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

Battram T[1,2]*, Richmond RC[1,2]*, Baglietto L[3†], Haycock P[1,2†], Perduca V[4], Bojesen S[5,6,7], Gaunt TR[1,2], Hemani G[1,2], Guida F[8], Carreras-Torres R[8], Hung R[9], Amos CI[10], Freeman JR[11], Sandanger TM[12], Nøst TH[13], Nordestgaard B[5,6,7], Teschendorff AE[14,15,16], Polidoro S[17], Vineis P[17,18], Severi G[19,20,21,22], Hodge A[22], Giles G[21,22], Grankvist K[23], Johansson MB[24], Johansson M[8], Davey Smith G[1,2$], Relton CL[1,2$]

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 Centre, 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

*joint first authors
[†]joint second authors
[$]joint last authors

Abstract

DNA methylation changes in peripheral blood have been identified in relation to lung cancer risk. However, the causal nature of these associations remains to be fully elucidated. Meta-analysis of four epigenome-wide association studies (918 cases, 918 controls) revealed differential methylation at 16 CpG sites (FDR < 0.05) in relation to lung cancer risk. A two-sample Mendelian randomization analysis, using genetic instruments for methylation at 14 of the 16 CpG sites, and 29,863 cases and 55,586 controls from the TRICL-ILCCO lung cancer consortium, was performed to appraise the causal role of methylation at these sites on lung cancer. This approach provided little evidence that DNA methylation in peripheral blood at the 14 CpG sites play a causal role in lung cancer development, including for cg05575921 *AHRR*, where methylation is strongly associated with lung cancer risk. Further studies are needed to investigate the causal role played by DNA methylation in lung tissue.

Background

Lung cancer is the most common cause of cancer-related death worldwide, estimated to be responsible for nearly one in five cancer-related deaths (1). Epigenetic changes have been implicated during the early stages of carcinogenesis and clonal expansion (2, 3) and several DNA methylation changes have been recently identified in relation to lung cancer risk (4-6). Interestingly, many of these identified methylation changes have also been associated with smoking, the most well-established risk factor for lung cancer. Therefore, DNA methylation may serve as a mediator of the influence of smoking on lung cancer, as previously reported (4); as an independent risk factor; or alternatively as a non-causal biomarker. Given the plasticity of epigenetic markers in response to modifiable exposures, any DNA methylation changes that are causally linked to lung cancer are potentially appealing targets for intervention (7, 8). However, these epigenetic markers are sensitive to reverse causation, being affected by cancer processes (8), and are also prone to confounding, for example by socio-economic and lifestyle factors (9, 10).

Epigenome-wide association studies (EWAS) for lung cancer have recently been conducted, with genome-wide DNA methylation measured using the Illumina Infinium® Human Methylation 450 BeadChip on DNA extracted from pre-diagnostic, peripheral blood samples (4, 5). The prospective design of these studies minimizes the potential impact of reverse causation, although it is possible that latent cancer undiagnosed at the time of blood draw impacted peripheral methylation changes in these individuals. One site that has been identified is cg05575921, within the aryl hydrocarbon receptor (*AHRR*) gene, which has been consistently replicated in relation to both smoking (11) and lung cancer (4, 5, 12). Functional evidence suggests that this region could be causally involved in lung cancer (13). 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 (14), including residual confounding after adjustment for self-reported smoking behaviour (15, 16).

One alternative approach to assess whether an association between exposure and disease reflects causation is the use of Mendelian randomization (MR) (17, 18). MR uses genetic variants robustly associated with modifiable factors as instruments to infer causality between the modifiable factor and disease outcome, overcoming most of unmeasured or residual confounding and reverse causation. MR may be adapted to the setting of DNA methylation (19-21) with the use of methylation quantitative trait loci (mQTLs), which are genetic variants that strongly correlate with the methylation state of nearby CpG sites (22). The degree of association of these mQTLs with lung cancer may be used to shed light upon the potential causal role of DNA methylation on lung cancer incidence.

In this study, we performed a meta-analysis of four lung cancer EWAS nested within prospective cohorts to identify CpG sites at which differential methylation is associated with lung cancer risk and applied Mendelian randomization to investigate whether the observed DNA methylation changes are causally linked to lung cancer.

Results

EWAS meta-analysis

To identify CpG sites at which methylation associated with lung cancer risk we conducted a meta-analysis of four EWAS (N = 1,836 (918 cases, 918 controls)). The basic meta-analysis adjusted for study-specific covariates identified 16 CpG sites associated with lung cancer at a false discovery rate (FDR) < 0.05 (Model 1) (**Figure 1**). Adjusting for 10 surrogate variables (Model 2) and derived cell counts (Model 3) gave similar results, but with slightly reduced effect sizes and larger P values at each of the 16 sites (**Table 1**). We also investigated heterogeneity in effect estimates between the four studies and found with the direction of effect did not vary between studies and the I^2 statistic ranged from 0 and 68.7, with a median of 38.6 for the top 16 sites in Model 1 (**Supplementary Table 1, Supplementary Figures 1-3**). We next stratified individuals by smoking status: never (N = 304), former (N = 648) and current smokers (N = 857); individuals without smoking status information (N = 27) were removed from the analysis. There was some evidence for heterogeneity in effect estimates at some of the CpG sites by smoking status (I^2 0.0-80.8%) (**Supplementary Table 2, Supplementary Figure 4**). At cg01901332 *ARRB1*, cg01940273 *ALPPL2*, and cg23771366 *PRSS23* the direction of effect differed in never smokers compared to former and current smokers, but it was consistent amongst the three strata for each of the other sites. The Tukey outlier identification method removed cg05951221 from NOWAC former smokers, while all other sites were retained.

mQTL analysis

From the 16 CpG sites identified in the EWAS meta-analysis, 14 had one or more corresponding mQTLs in the Accessible Resource for Integrated Epigenomics Studies (ARIES) ($P < 1x10^{-7}$) (**Supplementary Table 3**). Of the 1,097 mQTLs identified across the 14 CpG sites, 928 were present in an independent sample from NSHDS; 526 replicated at FDR < 0.05 and 98 replicated at a Bonferroni corrected threshold ($P < 0.05/928 = 5.39x10^{-5}$) (**Supplementary Table 4**). The 526 replicated SNPs were associated with 9 of the 14 CpG sites.

We next identified independent mQTLs (pruned for linkage disequilibrium [$r^2 < 0.01$]) for use as genetic instruments for DNA methylation in Mendelian randomization analyses (**Supplementary Table 5**). Replicated mQTLs were included where possible to reduce the effect of winner's curse, but for those CpGs without a replicated SNP-methylation association, we ran the analysis using the SNP-methylation associations identified in ARIES and have highlighted them in the tables and figures. Independent mQTLs were found to explain between 4 and 11% of the variance in methylation at the 14 CpG sites, and we had 100% power to detect each of the effect sizes from the EWAS (**Supplementary Table 6**).

Mendelian randomization

Two sample MR analysis was first conducted using mQTL-exposure effect estimates from ARIES (discovery) (**Supplementary Table 5**) and mQTL-outcome effect estimates from the Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium (TRICL-ILCCO), which has conducted genome-wide association study (GWAS) on lung cancer overall (29,863 cases, 55,586 controls) for individuals genotyped using the Illumina Infinium OncoArray-500K BeadChip (Illumina Inc. San Diego, CA) (23) and from prior studies, joined by meta-analysis.

There was little evidence for a causal effect of methylation at these 14 sites on lung cancer. For 9 out 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 (**Figure 2**); a Z-test gave evidence for a difference between the odds

ratios from the EWAS and MR analyses at each site (P < 0.001, **Supplementary Table 7**). Using SNP effect estimates from NSHDS for the 9 CpG sites that replicated (FDR < 0.05) (**Supplementary Tables 4 and 5**), findings were consistent with limited evidence for causal effect of peripheral blood-derived DNA methylation on lung cancer.

Further secondary MR analyses were performed for lung cancer among never and ever smokers separately (**Supplementary Figure 5**), as well as in 3 lung cancer subtypes: adenocarcinoma, squamous cell carcinoma and small cell carcinoma (**Supplementary Figure 6**). There was little evidence of differences between ever and never smokers at individual CpG sites (Z-test, P > 0.5). There was some evidence for a possible causal effect of methylation at cg21566642 *ALPPL2* and cg23771366 *PRSS23* on squamous cell lung cancer (OR = 0.85, [95% CI, 0.75; 0.97] and 0.91 [95% CI, 0.84; 1.0] per SD [14.4% and 5.8%] increase in methylation 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 in methylation respectively). However, none of the results withstood multiple testing correction (FDR < 0.05). Full results for this MR analysis can be found in **Supplementary Table 8**. For those CpGs where multiple mQTLs were used as instruments (cg05575921 *AHRR* and cg01901332 *ARRB1*), there was limited evidence for heterogeneity in causal effect estimates (Q-test, P = 1) (**Supplementary Table 9**).

Using data from the Tobacco and Genetics (TAG) consortium, we investigated whether our mQTLs were associated with four separate smoking behaviours: age of initiation, cigarettes smoked per day, smoking initiation (ever vs. never smoked) and smoking cessation (former vs. current smoker). This could violate the assumptions of MR that the genetic instruments are independent of confounding factors and are not subject to horizontal pleiotropy (**Supplementary Figure 7**). For most mQTLs, there was no clear evidence of association with the different smoking behaviours. However, single mQTLs for cg05575921 *AHRR,* cg27241845 *ALPPL2,* and cg26963277 *KCNQ1* showed some evidence for an association with smoking cessation (former vs. current smokers) (log(OR) = 0.059 [95% CI, 0.011; 0.106], -0.076 [95% CI, -0.131; -0.020], and -0.0363 [95% CI, -0.071; -0.002]) respectively, although these associations did not withstand multiple testing correction (FDR < 0.05).

Potential causal effect of *AHRR* methylation on lung cancer risk

Similar to previous studies analysing a number of the cohorts included here (4, 5), our meta-analysis found cg05575921 *AHRR* was most strongly associated with lung cancer risk (OR = 0.474 [95% CI, 0.397; 0.566] per SD increase in methylation). In contrast, there was little evidence for a strong causal effect of methylation at cg05575921 on lung cancer in the main MR analysis (OR = 0.921 [95% CI, 0.971; 0.998] per SD increase in methylation). However, this result was based on two mQTLs identified in ARIES that showed little evidence for replication in NSHDS (replication P = 0.629 and 0.243).

We chose to further investigate the possible causal effect of methylation at this site on lung cancer risk, using both one-sample and two-sample MR (24), using individual-level data from the Copenhagen City Heart Study (CCHS) and summary statistics from CCHS and TRICL-ILCCO, respectively.

For the one-sample MR analysis, 4 mQTLs were combined into an allele score ($r^2 < 0.2$). A per (average methylation-increasing) allele change in the 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 methylation in the CCHS (F statistic = 74.2) (**Supplementary Table 10**). Confounding factors were not strongly associated with the genotypes in this cohort (P values ≥ 0.11) (**Supplementary Table 11**).

In 8,758 participants in the CCHS (357 incident lung cancer cases), results provided some evidence for an effect of DNA methylation of cg05575921 *AHRR* on total lung cancer risk (HR = 0.30 [95% CI 0.10, 1.00] per SD (9.2%) increase in methylation) (**Supplementary Table 12**). Furthermore, this effect

estimate did not change substantively when individuals were stratified by smoking status (current, former, never) (**Supplementary Table 12**).

Given contrasting findings with the main MR analysis, we performed further two-sample MR based on the four independent SNPs associated with methylation at cg05575921 using data from the TRICL-ILCCO consortium. Results of the two-sample MR approach showed no strong evidence for a causal effect of DNA methylation on total lung cancer risk (OR = 1.00 (0.83, 1.10) per SD increase in methylation in random effects meta-analysis) (**Supplementary Figure 8**), indicating that DNA methylation at *AHRR* does not appear to be mediating the effect of smoking on all lung cancers. There was also limited evidence for a causal effect of *AHRR* methylation (**Supplementary Figure 8**) when stratified by cancer subtype and smoking status (**Supplementary Figure 8**). There was no strong evidence for heterogeneity of the SNP effects in the MR analysis (**Supplementary Table 13**) and conclusions were consistent when MR-Egger was used in sensitivity analyses (**Supplementary Figure 8**) and when a weighted generalized regression method was used to account for correlation structure between the SNPs (**Supplementary Table 14**).

Tumour and adjacent normal lung tissue methylation patterns

To investigate how methylation changes observed in peripheral blood in relation to lung cancer corresponded to changes in lung tissue, we analysed methylation differences at all 16 CpG sites of interest in lung cancer tissue and adjacent healthy lung tissue from the same patient using data from The Cancer Genome Atlas (TCGA) (n=40 lung squamous cell carcinoma and n=29 lung adenocarcinoma). For cg05575921 *AHRR,* there was no strong evidence for differential methylation between lung adenocarcinoma tissue and adjacent healthy tissue (P = 0.963), and weak evidence for hypermethylation in squamous cell lung cancer tissue compared to normal tissue (P = 0.035). All results can be found in **Figure 3** and a comparison between the MR analysis and these results can be found in **Supplementary Table 15**.

Gene expression associated with mQTLs in blood and lung tissue

Data from the Gene-Tissue Expression (GTEx) consortium was next used to evaluate the influence of the identified mQTLs on gene expression in blood and lung tissue. Of the 10 genes annotated to the 14 CpG sites, 8 genes were expressed sufficiently to be detected in lung (*AVPR1B* and *CASC21* were not) and 7 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, which are mQTLs for cg21566642 and cg05951221 respectively, were associated with increased expression of *ALPPL2* (P = 0.002 and P = 0.0001 respectively). 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 Tables 16-17).**

Discussion

Main findings

In this study, we verified findings from previous analyses which have identified DNA methylation changes in relation to incident lung cancer, predominantly at CpG sites previously implicated in relation to smoke exposure (4, 5). This meta-analysis of lung cancer EWAS nested within four prospective cohorts identified 16 CpG sites associated with risk of lung cancer, 14 of which have been previously identified in relation to smoke exposure (11) and 6 were highlighted in a previous study as being associated with lung cancer (5). This previous study used the same data from the four cohorts investigated here, but in a discovery and replication, rather than meta-analysis framework, which limited the number of CpG sites identified.

We further evaluated evidence for a causal effect of DNA methylation on lung cancer risk using Mendelian randomization. The data presented provide no supporting evidence that DNA methylation in peripheral blood at CpG sites identified in the lung cancer EWAS play a causal role in lung cancer development. These findings are in contrast to previous analyses which suggested that methylation at two CpG sites investigated (*AHRR* and *F2RL3*) mediated a large proportion of the effect of smoking on lung cancer risk (4). This previous study used methods which are sensitive to residual confounding and measurement error that may have biased effect estimates (14, 25). Furthermore, despite the longitudinal design, presence of subclinical disease could have also made the results liable to reverse causation. These limitations are largely overcome with the use of instrumental variables in the Mendelian randomization approach used here (14).

AHRR methylation

AHRR methylation is associated with smoke exposure in a range of tissue types (26-28), has been found to be associated with gene expression in lymphoblasts, pulmonary macrophages and lung tissue (26, 28, 29), and persists many years after smoking cessation (29-32). These features indicate that methylation at this site may be biologically relevant and functional work has implicated *AHRR* expression (and methylation status) in mediating the detoxification of polycyclic aromatic hydrocarbons, which are principle carcinogenic agents in tobacco smoke (29).

In a previous work that used mediation analysis, it was estimated that 32% of the total effect of smoking on lung cancer risk was mediated by differential methylation in the *AHRR* region implicated (4). cg05575921 *AHRR* is located in a gene enhancer and methylation at this site has been found to be inversely associated with gene expression in lymphoblasts (26). If this inverse relationship with expression extrapolates to other tissue types, including lung tissue, then previous findings would have implicated upregulation of *AHRR* in mediating the adverse effect of smoking on lung cancer risk (4). However, this conflicts with a previous *in vitro* study that found *AHRR* to be a putative tumour suppressor gene that was downregulated in lung cancer cell lines (13).

One possible explanation for the inconsistency between the observational and in vitro results is that residual confounding from tobacco smoking could explain the inverse association between *AHRR* methylation and lung cancer seen observationally and this possibility could not be fully excluded using the statistical methods applied in the previous study (4, 14). Results of this MR analysis suggest that differential methylation at *AHRR* does not have a strong causal effect on lung cancer, and the results are in stark contrast with the observational analysis. MR analysis provided some evidence for an effect of *AHRR* methylation on lung cancer within the Copenhagen City Heart Study in line with the observational analysis. However, when we applied a two-sample MR analysis to improve power and enable histological stratification of the analysis using summary GWAS data from the TRICL-ILCCO consortium, effect estimates were null and there was evidence of a difference compared with

observational findings. One possible explanation for a lack of causal effect 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 in squamous cell lung cancer tissue compared to normal tissue. 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) (33).

Null MR associations between cg05575921 *AHRR* methylation and lung cancer does not exclude the pathway from being involved in the disease process. AHRR and AHR form a regulatory feedback loop, with AHR acting to overexpress its repressor AHRR, which means that the actual effect of differential methylation or differential expression of *AHR*/*AHRR* on pathway activity is complex (34).Thus, the AHR pathway may still be causally implicated, with the CpGs mapping to *AHRR* merely reflecting a response to smoking exposure.

Methylation at other CpG sites

While there was some evidence for a causal 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.

Strengths

Major strengths of this analysis include the use of MR approaches to integrate an extensive, population-based epigenetic resource (ARIES) and summary data from a large lung cancer consortium GWAS (TRICL-ILCCO), in order to appraise the causal effect of methylation on lung cancer. The large GWAS from a lung cancer consortium ensured the MR analyses were 100% powered to detect each of the effect sizes observed in the EWAS.

In particular, we showed that the mQTLs used as instrumental variables were associated with methylation in our discovery sample (ARIES) and in replication cohorts (NSHDS and CCHS). We also found that for the most part, these genetic variants were not strongly associated with smoking behaviour, which was hypothesized to be the most likely confounding factor in this context.

We were also able to appraise the functional relevance of our findings for methylation in peripheral blood by investigating methylation patterns between tumour and adjacent normal lung tissue samples from The Cancer Genome Atlas. Potentially causal DNA methylation alterations would mark cells that become selected for and therefore enriched during tumorigenesis (27, 35). For the most part, this pattern was not observed for the CpG sites investigated and therefore were largely consistent with findings from the MR analysis, i.e. in accordance with no strong causal effect of methylation on lung cancer at the top CpG sites from the EWAS study, with the exception of *PRSS23* where methylation levels did appear to differ between normal versus lung cancer samples in the same direction as the MR and EWAS analyses.

Furthermore, we investigated expression of those genes annotated to the CpG sites identified in both blood and lung tissue. For the most part, mQTLs did not strongly influence expression within either tissue. This was with the exception of *ALPPL2* where two of the mQTLs were strongly related to expression in lung tissue, although methylation at this site appeared to have little impact on lung cancer risk.

Limitations

The MR analysis was limited by the lack of mQTLs for some of the identified CpG sites. Furthermore, only one or two mQTLs were identified at the threshold of $P < 1x10^{-7}$ after LD pruning. While these mQTLs explained a reasonable proportion of variation in methylation, the lack of instruments did not allow us to appraise potential pleiotropy for most of the analyses, using methods such as MR-Egger (36). This was a concern given that we had included some trans-SNPs as genetic instruments, which are typically thought to influence methylation at more CpG sites than cis-SNPs do, and could be potentially pleiotropic. Where possible we omitted these SNPs, but for cg01901332 and cg08709672 only trans-SNPs were available, which has been highlighted in **Supplementary Table 3/5**. However, analysis of *AHRR* methylation when we relaxed the criteria for mQTL selection and used more genetic instruments as part of an allele score, allowed us to start to appraise pleiotropy in this context and showed similar findings.

In two-sample MR, we primarily used SNP-methylation estimates from ARIES as this was our largest mQTL dataset. However, "winner's curse" 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 (37). To minimize the risk of this, we used SNP-methylation estimates from mQTLs that replicated in an independent cohort (NSHDS) in secondary analysis. Despite some potential sample overlap between NSHDS individuals and those in the TRICL-ILCCO dataset, which could bias results from two-sample MR towards the confounded observational association, the two sample MR analysis using the NSHDs mQTLS were similar to the findings based on ARIES, indicating the potential impact of the bias from either winner's curse or sample overlap was minimal (**Supplementary Figure 9**).

To investigate potential residual confounding of the MR analysis, we investigated genetic associations between the mQTLs and smoking status. While there was no strong evidence for SNP-smoking effects, this was with the exception of mQTLs being used to instrument cg05575921 *AHRR,* cg27241845 *ALPPL2* and cg26963277 *KCNQ1* methylation, which were nominally associated with smoking phenotypes. These associations may represent: true confounding of the methylation-lung cancer MR analysis by smoking, the causal influence of methylation on smoking status, pleiotropy of SNPs influencing both methylation and smoking, or chance associations. Nonetheless, these associations were nominal and not strong enough to bias our MR findings substantially. We have not assessed the potential confounding of the two-sample MR analysis by the other factors that were assessed in the one-sample setting (sex, alcohol consumption, occupational exposure to dust and/or welding fumes, and passive smoking). This is because GWAS summary statistics are not available for many of these other traits, and we had hypothesized that smoking would be the strongest confounding factor in the study.

Other possible limitations of MR include population stratification and canalization (17, 18, 38). Major population stratification is unlikely since all studies investigated here consist only of individuals of European ancestry. Canalization (or developmental compensation) could potentially bias the Mendelian randomization effect estimates towards the null and thus may explain discrepancies between MR and the observational results. However, the magnitude of bias would need to be very large to fully explain the differences.

This study predominantly investigated effects of methylation levels in peripheral blood, which may not be the most appropriate tissue in which to study causal associations with lung cancer and a more pronounced effect may be observed in buccal cells (a source of squamous epithelial cells) (27) and lung tissue (39). While a high degree of concordance in mQTLs has been observed across lung tissue, skin and peripheral blood DNA (40), we were unable to directly evaluate this here. However, it was of interest that the mQTLs we identified were not strongly related to gene expression in lung tissue, which might have explained some of our null findings. Further studies should therefore consider direct

links between methylation and expression in tissues of interest in prioritizing potentially causal CpG sites.

This study only assessed the causal effect of 14 single CpG sites. While some of the identified CpG sites reside in the same gene regions, our findings are typically not generalizable to methylation across the whole gene and it remains possible that other CpG sites do have a causal effect on lung cancer risk. For example, alternate sites (not cg05575921) within the *AHRR* region are correlated with *AHRR* expression and knockdown of *AHRR* was correlated with greater lung tumor cell invasiveness (13). Furthermore, if lung cancer risk is associated with a large number of DNA methylation changes across the genome, with changes at each CpG site contributing little to the increase in risk of disease then our power to detect differential methylation at individual sites is very low.

Some of the mQTLs used influence multiple CpGs in the same region, for example rs1048691 was used to instrument differential methylation at both cg23387569 and cg16823042 *AGAP2*, suggesting genomic control of methylation at a regional rather than single CpG level. Thus, methods to detect differentially methylated regions (DMRs) and identify genetic variants which proxy for them may be fruitful in probing the causal effect of methylation across gene regions rather than at single CpG sites.

Conclusion

While methylation at multiple CpG sites are observationally associated with lung cancer risk in EWAS meta-analysis, Mendelian randomization suggests that methylation at these sites is not causally implicated in lung cancer. Therefore, while methylation may be predictive of lung cancer risk (especially as it is a strong biomarker of smoking (4, 12)), 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 (14). Rather, MR, which takes genetic variants randomized with respect to confounding factors as proxies for an exposure of interest (i.e., DNA methylation), may be used to identify causal effects, or reject spurious findings.

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 cancers) (41). In addition, it is possible that DNA methylation at these CpG sites in lung tissue has a causal effect on lung cancer, which we were unable to directly evaluate here. Extending this robust causal modelling approach to other CpG sites recently reported to have been linked to tobacco smoke (11) is advocated.

Methods

EWAS Meta-analysis

This study involved a re-analysis of four lung cancer EWAS datasets previously investigated as part of a discovery and replication (look up) design (5). To identify CpG sites at which differential methylation associates with lung cancer risk with greater power (42), we conducted a meta-analysis of these four EWAS that assessed DNA methylation using the Illumina Infinium HumanMethylation450 BeadChip, which targets 485,512 CpG sites across the genome. All four studies are case-control studies 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).

The study populations, laboratory methods, data pre-processing and quality control methods for each of the four lung cancer studies have been described in detail elsewhere (5), but are briefly outlined in the **Supplementary Methods**. At the various laboratory sites, samples were distributed into 96-well plates and processed in chips of 12 arrays (8 chips per plate) with case-control pairs arranged randomly on the same chip. Methylation data were pre-processed and normalized in each study, and probe filtering was performed as previously described (5), leaving 465,886 CpGs suitable for the analysis in EPIC-Italy, 485,330 CpGs in MCCS, 450,890 CpGs in NOWAC and 482,867 CpGs in NSHDS. The total number of CpGs that were common across all 4 studies was 447,606.

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. Study-specific covariates had been adjusted for in each individual EWAS based on matching characteristics (**Supplementary Methods**), which included matching cases and controls on smoking status in two studies (MCCS and NSHDS). For EPIC-Italy and NOWAC, smoking status was included as a covariate in the EWAS model Surrogate variables were computed in all of the four studies using the SVA R package (43) and the proportion of CD8+ and CD4+ T cells, B cells, monocytes, natural killer cells, and granulocytes within whole blood were estimated using a method proposed by Houseman et al (44).

We performed an inverse-variance weighted fixed effects meta-analysis of the EWAS analysis in the four studies using the METAL software (http://csg.sph.umich.edu/abecasis/metal/), which computes pooled effect estimates, standard errors and P values for each CpG site. Direction of effect, effect estimates and the I^2 statistic were used to assess heterogeneity across the four studies. In total, the analysis consisted of 918 case-control pairs from the 4 prospective studies. When stratifying by smoking status

Before conducting the meta-analysis, all sites with missing values (due to non-convergence of models) in any of the cohorts were removed which included, in EPIC-Italy, 13 sites among never smokers, 300 among former smokers and 1326 from the current smokers, and in NOWAC, 2666 sites from never smokers, 793 from the former smokers, and 0 from the current smokers. A heterogeneity test was conducted to determine if there were differences between the effect estimates within each of the smoking groups.

Sites were taken forward from the unadjusted model (Model 1) with a false discovery rate (FDR) < 0.05, calculated using the Benjamini-Hochberg method (45), and examined in the other models. The Tukey outlier identification method (outlier = outside of lower/upper quartile -/+ 3 x interquartile range) was also applied to assess whether any of the top hits were outliers.

mQTL identification

Accessible Resource for Integrated Epigenomics Studies (ARIES)

We identified mQTLs for each CpG site of interest from an mQTL database (http://www.mqtldb.org), based on a genome-wide association study (GWAS) of DNA methylation in 1,018 mother-offspring pairs taken at multiple time points in ARIES (22, 46). Cord blood and peripheral blood samples (white blood, buffy coats or blood spots) in ARIES were collected according to standard procedures. Further details on the methylation pre-processing and quality control (QC) pipeline are outlined in the **Supplementary Methods**. The ARIES participants were previously genotyped as part of the larger Avon Longitudinal Study of Parents and Children (ALSPAC) study (http://www.bristol.ac.uk/alspac), with quality control, cleaning and imputation performed at the cohort level before extraction of the subset comprising ARIES. Further details on the genotyping and QC pipeline and the analysis are outlined in the **Supplementary Methods**. All associations at $P < 1x10^{-7}$ are stored in the mQTL database. We used this resource to identify genetic variants that can be used as proxy for the top CpG sites from the EWAS meta-analysis.

Northern Sweden Health and Disease Study (NSHDS)

We attempted to replicate the mQTLs identified in ARIES in an independent cohort, the Northern Sweden Health and Disease Study (NSHDS). If there was evidence for a SNP-CpG site association in the ARIES cohort (**Supplementary Methods**) in at least one time-point, we re-analysed this association using linear regression of methylation on each genotyped SNP available in the NSHDS, using rvtests (47). This prospective study, included in our EWAS meta-analysis, has genetic data as well as DNA methylation data available for 463 of its participants.

The NSHDS samples were genotyped using the Illumina Infinium OncoArray-500k BeadChip (Illumina Inc. San Diego, CA) and quality control parameters were applied under the recently published TRICL-ILCCO GWAS study on lung cancer (23). Genetic imputation was performed on these samples using the Haplotype Reference Consortium (HRC) Panel (release 1) (48) through the Michigan Imputation Server (49).

Mendelian randomization

To establish the potential causal effects of differential methylation on lung cancer risk we utilized a two-sample MR approach (50, 51). In this approach, information on the SNP-exposure (here DNA methylation) and SNP-outcome (here lung cancer) effects are derived from separate studies. As described above, ARIES and the NSHDS were used to estimate the SNP-methylation effects (i.e. the aforementioned mQTLs). 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) Consortium, to estimate the effect of each mQTL on risk of lung cancer. Summary data were also available for lung cancer subtypes, including squamous cell lung cancer, small cell lung cancer and lung adenocarcinoma, as well as for never and ever smokers separately. Power analyses indicated greater than 95% power to detect odds ratios (ORs) for lung cancer of the same magnitude as those detected in the EWAS meta-analysis of DNA-methylation and lung cancer. Furthermore, we calculated the maximum and minimum ORs we could detect for lung cancer with 80% power, using mQTL effect estimates from both ARIES and NSHDS cohorts (**Supplementary Table 6**).

Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium (TRICL-ILCCO)

Summary-level SNP effect estimates for lung cancer were obtained from the TRICL-ILCCO consortium, which has conducted GWAS on lung cancer overall (29,863 cases, 55,586 controls) for individuals genotyped using the Illumina Infinium OncoArray-500K BeadChip (Illumina Inc. San Diego, CA) and independent samples for which prior genotyping was performed.(23)

For the CpG sites identified in the EWAS meta-analysis which were associated with lung cancer at FDR < 0.05, we performed a look-up of the identified mQTLs in the lung cancer GWAS summary data from the TRICL-ILCCO consortium. We extracted the following summary data for each SNP: the log odds ratio for lung cancer per copy of the effect allele and its standard error, the effect allele and the reference allele and effect allele frequency. We combined information on the SNP-outcome associations from TRICL-ILLCO with information on the SNP-exposure associations from ARIES in instrumental variable analysis, described below.

The mQTLs were pruned for linkage disequilibrium ($r^2 < 0.01$) to prevent underestimation of standard errors. For each SNP, we calculated the log odds ratio per unit increase in methylation by the formula β_{GD}/β_{GP} (also known as a Wald ratio), where β_{GD} is the log odds ratio for disease per copy of the effect allele and β_{GP} is the standard deviation increase in methylation per copy of the effect allele. **Supplementary Table 18** illustrates the % methylation equivalent of a SD increase in methylation at each of the identified CpG sites. Standard errors of the Wald ratios were approximated by the delta method (52). Where multiple independent mQTLs were available for the same 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 SNPs was estimated using Cochran's Q test. The same two-sample MR analysis was conducted using SNP-exposure effect estimates from ARIES (discovery) and SNP-exposure effect estimates from NSHDS (replication).

Sensitivity analysis: Testing MR assumptions

The assumptions of the MR approach are: 1) the genetic instrument is associated with the exposure; 2) the genetic instrument is not related to confounding factors for the exposure-outcome association; and 3) the genetic instrument is related to the outcome only through its effect on the exposure (18). If these assumptions are true, then any association observed between the genetic instrument and outcome is best explained by a true causal effect of the exposure on the outcome. Here genetic instrument, exposure and outcome refer to mQTL, methylation and lung cancer, respectively.

To assess the 1st MR assumption and appraise instrument strength for each of the CpGs being investigated, we obtained an r^2 statistic for the proportion of variation in methylation explained by each mQTL identified in ARIES using NSHDS as an independent dataset.

A previous study found that mQTLs at smoking-related CpG sites have null associations with active smoking and therefore have the potential to be used as valid genetic instruments in MR analysis (53). Here we investigated the extent to which the mQTLs at cancer-related CpGs were associated with smoking behaviour, which could implicate confounding of the mQTL effect by smoking or horizontal pleiotropy, whereby the mQTLs have a causal effect on smoking (and thus lung cancer) independent of its effect on methylation, using summary genetic data on a large number of individuals as part of the Tobacco and Genetics (TAG) consortium.

The TAG consortium conducted a GWAS meta-analysis of four smoking traits: number of cigarettes per day, smoking cessation rate, smoking initiation and age of smoking initiation, using data from 16 cohorts and 74,053 individuals. Full details of the GWAS methods are described elsewhere (54). mQTL associations with the four smoking behaviours were obtained from the TAG consortium GWAS summary statistics, and two-sample MR analysis performed.

While various other sensitivity analyses exist for investigating possible pleiotropy in MR analysis (36), this approach relies upon the existence of multiple genetic instruments for each exposure and therefore the application of these approaches is restricted in the setting of evaluating methylation changes, where few independent mQTLs exist for individual CpG sites. Nonetheless, where there were multiple mQTLs available to instrument a CpG site, we assessed heterogeneity of the causal estimates which can be used to indicate pleiotropy, as well as violation of the other MR assumptions (36).

Supplementary analyses

Subgroup analyses

In addition to lung cancer (overall), the association between genetically increased methylation and lung cancer subtypes: lung adenocarcinoma (11,245 cases, 54,619 controls), small cell lung cancer (2,791 cases, 20,580 controls), and squamous cell lung cancer (7,704 cases, 54,763 controls) was assessed. We also assessed this association of lung cancer in never smokers (2,303 cases, 6,995 controls) and ever smokers (23,848 cases, 16,605 controls). GWAS summary statistics for these traits was obtained from the TRICL-ILCCO consortium (23).

Assessing the potential causal effect of AHRR methylation

We first used a one-sample MR approach to establish whether methylation at *AHRR* has a causal effect on lung cancer incidence in 8,758 participants in the Copenhagen City Heart Study (CCHS) (357 incident lung cancer cases, 8,401 remaining free of lung cancer and methylation data available at the cg05575921 *AHRR*). This prospective study (55), representing the Danish general population, examined individuals during 1991-1994. Details of the phenotypic, methylation and genetic data, as well as the linked lung cancer data, are outlined in the **Supplementary Methods**.

As just two mQTLs were identified in relation to cg05575921 methylation in ARIES at $P < 1x10^{-7}$ in the main analysis, we assessed whether the null result was due to lack of power. We sought to identify further genetic instruments located within 1 Mb of the index *AHRR* CpG site in ARIES and associated with cg05575921 methylation with a less stringent p-value based on the total number of SNPs in this region ($P < 0.05/4414 = 1.1x10^{-5}$) (**Supplementary Table 19**). 132 mQTLs surpassed this threshold and were taken forward for re-analysis in linear regression of methylation on each genotyped SNP available (13 of 132) in the CCHS. For those genetic variants which replicated within the CCHS (p-value below Bonferroni threshold for replication [$P < 0.05/13 = 0.004$]) (**Supplementary Table 20**), we performed an LD pruning step using the 1000 Genomes reference set (limited to those of European ancestry) and a less stringent r^2 threshold of 0.2.

Taking those SNPs, an unweighted allele score was created to act as an instrumental variable for *AHRR* methylation in Mendelian randomization analysis. This was calculated by coding and then summing the alleles to reflect the average number of methylation-increasing alleles carried by an individual. For investigating associations between the allele score and methylation, continuous effects were estimated using linear regression. We examined associations between a number of confounding factors (sex, alcohol consumption, smoking status, current and cumulative consumption of tobacco, occupational exposure to dust and/or welding fumes, passive smoking) previously considered in the association between methylation and lung cancer risk to check the core instrumental variable assumption that the instrument (genotype) is independent of factors that potentially confound the observational association.

Next, we performed a two-stage regression analysis: the first stage was a linear regression of the allele score as a continuous instrument on methylation levels and the second stage was a Cox regression of the predicted values of methylation on lung cancer incidence. This model was adjusted for sex and age

of the participants. Individuals (n=8758) free of lung cancer at time of blood sampling (1991-94) were followed until Dec. 31st 2012, death or event of lung cancer, whichever came first. Data on smoking status of the CCHS participants was also available and so we also performed MR analysis stratified by never, former and current smoking status at the time of blood draw.

Given that we only had data for 357 lung cancer events in the CCHS, we also used a two-sample MR approach to harness the power of larger sample sizes and obtain more precise causal estimates for the effect of *AHRR* CpG sites on risk of lung cancer. In this approach, the SNP-methylation associations for the replicated mQTLs, corresponding to the *AHRR* region, were established in 8,780 individuals from the CCHS and the SNP-lung cancer associations were taken from TRICL-ILCCO. Given the larger number of mQTLs in this analysis, we also investigated possible pleiotropy using the MR Egger regression method and appraised causal estimates using a weighted generalized regression method to account for the correlation structure between the SNPs (given the less stringent LD pruning threshold).

Tumour and adjacent normal methylation patterns

To further assess the effect of methylation on lung cancer, we obtained DNA methylation data from lung cancer tissue and matched normal adjacent tissue (n=40 lung squamous cell carcinoma and n=29 lung adenocarcinoma), profiled as part of The Cancer Genome Atlas and assessed methylation in the normal vs lung cancer samples, as outlined previously (27). This study design was used in addition to Mendelian randomization in order to triangulate findings regarding the potential causal effect of methylation.

mQTL association with gene expression

For the genes annotated to CpG sites identified in the lung cancer EWAS, we examined gene expression in both whole blood and lung tissue using data from the gene-tissue expression (GTEx) consortium (56).

All analyses were conducted in Stata (version 14) and R (version 3.2.2). For the two-sample Mendelian randomization analysis we used the MR-Base R package TwoSampleMR (57). All P values were two-sided. To maximize true positive ascertainment a false discovery rate (FDR) < 0.05 calculated using the Benjamini-Hochberg method was used for the analyses. This has been indicated in the text.

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. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet. 2006;7(1):21-33.
3. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
4. 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.
5. 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.
6. Zhang Y, Breitling LP, Balavarca Y, Holleczek B, Schottker B, Brenner H. Comparison and combination of blood DNA methylation at smoking-associated genes and at lung cancer-related genes in prediction of lung cancer mortality. Int J Cancer. 2016;139(11):2482-92.

7. Strathdee G, Brown R. Aberrant DNA methylation in cancer: potential clinical interventions. Expert Rev Mol Med. 2002;4(4):1-17.
8. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002;3(6):415-28.
9. 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.
10. 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.
11. 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.
12. Bojesen SE, Timpson N, Relton C, Davey Smith G, Nordestgaard BG. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. Thorax. 2017.
13. 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.
14. 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.
15. 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.
16. 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.
17. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23(R1):R89-98.
18. 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.
19. 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.
20. Relton CL, Davey Smith G. Mendelian randomization: applications and limitations in epigenetic studies. Epigenomics. 2015;7(8):1239-43.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. Monick MM, Beach SR, Plume J, Sears R, Gerrard M, Brody GH, et al. Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. Am J Med Genet B Neuropsychiatr Genet. 2012;159B(2):141-51.

27. 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.
28. Stueve TR, Li WQ, Shi J, Marconett CN, Zhang T, Yang C, et al. Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. Hum Mol Genet. 2017;26(15):3014-27.
29. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet. 2013;22(5):843-51.
30. Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, Palli D, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Human Molecular Genetics. 2015;24(8):2349-59.
31. Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet. 2012;21(13):3073-82.
32. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9(10):1382-96.
33. 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.
34. 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.
35. Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, et al. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nat Commun. 2016;7:10478.
36. 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.
37. Burgess S, Thompson SG, Collaboration CCG. Avoiding bias from weak instruments in Mendelian randomization studies. International Journal of Epidemiology. 2011;40(3):755-64.
38. Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. International Journal of Epidemiology. 2004;33(1):30-42.
39. Birney E, Smith GD, Greally JM. Epigenome-wide Association Studies and the Interpretation of Disease -Omics. PLoS Genet. 2016;12(6):e1006105.
40. 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.
41. 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.
42. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. Am J Hum Genet. 2016;98(4):680-96.
43. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD, et al. sva: Surrogate Variable Analysis. R pacakge version 3260. 2017.
44. 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.
45. 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.
46. Relton CL, Gaunt T, McArdle W, Ho K, Duggirala A, Shihab H, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). Int J Epidemiol. 2015;44(4):1181-90.

47. 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.
48. 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.
49. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature Genetics. 2016;48(10):1284-7.
50. Inoue A, Solon G. Two-Sample Instrumental Variables Estimators. Rev Econ Stat. 2010;92(3):557-61.
51. 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.
52. 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.
53. Gao X, Thomsen H, Zhang Y, Breitling LP, Brenner H. The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes. Clin Epigenetics. 2017;9:87.
54. Tobacco, Genetics C. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010;42(5):441-7.
55. Kaur-Knudsen D, Bojesen SE, Tybjaerg-Hansen A, Nordestgaard BG. Nicotinic acetylcholine receptor polymorphism, smoking behavior, and tobacco-related cancer and lung and cardiovascular diseases: a cohort study. J Clin Oncol. 2011;29(21):2875-82.
56. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.
57. 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. 2016.

				Basic meta-analysis		SV adjusted meta-analysis		count + SV adjusted meta-anal		Never smokers meta-analysis		Former smokers meta-analysis		Current smokers meta-analysis	
CpG	**Gene region**	**Chr**	**Position**	**OR (95% CI)**	**P value**	**OR (95% CI)**	**P value**	**OR (95% CI)**	**P value**	**OR (95% CI)**	**P value**	**OR (95% CI)**	**P value**	**OR (95% CI)**	**P value**
cg05575921	AHRR	5	373378	0.474 (0.397, 0.566)	1.45E-16	0.452 (0.367, 0.556)	6.27E-14	0.452 (0.365, 0.560)	3.60E-13	0.932 (0.638, 1.36)	7.17E-01	0.458 (0.337, 0.622)	6.10E-07	0.708 (0.599, 0.837)	5.36E-05
cg21566642	ALPPL2	2	233284661	0.535 (0.459, 0.624)	1.70E-15	0.525 (0.442, 0.624)	2.49E-13	0.513 (0.429, 0.614)	3.12E-13	0.892 (0.677, 1.18)	4.18E-01	0.522 (0.401, 0.680)	1.42E-06	0.746 (0.635, 0.876)	3.67E-04
cg06126421	IER3	6	30720080	0.585 (0.507, 0.675)	2.08E-13	0.544 (0.455, 0.651)	2.49E-11	0.513 (0.425, 0.620)	3.92E-12	0.783 (0.529, 1.16)	2.22E-01	0.561 (0.431, 0.731)	1.88E-05	0.727 (0.558, 0.947)	1.79E-02
cg03636183	F2RL3	19	17000585	0.636 (0.559, 0.724)	7.99E-12	0.615 (0.526, 0.718)	8.21E-10	0.610 (0.520, 0.717)	1.61E-09	0.909 (0.663, 1.25)	5.53E-01	0.624 (0.494, 0.788)	7.50E-05	0.786 (0.671, 0.921)	2.92E-03
cg05951221	ALPPL2	2	233284402	0.660 (0.582, 0.749)	9.68E-11	0.642 (0.555, 0.741)	1.77E-09	0.629 (0.541, 0.731)	1.50E-09	0.868 (0.621, 1.21)	4.09E-01	0.634 (0.506, 0.794)	7.21E-05	0.819 (0.708, 0.948)	7.42E-03
cg01940273	ALPPL2	2	233284934	0.692 (0.606, 0.789)	4.20E-08	0.675 (0.577, 0.788)	7.32E-07	0.685 (0.583, 0.804)	3.58E-06	1.14 (0.821, 1.59)	4.28E-01	0.575 (0.445, 0.744)	2.57E-05	0.876 (0.761, 1.01)	6.59E-02
cg23771366	PRSS23	11	86510998	0.769 (0.697, 0.847)	1.10E-07	0.729 (0.641, 0.829)	1.45E-06	0.709 (0.620, 0.811)	5.60E-07	1.09 (0.849, 1.41)	4.90E-01	0.621 (0.501, 0.770)	1.40E-05	0.856 (0.751, 0.975)	1.97E-02
cg11660018	PRSS23	11	86510915	0.788 (0.721, 0.861)	1.18E-07	0.700 (0.612, 0.801)	1.97E-07	0.678 (0.588, 0.782)	8.86E-08	0.935 (0.734, 1.19)	5.86E-01	0.753 (0.636, 0.892)	1.01E-03	0.844 (0.752, 0.948)	4.15E-03
cg26963277	KCNQ1	11	2722407	0.668 (0.576, 0.776)	1.21E-07	0.640 (0.530, 0.773)	3.79E-06	0.623 (0.511, 0.758)	2.53E-06	0.539 (0.330, 0.883)	1.40E-02	0.724 (0.557, 0.940)	1.54E-02	0.707 (0.570, 0.877)	1.59E-03
cg27241845	ALPPL2	2	233250370	0.669 (0.576, 0.777)	1.45E-07	0.679 (0.570, 0.810)	1.67E-05	0.673 (0.560, 0.809)	2.47E-05	0.750 (0.486, 1.16)	1.93E-01	0.677 (0.516, 0.889)	5.01E-03	0.726 (0.588, 0.898)	3.09E-03
cg23387569	AGAP2	12	58120011	0.713 (0.628, 0.809)	1.53E-07	0.702 (0.604, 0.816)	3.69E-06	0.683 (0.584, 0.799)	1.89E-06	0.786 (0.557, 1.11)	1.69E-01	0.714 (0.552, 0.923)	1.02E-02	0.749 (0.621, 0.903)	2.48E-03
cg09935388	GFI1	1	92947588	0.676 (0.583, 0.785)	2.48E-07	0.669 (0.560, 0.800)	9.67E-06	0.674 (0.561, 0.812)	3.00E-05	0.961 (0.643, 1.44)	8.44E-01	0.740 (0.553, 0.990)	4.22E-02	0.681 (0.560, 0.827)	1.06E-04
cg01901332	ARRB1	11	75031054	0.725 (0.642, 0.820)	2.82E-07	0.686 (0.580, 0.812)	1.12E-05	0.658 (0.553, 0.782)	2.20E-06	1.02 (0.720, 1.44)	9.22E-01	0.599 (0.460, 0.781)	1.48E-04	0.783 (0.663, 0.924)	3.92E-03
cg25305703	CASC21	8	128378218	0.725 (0.639, 0.821)	4.46E-07	0.717 (0.606, 0.849)	1.11E-04	0.715 (0.601, 0.850)	1.48E-04	0.801 (0.567, 1.13)	2.10E-01	0.761 (0.598, 0.968)	2.58E-02	0.769 (0.645, 0.916)	3.20E-03
cg16823042	AGAP2	12	58119992	0.739 (0.654, 0.835)	1.14E-06	0.726 (0.628, 0.839)	1.51E-05	0.701 (0.601, 0.817)	5.90E-06	0.830 (0.580, 1.19)	3.09E-01	0.720 (0.566, 0.915)	7.36E-03	0.799 (0.668, 0.955)	1.35E-02
cg08709672	AVPR1B	1	206224334	0.749 (0.666, 0.842)	1.36E-06	0.759 (0.660, 0.873)	1.14E-04	0.739 (0.638, 0.856)	5.33E-05	0.729 (0.499, 1.06)	1.02E-01	0.738 (0.602, 0.905)	3.47E-03	0.816 (0.687, 0.970)	2.13E-02

Table 1. Meta-analysis of EWAS of lung cancer using 4 separate cohorts: 16 CpG sites associated with lung cancer at FDR < 0.05. OR = odds ratio per standard deviation increase in DNA methylation, SV = surrogate variable.

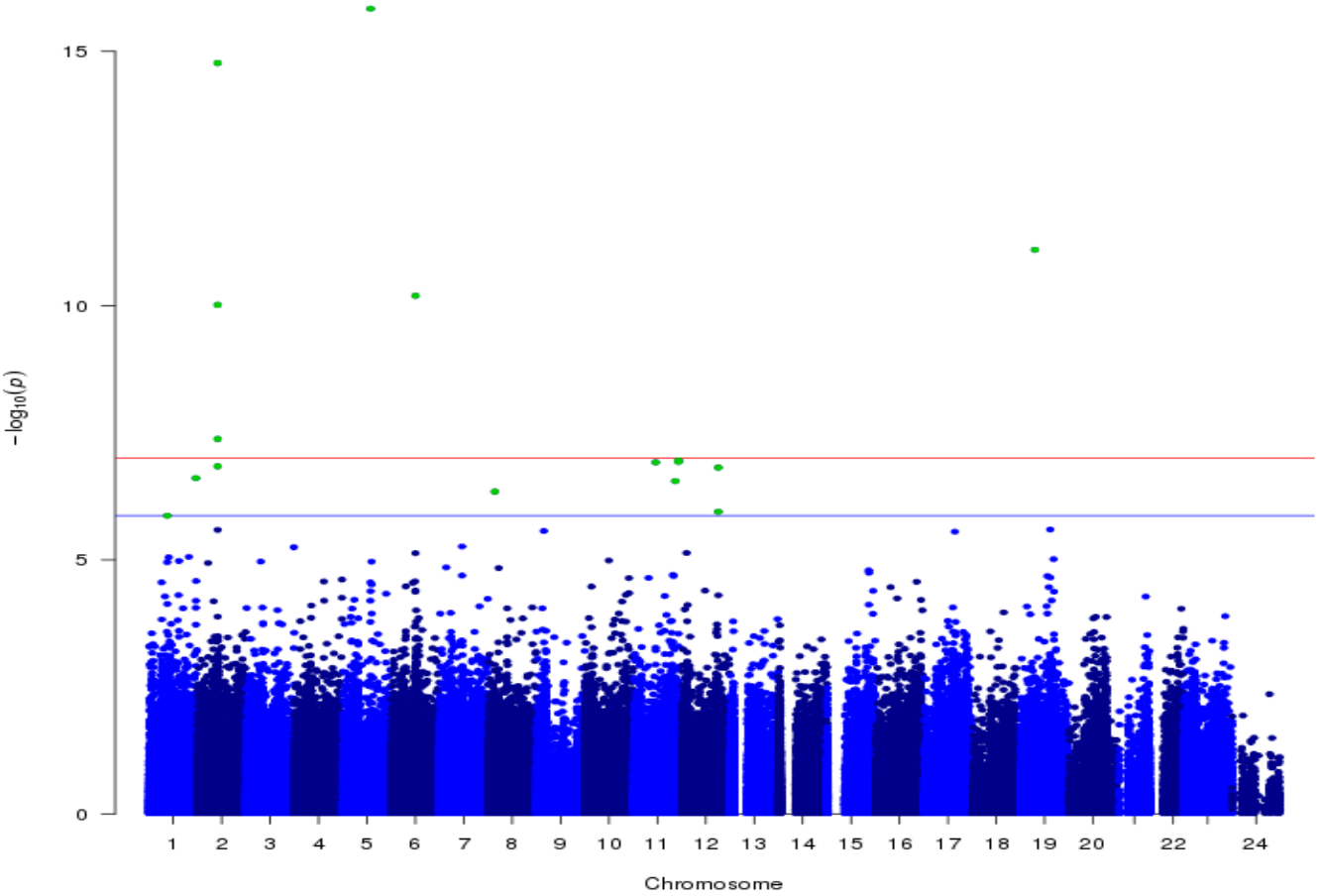

Figure 1. Observational associations of DNA methylation and lung cancer: A fixed effects meta-analysis of lung cancer EWAS weighted on the inverse variance was performed to establish the observational association between differential DNA methylation and lung cancer. All points above the red line are at $P < 1x10^{-7}$ and all points above the blue line (and those in green) are at FDR < 0.05. In total 16 CpG sites are associated with lung cancer (FDR < 0.05).

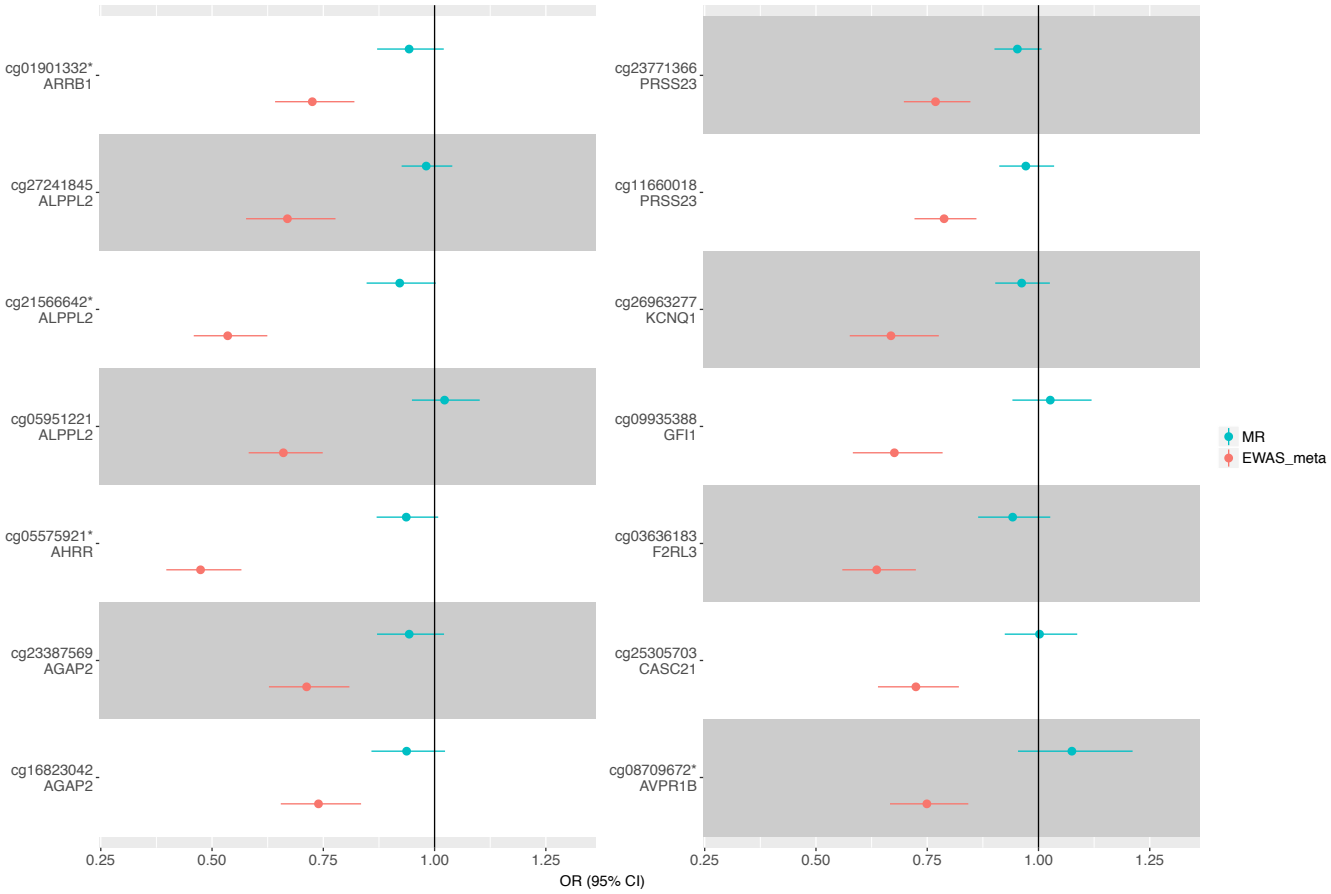

Figure 2. Mendelian randomisation (MR) vs. observational analysis. Two-sample MR was carried out with methylation at 14/16 CpG sites identified in the EWAS meta-analysis as the exposure and lung cancer as the outcome. cg01901332 and cg05575921 had 2 instruments are so the estimate was calculated using 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). OR = odds ratio per SD increase in DNA methylation.

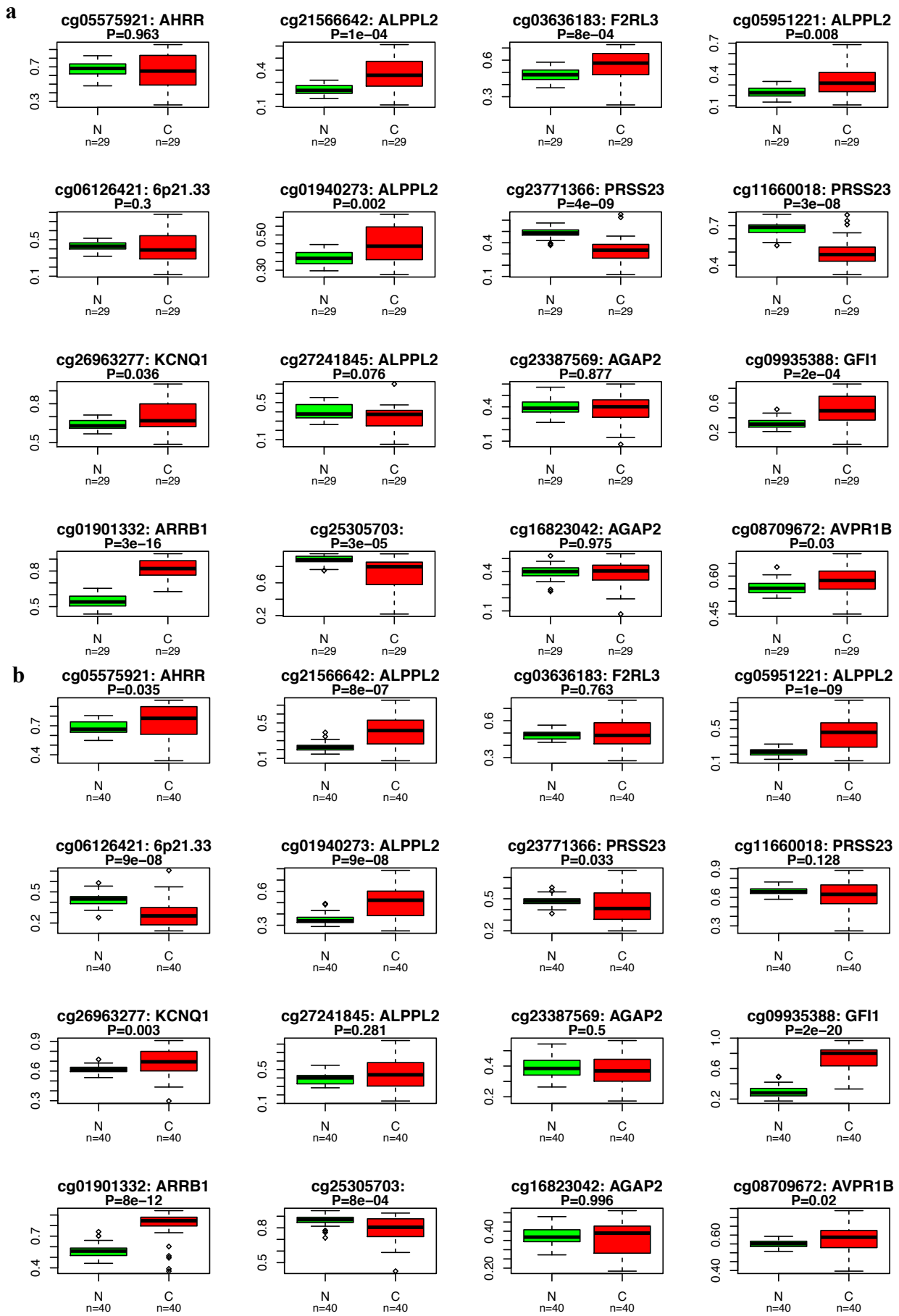

Figure 3. Differential DNA methylation in lung cancer tissue: A comparison of methylation at each of the 16 CpG sites identified in our meta-analysis was made between lung cancer tissue and adjacent healthy lung tissue for patients with a) lung adenocarcinoma and b) squamous cell lung cancer. Data from The Cancer Genome Atlas was used for this analysis.