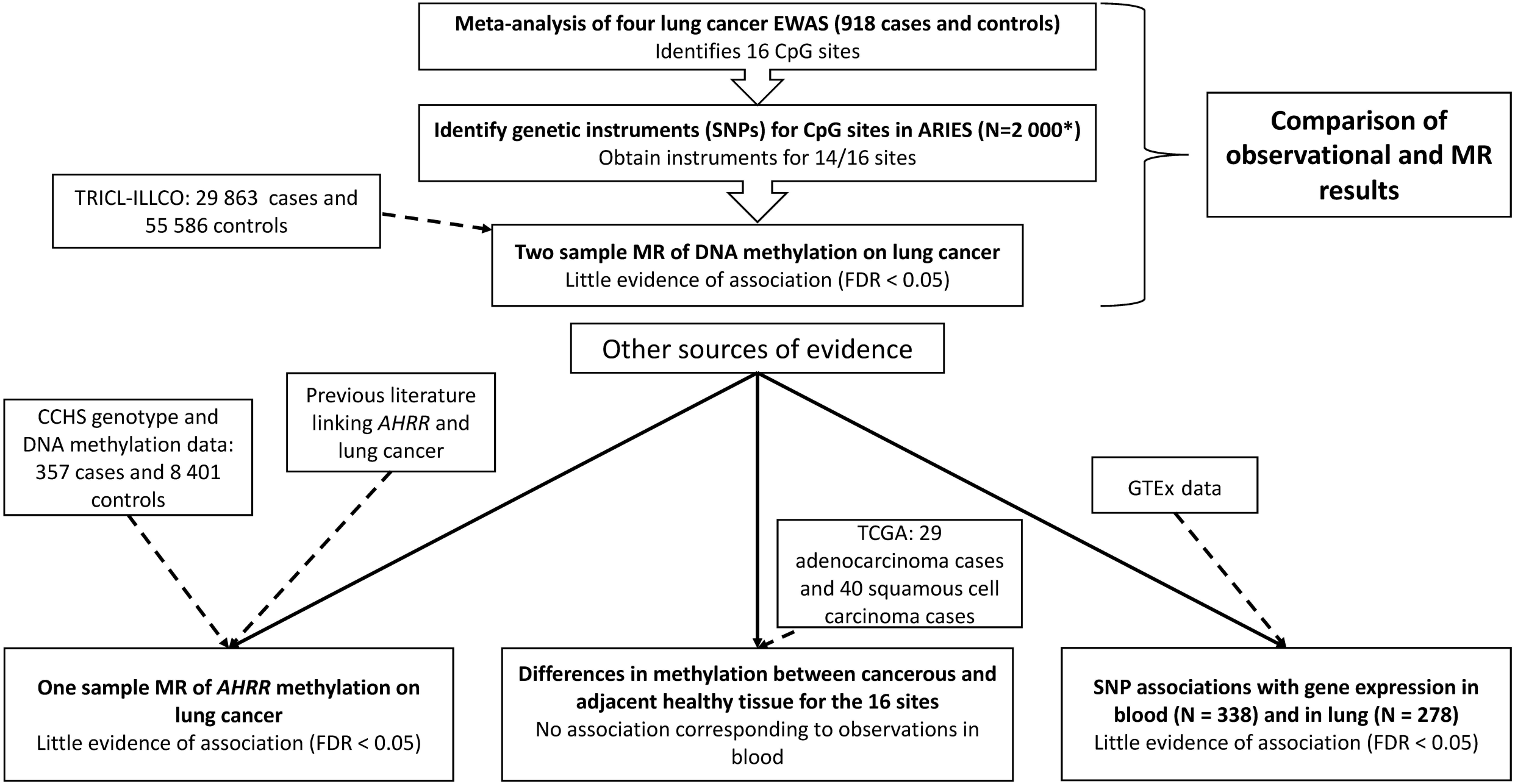
				Basic			SV adjusted			Cell count + SV adjusted			Never smokers			Former smokers			Current smokers			Smoker group comparison		
CpG	**Gene**	**Chr**	**Position**	**OR**	**SE**	**P**	**OR**	**SE**	**P**	**OR**	**SE**	**P**	**OR**	**SE**	**P**	**OR**	**SE**	**P**	**OR**	**SE**	**P**	**Dir**	**I2**	**P**
cg05575921	*AHRR*	5	373378	0.474	0.047	1.45E-16	0.452	0.053	6.27E-14	0.452	0.055	3.60E-13	0.932	0.22	7.17E-01	0.458	0.084	6.10E-07	0.708	0.066	5.36E-05	+--	63	0.07
cg21566642	*ALPPL2*	2	233284661	0.535	0.045	1.70E-15	0.525	0.05	2.49E-13	0.513	0.051	3.12E-13	0.892	0.145	4.18E-01	0.522	0.081	1.42E-06	0.746	0.067	3.67E-04	+--	81	0.01
cg06126421	*IER3*	6	30720080	0.585	0.046	2.08E-13	0.544	0.054	2.49E-11	0.513	0.054	3.92E-12	0.783	0.192	2.22E-01	0.561	0.087	1.88E-05	0.727	0.112	1.79E-02	---	33	0.23
cg03636183	*F2RL3*	19	17000585	0.636	0.045	7.99E-12	0.615	0.053	8.21E-10	0.61	0.054	1.61E-09	0.909	0.172	5.53E-01	0.624	0.084	7.50E-05	0.786	0.069	2.92E-03	---	71	0.03
cg05951221	*ALPPL2*	2	233284402	0.66	0.045	9.68E-11	0.642	0.051	1.77E-09	0.629	0.052	1.50E-09	0.868	0.176	4.09E-01	0.634	0.082	7.21E-05	0.819	0.066	7.42E-03	---	44	0.17
cg01940273	*ALPPL2*	2	233284934	0.692	0.05	4.20E-08	0.675	0.058	7.32E-07	0.685	0.061	3.58E-06	1.144	0.23	4.28E-01	0.575	0.086	2.57E-05	0.876	0.068	6.59E-02	---	22	0.28
cg23771366	*PRSS23*	11	86510998	0.769	0.04	1.10E-07	0.729	0.051	1.45E-06	0.709	0.052	5.60E-07	1.093	0.16	4.90E-01	0.621	0.076	1.40E-05	0.856	0.061	1.97E-02	---	0	0.66
cg11660018	*PRSS23*	11	86510915	0.788	0.037	1.18E-07	0.7	0.051	1.97E-07	0.678	0.053	8.86E-08	0.935	0.131	5.86E-01	0.753	0.071	1.01E-03	0.844	0.053	4.15E-03	---	0	0.53
cg26963277	*KCNQ1*	11	2722407	0.668	0.055	1.21E-07	0.64	0.068	3.79E-06	0.623	0.069	2.53E-06	0.539	0.175	1.40E-02	0.724	0.11	1.54E-02	0.707	0.087	1.59E-03	---	16	0.31
cg27241845	*ALPPL2*	2	233250370	0.669	0.055	1.45E-07	0.679	0.067	1.67E-05	0.673	0.069	2.47E-05	0.75	0.208	1.93E-01	0.677	0.108	5.01E-03	0.726	0.087	3.09E-03	---	0	0.65
cg23387569	*AGAP2*	12	58120011	0.713	0.049	1.53E-07	0.702	0.058	3.69E-06	0.683	0.059	1.89E-06	0.786	0.164	1.69E-01	0.714	0.107	1.02E-02	0.749	0.079	2.48E-03	---	69	0.04
cg09935388	*GFI1*	1	92947588	0.676	0.055	2.48E-07	0.669	0.066	9.67E-06	0.674	0.07	3.00E-05	0.961	0.242	8.44E-01	0.74	0.127	4.22E-02	0.681	0.075	1.06E-04	---	0	0.89
cg01901332	*ARRB1*	11	75031054	0.725	0.048	2.82E-07	0.686	0.064	1.12E-05	0.658	0.064	2.20E-06	1.017	0.214	9.22E-01	0.599	0.093	1.48E-04	0.783	0.072	3.92E-03	+--	81	0.01
cg25305703	*CASC21*	8	128378218	0.725	0.049	4.46E-07	0.717	0.067	1.11E-04	0.715	0.069	1.48E-04	0.801	0.169	2.10E-01	0.761	0.106	2.58E-02	0.769	0.075	3.20E-03	---	0	0.98
cg16823042	*AGAP2*	12	58119992	0.739	0.049	1.14E-06	0.726	0.058	1.51E-05	0.701	0.059	5.90E-06	0.83	0.183	3.09E-01	0.72	0.1	7.36E-03	0.799	0.08	1.35E-02	---	10	0.33
cg08709672	*AVPR1B*	1	206224334	0.749	0.048	1.36E-06	0.759	0.058	1.14E-04	0.739	0.06	5.33E-05	0.729	0.171	1.02E-01	0.738	0.085	3.47E-03	0.816	0.079	2.13E-02	---	0	0.85

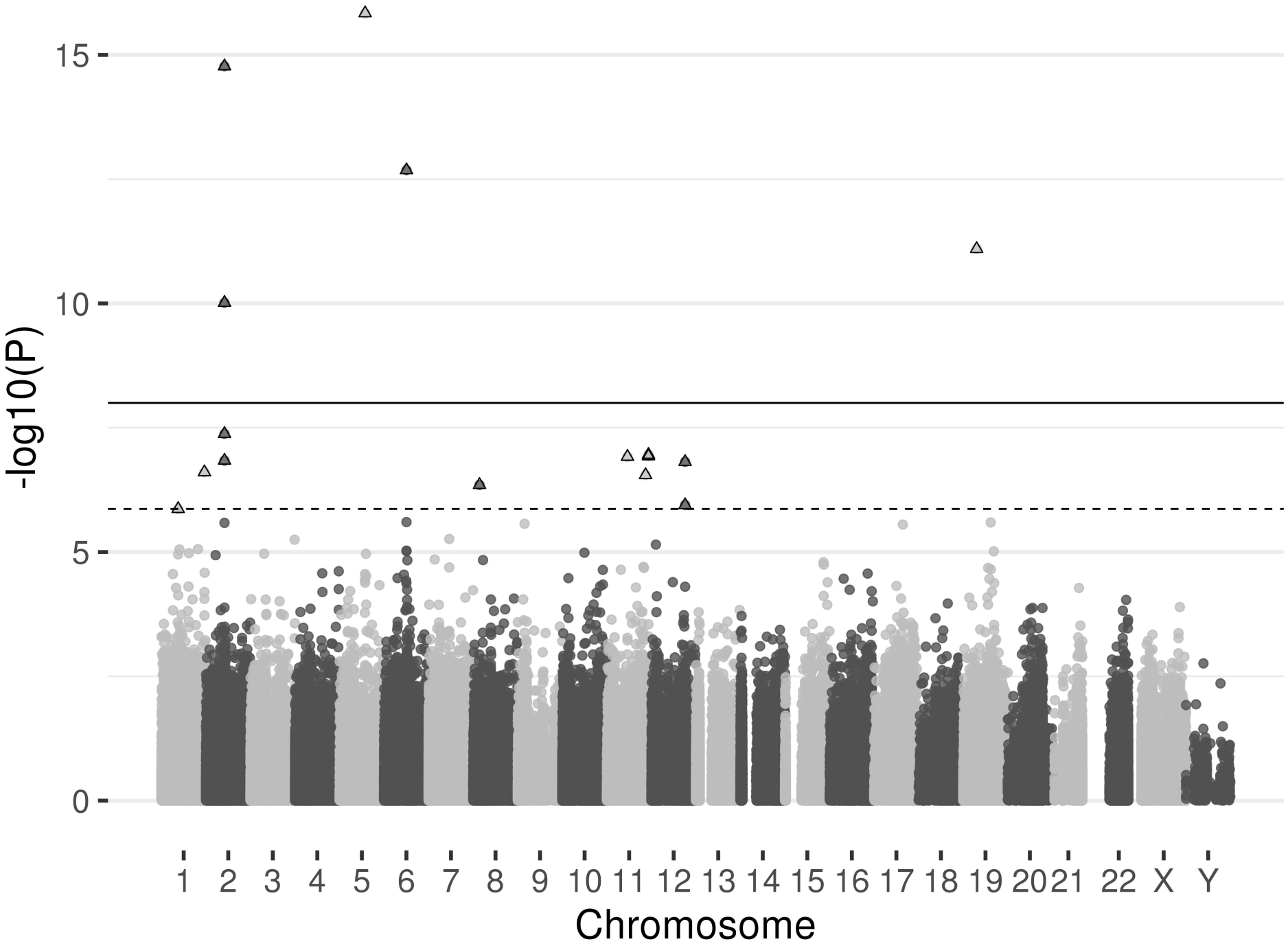
Meta-analyses of epigenome-wide association studies of lung cancer adjusted for study specific covariates (Basic, N = 1809), basic model + surrogate variables (SV adjusted, N = 1809), basic model + surrogate variables + derive

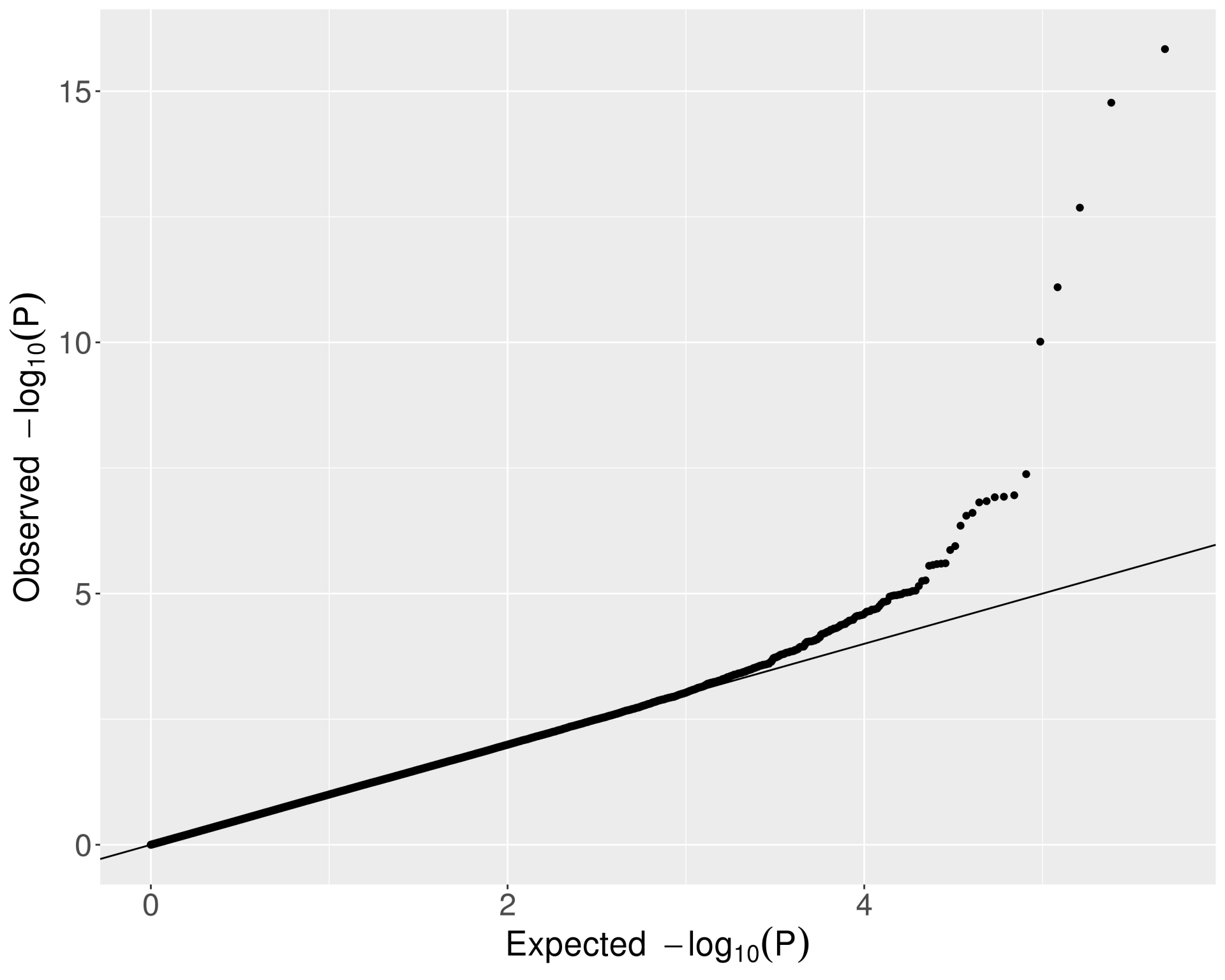
Meta-analyses were also conducted stratified by smoking status (never (N = 304), former (N = 648), current (N = 857)) using the basic model

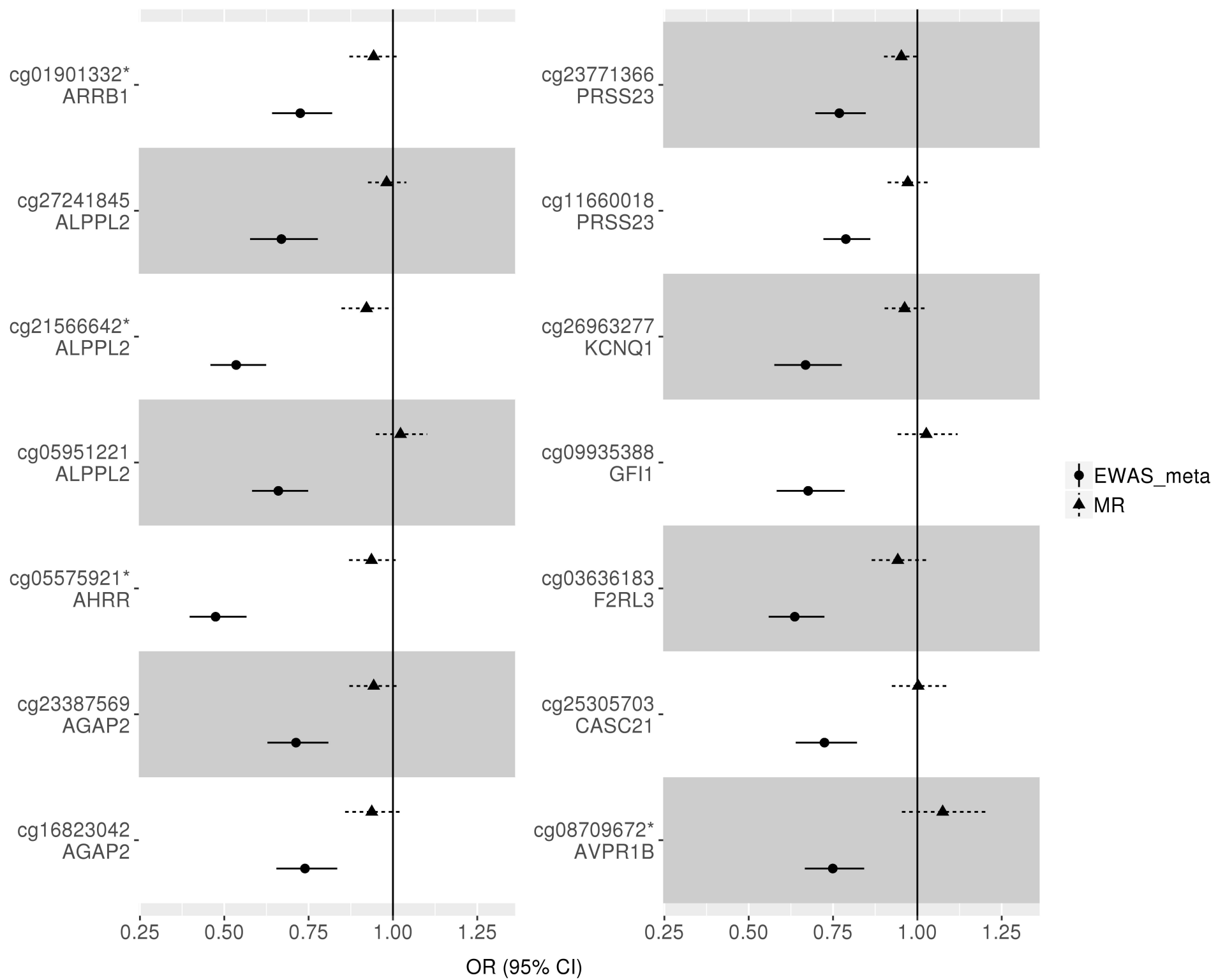
OR = odds ratio per SD increase in DNA methylation, SE = standard error, Chr = chromosome

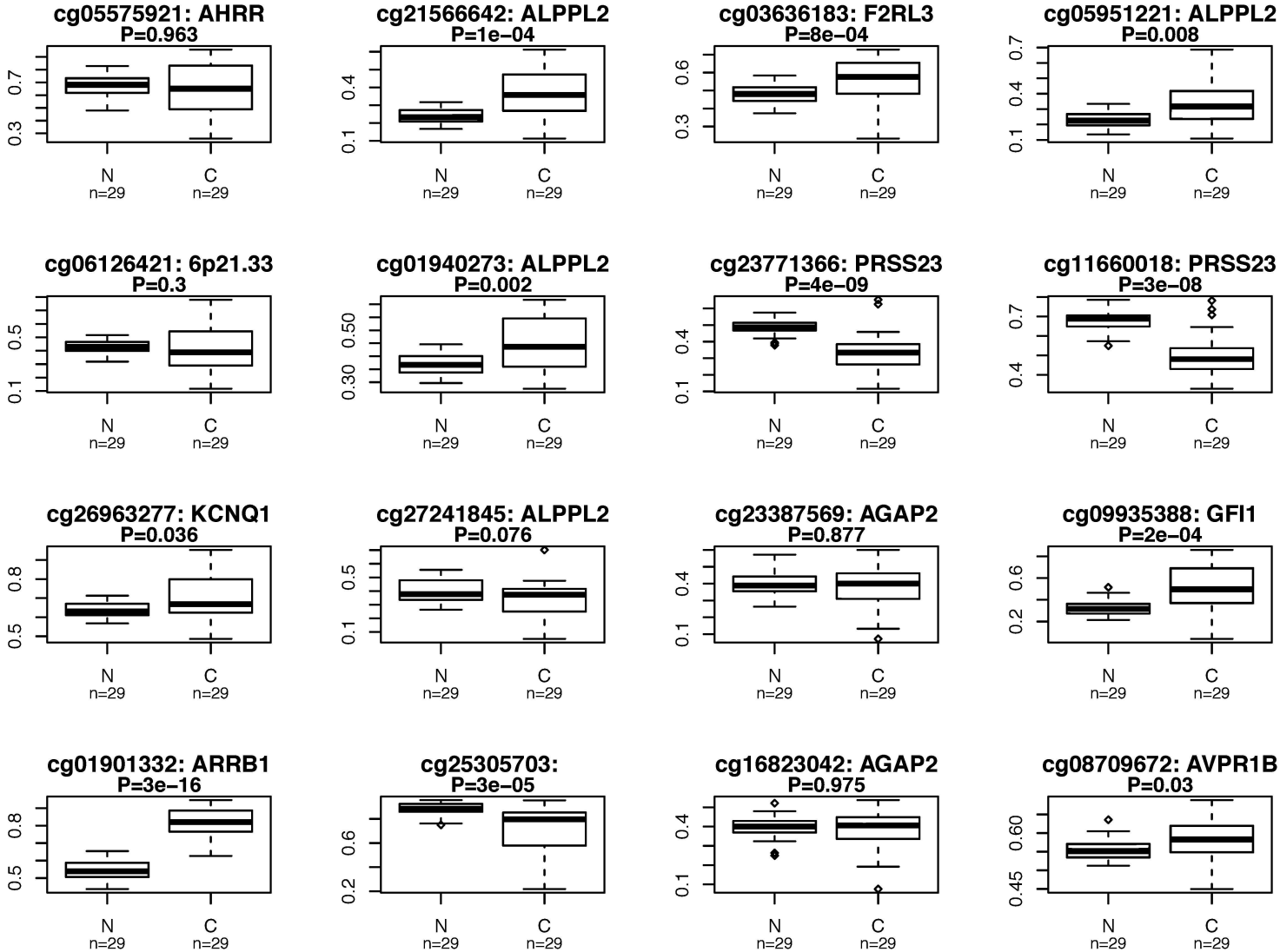
Smoker group comparisson = heterogeneity across meta-analyses when stratifying by smoking status, Dir = direction of effect

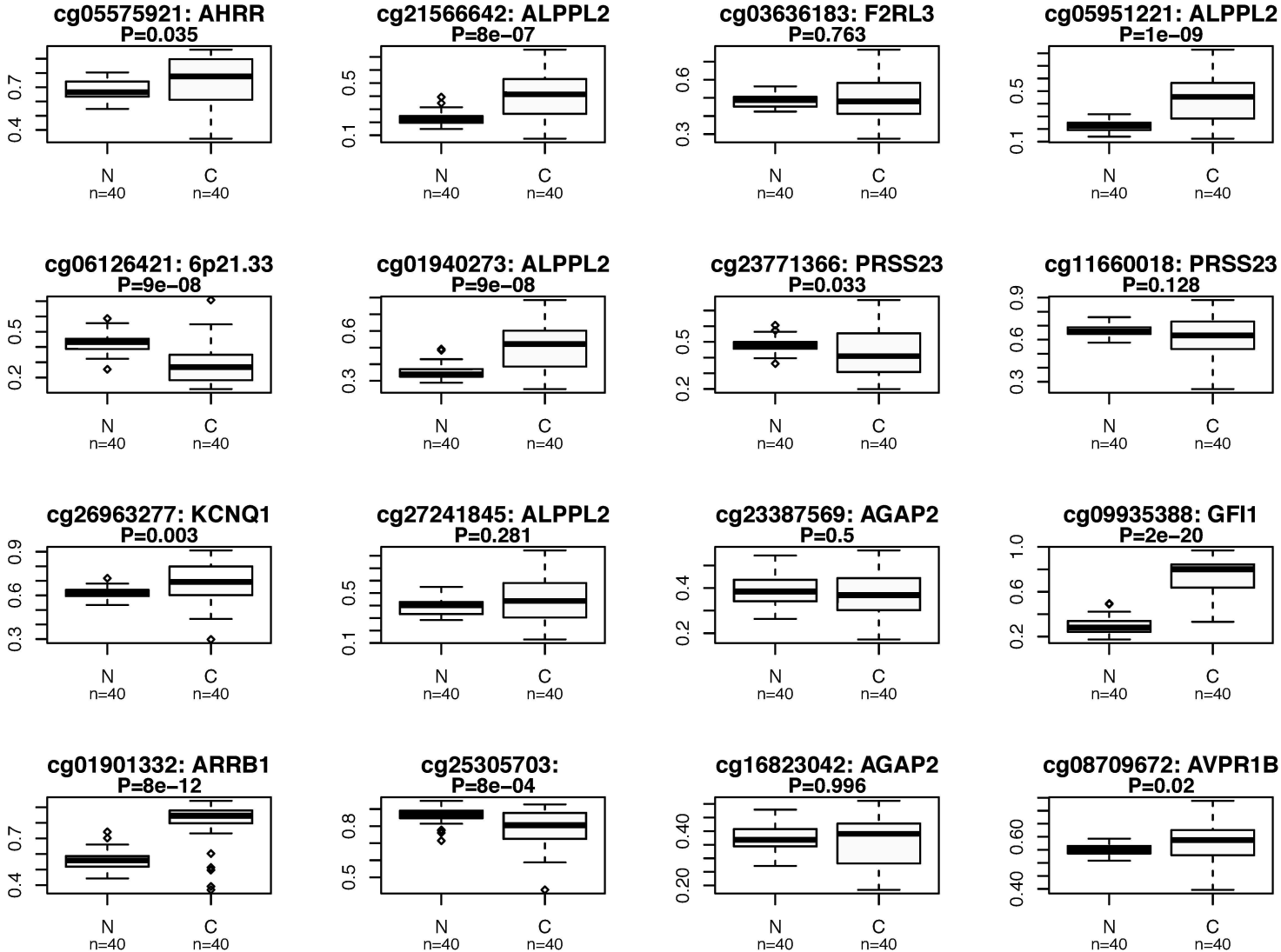

Meta-analysis of four lung cancer EWAS (918 cases and controls)
Identifies 16 CpG sites
Identify genetic instruments (SNPs) for CpG sites in ARIES (N=2 000*)
Obtain instruments for 14/16 sites
TRICL-ILLCO: 29 863 cases and 55 586 controls
Two sample MR of DNA methylation on lung cancer
Little evidence of association (FDR < 0.05)
Comparison of observational and MR results
Other sources of evidence
CCHS genotype and DNA methylation data: 357 cases and 8 401 controls
Previous literature linking AHRR and lung cancer
GTEx data
TCGA: 29 adenocarcinoma cases and 40 squamous cell carcinoma cases
One sample MR of AHRR methylation on lung cancer
Little evidence of association (FDR < 0.05)
Differences in methylation between cancerous and adjacent healthy tissue for the 16 sites
No association corresponding to observations in blood
SNP associations with gene expression in blood (N = 338) and in lung (N = 278)
Little evidence of association (FDR < 0.05)


15
10
5
0
-log10(P)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Chromosome


15
10
5
0
Observed $-\log_{10}(P)$
0
2
4
Expected $-\log_{10}(P)$


cg01901332*
ARRB1
cg27241845
ALPPL2
cg21566642*
ALPPL2
cg05951221
ALPPL2
cg05575921*
AHRR
cg23387569
AGAP2
cg16823042
AGAP2
cg23771366
PRSS23
cg11660018
PRSS23
cg26963277
KCNQ1
cg09935388
GFI1
cg03636183
F2RL3
cg25305703
CASC21
cg08709672*
AVPR1B
0.25
0.50
0.75
1.00
1.25
OR (95% CI)
EWAS_meta
MR


cg05575921: AHRR
P=0.963
cg21566642: ALPPL2
P=1e−04
cg03636183: F2RL3
P=8e−04
cg05951221: ALPPL2
P=0.008
cg06126421: 6p21.33
P=0.3
cg01940273: ALPPL2
P=0.002
cg23771366: PRSS23
P=4e−09
cg11660018: PRSS23
P=3e−08
cg26963277: KCNQ1
P=0.036
cg27241845: ALPPL2
P=0.076
cg23387569: AGAP2
P=0.877
cg09935388: GFI1
P=2e−04
cg01901332: ARRB1
P=3e−16
cg25305703:
P=3e−05
cg16823042: AGAP2
P=0.975
cg08709672: AVPR1B
P=0.03
N
n=29
C
n=29


cg05575921: AHRR
P=0.035
cg21566642: ALPPL2
P=8e−07
cg03636183: F2RL3
P=0.763
cg05951221: ALPPL2
P=1e−09
cg06126421: 6p21.33
P=9e−08
cg01940273: ALPPL2
P=9e−08
cg23771366: PRSS23
P=0.033
cg11660018: PRSS23
P=0.128
cg26963277: KCNQ1
P=0.003
cg27241845: ALPPL2
P=0.281
cg23387569: AGAP2
P=0.5
cg09935388: GFI1
P=2e−20
cg01901332: ARRB1
P=8e−12
cg25305703:
P=8e−04
cg16823042: AGAP2
P=0.996
cg08709672: AVPR1B
P=0.02
N
n=40
C
n=40