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

Thomas Battram, Rebecca Richmond, Laura Baglietto, Philip Haycock, Vittorio Perduca, Stig Bojesen, Tom Gaunt, Gibran Hemani, Florence Guida, Robert Carreras-Torres, et al.

International Journal of Epidemiology, 2019, 1493–1504
doi: 10.1093/ije/dyz190
Advance Access Publication Date: 24 September 2019
Original article

Mendelian Randomization

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

Thomas Battram [1,2*†] Rebecca C Richmond [1,2†] 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,[12] Børge G Nordestgaard,[5,6,7] Andrew E Teschendorff,[13,14,15] Silvia Polidoro,[16] Paolo Vineis,[16,17] Gianluca Severi,[18,19,20] Allison M Hodge,[19,20] Graham G Giles,[19,20] Kjell Grankvist,[21] Mikael B Johansson,[22] Mattias Johansson,[8] George Davey Smith[1,2$] and Caroline L Relton[1,2$]

[1]MRC Integrative Epidemiology Unit, [2]Population Health Sciences, University of Bristol, Bristol, UK, [3]Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy, [4]Laboratoire de Mathématiques Appliquées, Université Paris Descartes, Paris, France, [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]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, ON, Canada, [10]Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA, [11]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, USA, [12]Department of Community Medicine,Arctic University of Norway, Tromso, Norway, [13]Department of Women's Cancer, Institute for Women's Health, University College London, London, UK, [14]UCL Cancer Institute, University College London, London, UK, [15]Chinese Academy of Sciences (CAS) Key Laboratory of Computational Biology, CAS–Max Planck Gesellschaft (MPG) Partner Institute for Computational Biology, Shanghai, China, [16]Molecular and Genetic Epidemiology Unit, Italian Institute for Genomic Medicine (IIGM), Turin, Italy, [17]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK, [18]CESP (Inserm U1018), Facultés de Médicine Université Paris-Sud, UVSQ, Université Paris-Saclay, Gustave Roussy, 94805, Villejuif, France, [19]Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, VIC, Australia, [20]Centre for Epidemiology and Biostatistics, Melbourne School of Population & Global Health, University of Melbourne, Melbourne, VIC, Australia, [21]Department of Medical Biosciences, Clinical Chemistry and [22]Department of Radiation Sciences, Umeå University, Umeå, Sweden

*Corresponding author. MRC Integrative Epidemiology Unit, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK. E-mail: thomas.battram@bristol.ac.uk

[†]Joint first authors; [‡]joint second authors; [$]joint last authors.

Editorial decision 1 August 2019; Accepted 2 September 2019

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: Sixteen CpG sites were identified from the EWAS meta-analysis [false discovery rate (FDR) < 0.05], for 14 of which we could identify genetic instruments. Mendelian randomization provided little evidence that DNA methylation in peripheral blood at the 14 CpG sites plays 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.

Key words: 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 that 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 randomization analysis gave little evidence that any of the 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* loci, 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 socioeconomic 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: (i) the instrument is associated with the exposure; (ii) the instrument is not associated with any confounders; and (iii) 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 we applied MR to investigate whether the observed DNA methylation changes at these sites are causally linked to lung cancer.

Methods

EWAS meta-analysis

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

To quantify the association between the methylation level at each CpG and the risk of lung cancer, we fitted conditional logistic regression models for beta values of methylation [which ranges from 0 (no cytosines methylated) to 1 (all cytosines methylated)] on lung cancer status for the four studies. The cases and controls in each study were matched; details of this are in the Supplementary Methods, available as Supplementary data at *IJE* online. Surrogate variables were computed in the four studies using the SVA R package,[21] and the proportion of CD8+ and CD4+ T cells, B cells, monocytes, natural killer cells and granulocytes within whole blood were derived from DNA methylation.[22] The following EWAS models were included in the meta-analysis: Model 1—unadjusted; Model 2—adjusted for 10 surrogate variables (SVs); Model 3—adjusted for 10 SVs and derived cell proportions. Stratification of EWAS by smoking status was also conducted [never ($N = 304$), former ($N = 648$) and current smoking ($N = 857$)]. For Model 1, 2 and 3, 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 Models 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 effects 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 preprocessing, genotyping and quality control (QC) pipelines are outlined in the Supplementary Methods, available as Supplementary data at *IJE* online. 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]

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

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

MR analyses were also performed to investigate the impact of methylation on lung cancer subtypes in TRICL-ILCCO: adenocarcinoma (11 245 cases, 54 619 controls), small cell carcinoma (2791 cases, 20 580 controls) and squamous cell carcinoma (7704 cases, 54 763 controls). We also assessed the association in never smokers (2303 cases, 6995 controls) and ever smokers (23 848 cases, 16 605 controls).[25] Differences between the smoking subgroups were assessed using a Z test for difference.

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

Supplementary analyses

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

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

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

Tumour and adjacent normal methylation patterns

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

mQTL association with gene expression

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

Analyses were conducted in Stata (version 14) and R (version 3.2.2). For the two-sample MR analysis we used the MR-Base R package TwoSampleMR.[33] An adjusted *P*-value that limited the FDR was calculated using the Benjamini-Hochberg method.[34] All statistical tests were two-sided.

Results

A flowchart representing our study design along with a summary of our results at each step is displayed in Figure 1.

EWAS meta-analysis

The basic meta-analysis adjusted for study-specific covariates identified 16 CpG sites that were hypomethylated in

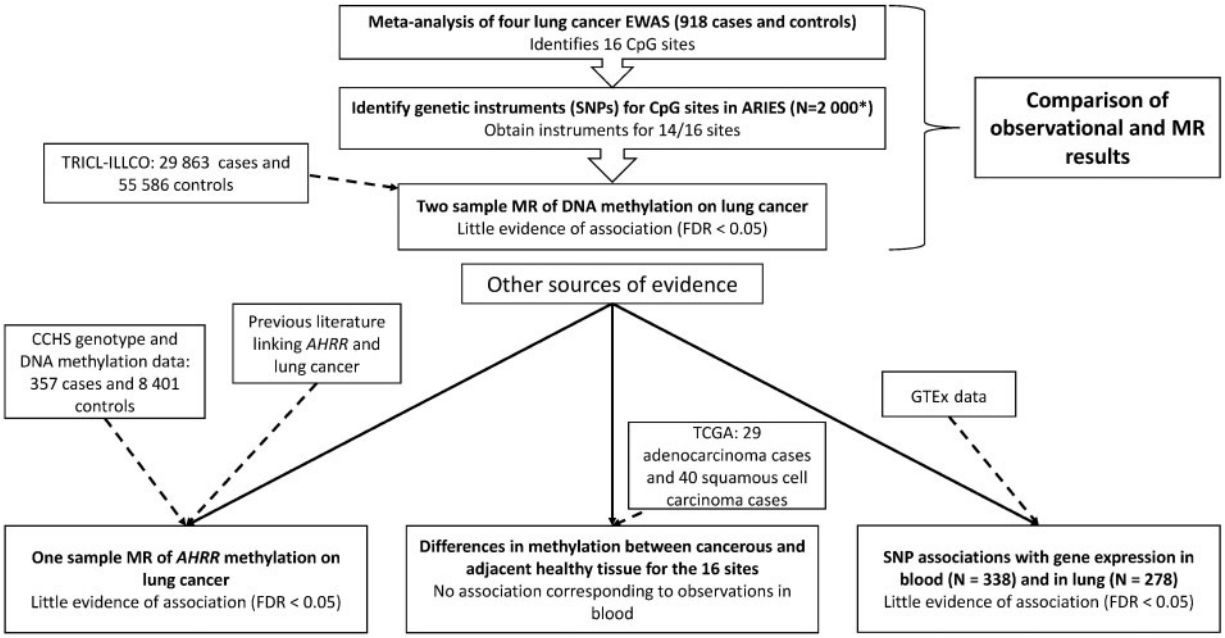

Figure 1. Study design with results summary. ARIES, Accessible Resource for Integrated Epigenomic Studies; TRICL-ILLCO, Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium; MR, Mendelian randomization; CCHS, Copenhagen City Heart Study; TCGA, The Cancer Genome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

relation to lung cancer (FDR < 0.05, Model 1, Figure 2). Adjusting for 10 surrogate variables (Model 2) and derived cell counts (Model 3) gave similar results (Table 1). The direction of effect at the 16 sites did not vary between studies (median I^2 = 38.6) (Supplementary Table 2, available as Supplementary data at *IJE* online), but there was evidence for heterogeneity of effect estimates at some sites when stratifying individuals by smoking status (Table 1).

Mendelian randomization

We identified 15 independent mQTLs (r^2<0.01) associated with methylation at 14 of 16 CpGs. Ten mQTLs replicated at FDR < 0.05 in NSHDS (Supplementary Table 3, available as Supplementary data at *IJE* online). MR power analyses indicated >99% power to detect ORs for lung cancer of the same magnitude as those in the meta-analysis of EWAS.

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

For nine of out the 16 mQTL-CpG associations, there was strong replication across time points (Supplementary Table 5, available as Supplementary data at *IJE* online) and 10 out of 16 mQTL-CpG associations replicated at FDR < 0.05 in an independent adult cohort (NSHDS). Using mQTL effect estimates from NSHDS for the 10 CpG sites that replicated (FDR < 0.05), findings were consistent with limited evidence for a causal effect of peripheral blood-derived DNA methylation on lung cancer (Supplementary Figure 1, available as Supplementary data at *IJE* online).

There was little evidence of different effect estimates between ever and never smokers at individual CpG sites (Supplementary Figure 2, available as Supplementary data at *IJE* online, Z test for difference, $P > 0.5$). There was some evidence for a possible effect of methylation at cg21566642-*ALPPL2* and cg23771366-*PRSS23* on squamous cell lung cancer {OR = 0.85 [95% confidence interval (CI)=0.75, 0.97] and 0.91 (95% CI = 0.84, 1.00) per SD (14.4% and 5.8%) increase, respectively} as well as methylation at cg23387569-*AGAP2*, cg16823042-*AGAP2*, and cg01901332-*ARRB1* on lung adenocarcinoma [OR = 0.86 (95% CI = 0.77, 0.96), 0.84 (95% CI = 0.74, 0.95), and 0.89 (95% CI = 0.80, 1.00) per SD (9.47%, 8.35%, and 8.91%) increase, respectively]. However, none of the results withstood multiple testing correction (FDR < 0.05) (Supplementary Figure 3, available as Supplementary data at *IJE* online). For those CpGs where multiple mQTLs were used as instruments (cg05575921-*AHRR* and cg01901332-*ARRB1*), there was limited evidence for heterogeneity in MR effect estimates (Q test, $P > 0.05$, Supplementary Table 6, available as Supplementary data at *IJE* online).

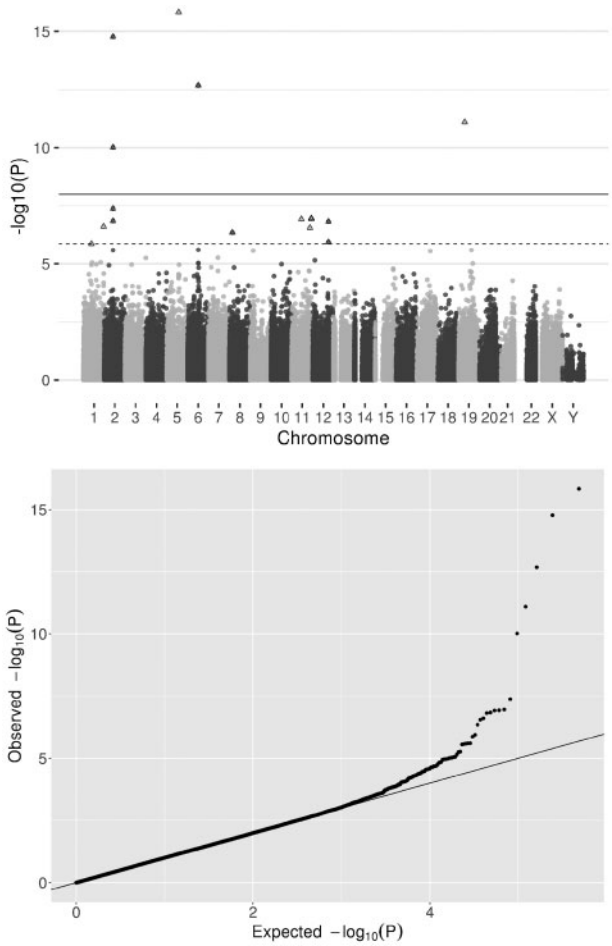

Figure 2. Observational associations of DNA methylation and lung cancer: a fixed effects meta-analysis of lung cancer EWAS weighted on the inverse variance was performed to establish the observational association between differential DNA methylation and lung cancer. a) Manhattan plot, all points above the solid line are at $P < 1 \times 10^{-7}$ 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 (a) Manhattan plot].

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

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

In the CCHS, a per (average methylation-increasing) allele change in a four-mQTL allele score was associated with a 0.73% (95% CI = 0.56, 0.90) increase in methylation ($P < 1 \times 10^{-10}$) and explained 0.8% of the variance in cg05575921-*AHRR* methylation (F statistic = 74.2). Confounding factors were not strongly associated with the genotypes in this cohort ($P \geq 0.11$) (Supplementary Table 7, available as Supplementary data at *IJE* online). Results provided some evidence for an effect of cg05575921 methylation on total lung cancer risk [hazard ratio (HR) = 0.30 (95% CI = 0.10, 1.00) per SD (9.2%) increase] (Supplementary Table 8, available as Supplementary data at *IJE* online). The effect estimate did not change substantively when stratified by smoking status (Supplementary Table 8, available as Supplementary data at *IJE* online).

Given contrasting findings with the main MR analysis, where cg05575921-*AHRR* methylation was not causally implicated in lung cancer, and the lower power in the one-sample analysis to detect an effect of equivalent size to the observational results (power = 19% at alpha = 0.05), we performed further two-sample MR based on the four mQTLs using data from both CCHS (sample one) and the TRICL-ILCCO consortium (sample two). 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, available as Supplementary data at *IJE* online). There was also limited evidence for an effect of cg05575921-*AHRR* methylation when stratified by cancer subtype and smoking status (Supplementary Figure 5, available as Supplementary data at *IJE* online) and no strong evidence for heterogeneity of the mQTL effects (Supplementary Table 9, available as Supplementary data at *IJE* online). Conclusions were consistent when MR-Egger[27] was applied (Supplementary Figure 5, available as Supplementary data at *IJE* online) and when accounting for correlation structure between the mQTLs (Supplementary Table 9, available as Supplementary data at *IJE* online).

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, available as Supplementary data at *IJE* online). 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.

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 + derived cell counts (cell count + SV adjusted, $N = 1809$).

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

Smoker group comparison = heterogeneity across meta-analyses when stratifying by smoking status.

Dir, direction of effect; OR, odds ratio per SD increase in DNA methylation; SE, standard error; Chr, chromosome.

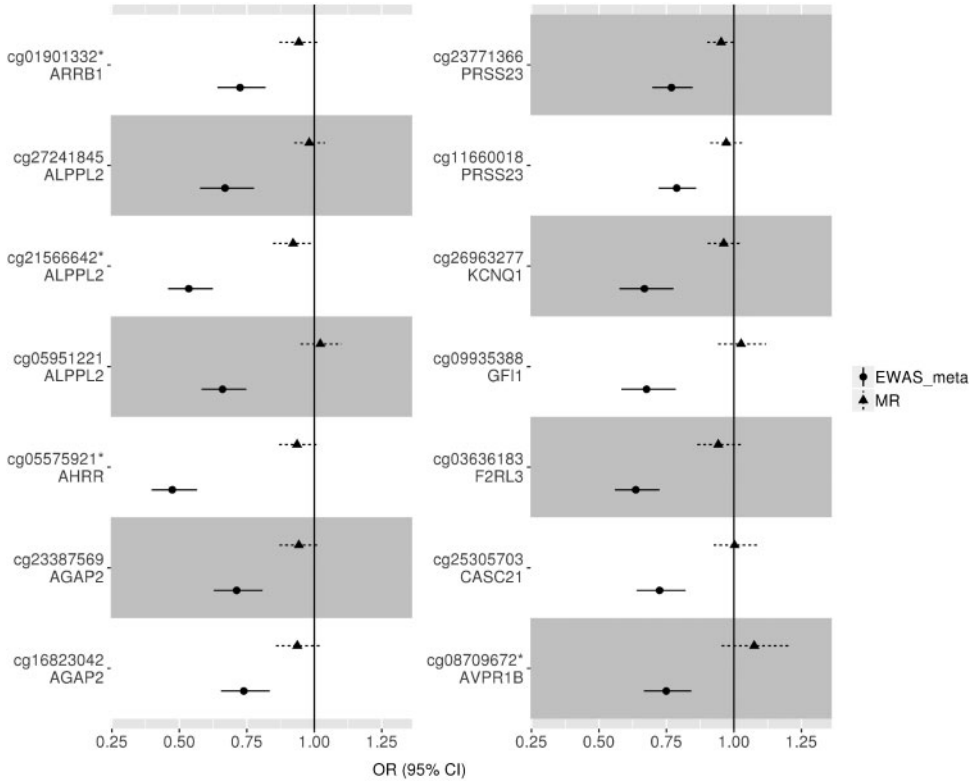

Figure 3. Mendelian randomization (MR) vs observational analysis. Two-sample MR was carried out with methylation at 14/16 CpG sites identified in the EWAS meta-analysis as the exposure and lung cancer as the outcome. cg01901332 and cg05575921 had two instruments, so the estimate was calculated using 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]. OR, odds ratio per SD increase in DNA methylation. *Instrumental variable not replicated in independent dataset (NSHDS). The sites for which instrumental variables have not been replicated are cg01901332, cg21566642, cg05575921 and cg08709672.

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, available as Supplementary data at *IJE* online).

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] Although 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%

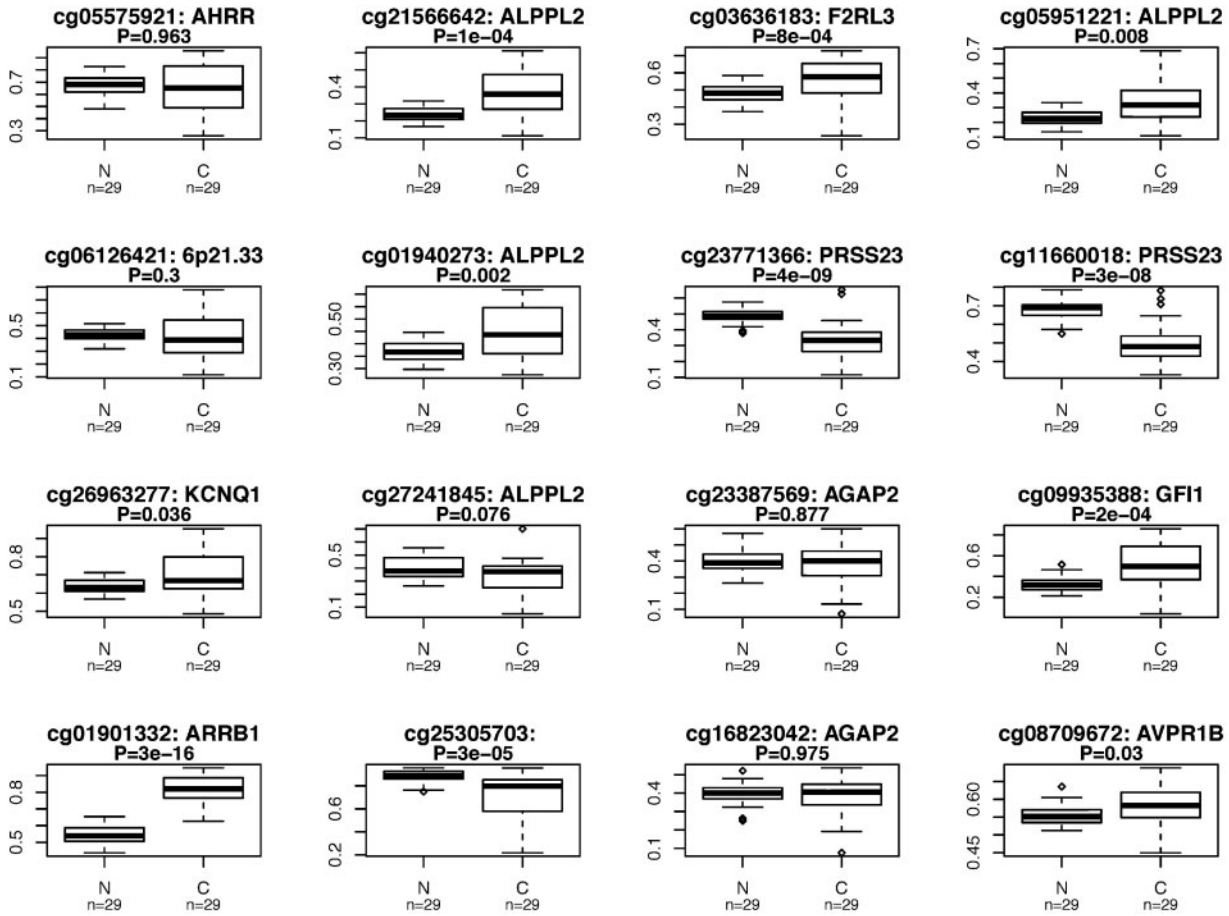

Figure 4. Differential DNA methylation in lung cancer tissue: a comparison of methylation at each of the 16 CpG sites identified in our meta-analysis was made between lung cancer tissue and adjacent healthy lung tissue for patients with: a) lung adenocarcinoma; and b) squamous cell lung cancer. Publicly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

power. Evidence against the observational findings was 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 that the potential impact of this bias was minimal (Supplementary Figure 1, available as Supplementary data at *IJE* online). 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 (four out of 16), there was limited replication across time points and in particular, six mQTLs were not strongly associated with DNA methylation in adults. Further, our primary data used for the first sample in the two-sample MR were ARIES, which contains no male adults. If the mQTLs identified vary by sex and time, then this could bias our results. However, our replication cohort NSHDS contains adult males. Therefore, the 10 mQTLs that replicated in NSHDS are unlikely to be biased by the sex discordance. Also, we replicated the findings for cg05575921 *AHRR* in CCHS, which contains both adult males and females, in a two-sample MR analysis, suggesting that these results also are not influenced by sex discordance. Caution is therefore warranted when interpreting the null results for the two-sample MR estimates for the CpG sites for which mQTLs were not replicated, 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 minimize 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,

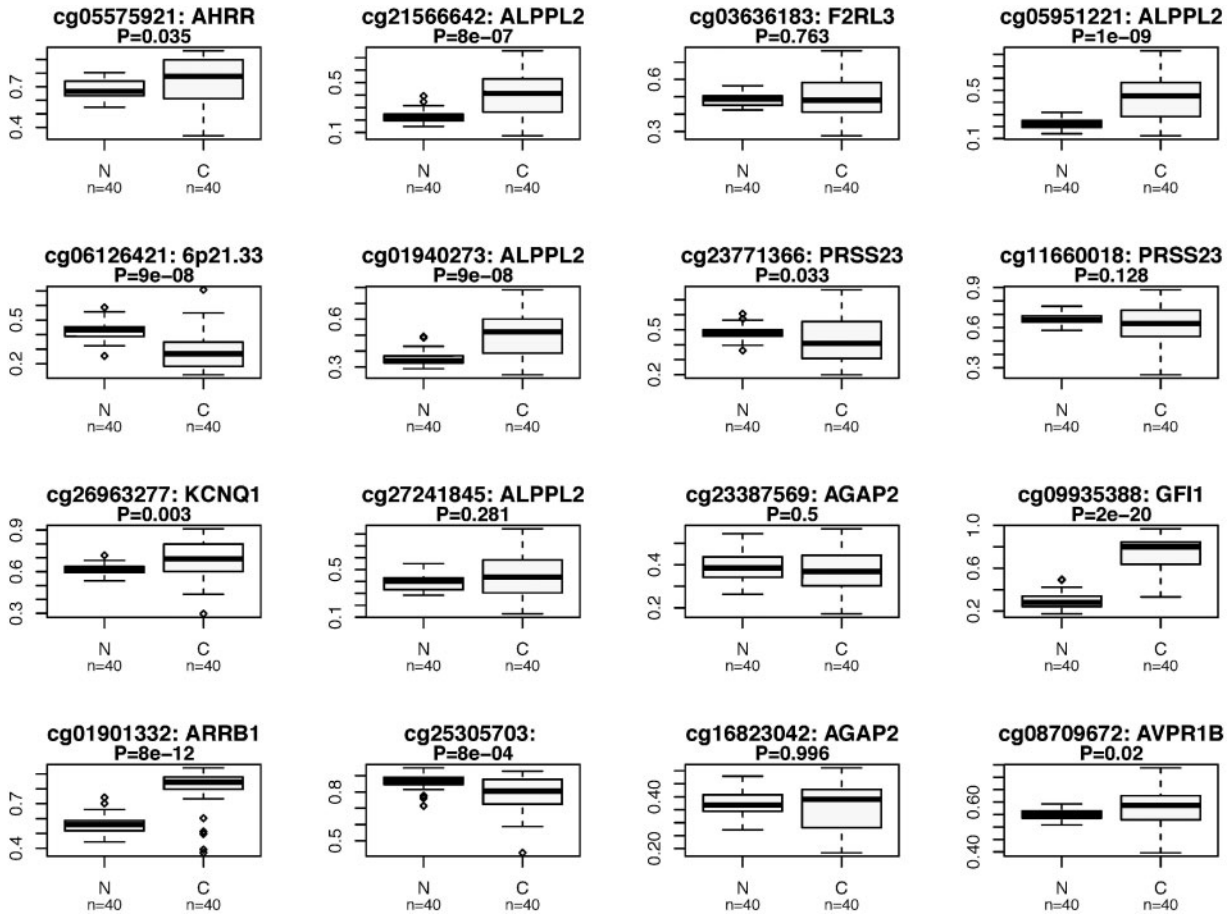

Figure 4. Continued.

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. 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], but in two-sample MR this site was not causally implicated [OR = 1.00 (95% CI = 0.83, 1.10) per SD increase]. We verified that this was not due to differences in the genetic instruments used, nor due to issues of weak instrument bias. Given that the CCHS one-sample MR had little power (19% at alpha = 0.05) to detect a causal effect with a size equivalent to that of the observational analysis, 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. A high degree of concordance in mQTLs has been observed across lung tissue, skin and peripheral blood DNA,[37] but 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, which 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. Whereas 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, although 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 does not give conclusive evidence that DNA methylation changes at these sites are not relevant to the development of lung cancer.

Whereas 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]

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was partly supported by a Wellcome Trust PhD studentship to T.B. (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_00011/1 and MC_UU_00011/5), which funds a Unit at the University of Bristol where T.B., R.C.R., P.C.H., T.R.G., G.D.S. and C.L.R. 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. 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].

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.

Author Contributions

This publication is the work of the authors and T.B., R.C.R. and C.L.R. will serve as guarantors for the contents of this paper.

Conflict of interest: None declared.

References

1. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C. *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11*. 2013. http://globocan.iarc.fr (9 December 2017, date last accessed).
2. Fasanelli F, Baglietto L, Ponzi E *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 *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**:50–61.
4. McCarthy S, Das S, Kretzschmar W *et al*. A reference panel of 64, 976 haplotypes for genotype imputation. *Nat Genet* 2016;**48**:1279–83.
5. Strathdee G, Brown R. Aberrant DNA methylation in cancer: potential clinical interventions. *Expert Rev Mol Med* 2002;**4**:1–17.
6. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 2002;**3**:415–28.
7. Borghol N, Suderman M, McArdle W *et al*. Associations with early life socioeconomic position in adult DNA methylation. *Int J Epidemiol* 2012;**41**:62–74.
8. Elliott HR, Tillin T, McArdle WL *et al*. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin Epigenetics* 2014;**6**:4.
9. Joehanes R, Just AC, Marioni RE *et al*. Epigenetic signatures of cigarette smoking. *Circ Cardiovasc Genet* 2016;**9**: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;**72**:646-53.
11. Zudaire E, Cuesta N, Murty V *et al*. The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers. *J Clin Invest* 2008;**118**: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**:R149–56.
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**:646–55.
14. Munafo MR, Timofeeva MN, Morris RW *et al*. Association between genetic variants on chromosome 15q25 locus and objective measures of tobacco exposure. *J Natl Cancer Inst* 2012;**104**:740–48.
15. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 2014;**23**:R89–98.

16. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Intl J Epidemiol* 2003;**32**: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**: 161–76.
18. Relton CL, Davey Smith G. Mendelian randomization: applications and limitations in epigenetic studies. *Epigenomics* 2015;**7**: 1239–43.
19. Richardson TG, Zheng J, Davey Smith G *et al*. Mendelian randomization analysis identifies CpG sites as putative mediators for genetic influences on cardiovascular disease risk. *Am J Hum Genet* 2017;**101**:590–602.
20. Gaunt TR, Shihab HA, Hemani G *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. http://www.genomine.org/sva/.
22. Houseman EA, Accomando WP, Koestler DC *et al*. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012;**13**:86.
23. Inoue A, Solon G. Two-sample instrumental variables estimators. *Rev Econ Stat* 2010;**92**: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**:1177–84.
25. McKay JD, Hung RJ, Han Y *et al*. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* 2017;**49**:1126.
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**:511–13.
27. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015;**44**: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**: 1423–26.
29. Tobacco and Genetics Consortium *et al*. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010;**42**:441–47.
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**:965–78.
31. Teschendorff AE, Yang Z, Wong A *et al*. Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. *JAMA Oncol* 2015;**1**:476–85.
32. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;**45**:580–85.
33. Hemani G, Zheng J, Wade KH *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 R Stat Soc B Methodol* 1995;**57**: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**:e1007081.
36. Burgess S, Thompson SG; CRP CHD Genetics Collaboration . Avoiding bias from weak instruments in Mendelian randomization studies. *Intl J Epidemiol* 2011;**40**:755–64.
37. Shi J, Marconett CN, Duan J *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**: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 Biol* 2017;**18**:236.
40. Gao X, Zhang Y, Breitling LP, Brenner H. Tobacco smoking and methylation of genes related to lung cancer development. *Oncotarget* 2016;**7**:59017–28.