The association of cigarette smoking with DNA methylation and gene expression in human tissue samples

Authors

James L. Li, Niyati Jain, Lizeth I. Tamayo, ..., Meritxell Oliva, Lin S. Chen, Brandon L. Pierce

Correspondence

brandonpierce@uchicago.edu

Li et al., 2024, The American Journal of Human Genetics *111*, 636–653
April 4, 2024
https://doi.org/10.1016/j.ajhg.2024.02.012

CellPress

ARTICLE

The association of cigarette smoking with DNA methylation and gene expression in human tissue samples

James L. Li,[1,2,10] Niyati Jain,[1,3,10] Lizeth I. Tamayo,[1] Lin Tong,[1] Farzana Jasmine,[4] Muhammad G. Kibriya,[1] Kathryn Demanelis,[5,6] Meritxell Oliva,[1,7] Lin S. Chen,[1] and Brandon L. Pierce[1,8,9,*]

Summary

Cigarette smoking adversely affects many aspects of human health, and epigenetic responses to smoking may reflect mechanisms that mediate or defend against these effects. Prior studies of smoking and DNA methylation (DNAm), typically measured in leukocytes, have identified numerous smoking-associated regions (e.g., *AHRR*). To identify smoking-associated DNAm features in typically inaccessible tissues, we generated array-based DNAm data for 916 tissue samples from the GTEx (Genotype-Tissue Expression) project representing 9 tissue types (lung, colon, ovary, prostate, blood, breast, testis, kidney, and muscle). We identified 6,350 smoking-associated CpGs in lung tissue (n = 212) and 2,735 in colon tissue (n = 210), most not reported previously. For all 7 other tissue types (sample sizes 38–153), no clear associations were observed (false discovery rate 0.05), but some tissues showed enrichment for smoking-associated CpGs reported previously. For 1,646 loci (in lung) and 22 (in colon), smoking was associated with both DNAm and local gene expression. For loci detected in both lung and colon (e.g., *AHRR*, *CYP1B1*, *CYP1A1*), top CpGs often differed between tissues, but similar clusters of hyper- or hypomethylated CpGs were observed, with hypomethylation at regulatory elements corresponding to increased expression. For lung tissue, 17 hallmark gene sets were enriched for smoking-associated CpGs, including xenobiotic- and cancer-related gene sets. At least four smoking-associated regions in lung were impacted by lung methylation quantitative trait loci (QTLs) that co-localize with genome-wide association study (GWAS) signals for lung function (FEV1/FVC), suggesting epigenetic alterations can mediate the effects of smoking on lung health. Our multi-tissue approach has identified smoking-associated regions in disease-relevant tissues, including effects that are shared across tissue types.

Introduction

Cigarette smoking has many detrimental effects on human health, including increased risk for cancer, cardiovascular diseases, and respiratory diseases.[1] Tobacco smoke contains thousands of chemicals, dozens of which are known carcinogens, and the potential mechanisms underlying the adverse effects of these chemicals on health include DNA damage, inflammation, and oxidative stress.[2] The effects of smoking on specific features of the human epigenome have been described previously, including studies that identify associations between smoking behaviors and epigenetic features, such as DNA methylation (DNAm) at cytosine-guanine (CpG) dinucleotides.[3,4]

The association between cigarette smoking and DNAm has been characterized in prior epigenome-wide association studies (EWASs).[5–22] Gene regions where DNAm in leukocytes shows consistent association with smoking status include the aryl-hydrocarbon receptor repressor (*AHRR*),[7,9,19–22] coagulation factor II (thrombin) receptor-like 3 (*F2RL3*),[7–9,19–21] G protein-coupled receptor 15 (*GPR15*),[7,19,21] 2q37.1 (containing *ALPPL2*),[7,9,19] and 6p21.33 (major histocompatibility complex) regions.[7,9,19,20] These studies demonstrate that DNAm changes in blood can be used as biomarkers of tobacco exposure and smoking history,[6] and subsequent studies have reported associations between smoking-related DNAm features and risk for smoking-related diseases, such as lung cancer.[23,24]

While the majority of prior studies on this topic focus on the effects of smoking on DNAm in leukocytes, there have been a small number of studies focusing on other tissue types, including lung,[6,15] cord blood,[25] placenta,[17] and blood from newborns with prenatal exposure.[11,16] These studies highlight regions affected by tobacco exposure in multiple tissue types (e.g., *AHRR*),[10,16,18,22,25,26] as well as effects that are potentially tissue-specific (e.g., cg27402634 near *LEKR1* and long noncoding RNA *LINC00886*, a hallmark of maternal smoking in placenta).[17] However, the association of cigarette smoking with DNAm in non-blood tissue types has received relatively little attention, as most tissue types are typically inaccessible in human studies.

[1]Department of Public Health Sciences, University of Chicago, Chicago, IL 60637, USA; [2]Interdisciplinary Scientist Training Program, University of Chicago, Chicago, IL 60637, USA; [3]Committee on Genetics, Genomics, Systems Biology, University of Chicago, Chicago, IL 60637, USA; [4]Institute for Population and Precision Health (IPPH), Biological Sciences Division, University of Chicago, Chicago, IL 60637, USA; [5]Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15261, USA; [6]UPMC Hillman Cancer Center, Pittsburgh, PA 15232, USA; [7]Genomics Research Center, AbbVie, North Chicago, IL 60064, USA; [8]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA; [9]Comprehensive Cancer Center, University of Chicago, Chicago, IL 60637, USA
[10]These authors contributed equally
*Correspondence: brandonpierce@uchicago.edu
https://doi.org/10.1016/j.ajhg.2024.02.012.

In this study, we generate genome-wide array-based DNAm data using human tissue samples from the Genotype-Tissue Expression (GTEx) project to assess the association of smoking and DNAm in the lung, colon, and seven additional tissue types.

Material and methods

Sample collection and ethics approval

As detailed in Oliva et al.,[27] the GTEx research protocol was reviewed by Chesapeake Bay Review, Roswell Park Comprehensive Cancer Center's Office of Research Subject Protection, and the institutional review boards at the University of Pennsylvania. Analyses of DNA samples from GTEx participants at the University of Chicago was not considered human subjects research by the university's institutional review board since only deidentified data on deceased individuals were utilized in this study.

The GTEx project

The GTEx project has established a biobank of human tissue samples from >950 postmortem multi-tissue donors for the study of molecular phenotypes.[28] The GTEx v8 dataset consists of tissue-specific RNA-sequencing and genotyping data from 838 donors and 17,382 unique samples from 52 tissue types.[29,30] GTEx also provides information on sex, age, and race/ethnicity based on questionnaire data, as well as measurements of ischemic time for all samples. For this project, we obtained DNAm measurements for 916 GTEx tissue samples representing nine tissue types (lung, colon, ovary, prostate, whole blood, breast, testis, kidney, and muscle), described previously.[27] These nine tissue types were selected based on several criteria reflecting our research interests, including inclusion of cancer-relevant tissues (lung, colon, prostate, ovary, breast, and kidney), tissues with unique aging biology (testis and skeletal muscle), and tissues commonly used in epidemiological research (whole blood). With resources available to profile DNAm for ~1,000 samples, we selected larger numbers of samples for some tissue types of strong public health interest (e.g., lung, colon, and ovary) and to assess the impact of sample size on power for DNAm quantitative trait loci (mQTL) detection.[27]

Determination of smoking status for GTEx donors

Smoking status was assigned "ever cigarette smoker" for GTEx donors with a reported history of cigarette smoking and "never cigarette smoker" for donors with no reported history of cigarette smoking. Assignment was based on the MHSMKSTS variable (smoking status: yes, no, unknown) and the MHSMKTP variable (smoke type: cigarette, cigar, pipe, others) provided by GTEx. We were able to assign ever/never status to 396 donors with DNAm data (269 cigarette smokers and 127 non-cigarette smokers), with 21 donors (46 samples) lacking data on cigarette smoking status. Ever cigarette smokers include both current and former smokers; however, the distinction between these two smoking categories was not assessed in our primary analyses due to incomplete information of prior smoking and smoking duration for many GTEx donors. However, we constructed a "current smoker" variable, which was used for secondary analyses. This variable was constructed using free text comments from family members of the tissue donors, recorded in the MHSMKCMT variable provided by GTEx.

DNAm data and quality control

DNA was extracted from GTEx tissue samples via the Qiagen Gentra Puregene method at GTEx Laboratory Data, Analysis and Coordinating Center (LDACC). The LDACC shipped DNA from 1,000 unique tissue samples to the University of Chicago. These 1,000 samples represent 424 unique GTEx donors and 9 unique GTEx tissue types. Genome-wide DNAm at >850,000 CpG sites was assessed using the Infinium MethylationEPIC array (Illumina, San Diego, CA, USA). All DNA samples were prepared and analyzed in accordance with the manufacturer's guidelines and protocols. For sample quality control (QC), we excluded 3 samples with undetectable or missing methylation values (detection $p > 0.01$) in $\geq$5% of CpGs, 6 samples with mismatched sex, and 13 samples that did not show clear clustering with their corresponding tissue type. The EPIC array measures 59 high-frequency SNPs that can be used as a genetic fingerprint.[31] Using these SNPs, we identified one sample that did not match the donor's existing genotype data, and this sample was excluded. The 15 male samples from breast tissue were excluded from the DNA methylation data. The 46 samples lacking data on cigarette smoking status were also excluded from the analyses. In total, 84 samples were excluded. After excluding samples due to quality control or missing data issues, there were 916 samples used for analysis (representing 9 tissue types and 398 GTEx donors).

For CpG QC, we followed guidance from Pidsley et al.[32] We excluded CpGs measured by probes with potential non-specific binding (n = 43,254), CpGs overlapping genetic variants or with variants that overlap single-base extension sites for type 1 probes (n = 7,708), CpGs mapping to the X and Y chromosomes (n = 16,037), and poorly performing CpGs according to Illumina (n = 169). We also excluded CpGs that had detection $p > 0.01$ in at least one sample (n = 44,135). A total of 754,119 CpGs passed QC and were retained for analyses. Genomic positions for all CpGs (and for all SNP and gene expression analyses described below) are based on human reference genome build hg19.

GTEx gene expression data

Gene-level expression data (v8) derived from RNA sequencing was obtained from the GTEx portal. The expression value for each gene was estimated as reads per kilobase of transcript per million mapped reads (RPKM) using RNA-SeQC on uniquely mapped, properly paired reads fully contained within exon boundaries and with alignment distances $\leq$6.[33] Samples with <10 million mapped reads or with outlier expression measurements based on the *D* statistic were removed.[34] A total of 56,200 genes in the v8 dataset had expression levels recorded in both read counts and transcripts per million (TPM). Read counts from these genes were normalized across samples using the trimmed mean of M-values (TMM) normalization method in the "edgeR" package to generate TMM-normalized TPM for each gene.[35] Following TMM normalization, genes were selected based on the expression threshold of >0.1 TPM in at least 20% of samples and $\geq$6 reads in at least 20% of the samples. We then restricted to the fully processed, filtered, and normalized autosomal genes from the GTEx v8 dataset, which resulted in 25,272 genes expressed in lung (n = 541) and 24,580 genes expressed in colon (n = 382).

Association of cigarette smoking status with DNAm

The beta values for each CpG were logit transformed into M-values prior to analyses using the following formula: $\log_2$[-beta/(1 − beta)]. For each tissue type, the association between

smoking status (ever/never) and DNAm at each CpG site was estimated using a linear model implemented in the "limma" package[36] in R, with age, sex, body mass index (BMI), race/ethnicity, ischemic time, batch/plate, and surrogate variables (SVs) included as covariates. For analyses of lung tissue, we also adjusted for common lung-related health conditions, including asthma (n = 24), chronic obstructive pulmonary disease (COPD) (n = 34), and pneumonia (n = 24). The R "sva" package[37] was used to generate the SVs for each tissue type. We included the smoking variable in the full model matrix but omitted the smoking variable from the null model matrix to prevent the effects of smoking from being captured by SVs. The resulting SVs were used to control for technical variation and other biologic sources of variability (i.e., cell-type composition). As a general rule, we adjusted for 10 SVs for tissue types with $n > 100$ and 5 SVs for tissue types with $n < 100$. To ensure the SVs captured all variability due to cell-type composition (but not effects of smoking), we examined correlations of the first 20 SVs (per tissue type) with smoking status and cell-type composition estimates (derived using the EPISCORE method as described below). SVs showing clear association with EPISCORE cell-type estimates were typically among the top 5 SVs (Table S1). We considered the exclusion of SVs associated with smoking status. For example, for lung tissue, four of the top 10 SVs showed association with smoking status ($p < 0.05$); however, including these SVs as covariates resulted in only mild attenuation of associations observed, so all 10 SVs were retained. For colon DNAm data, no SVs were associated with smoking status, so all 10 SVs were retained. The false discovery rate (FDR) was estimated using the Benjamini-Hochberg method.[38]

Estimating power to detect smoking-related CpGs at varying sample sizes

We estimated the power to detect the effect sizes observed for CpGs identified in lung tissue (FDR 0.05) at sample sizes similar to other tissues by first generating 1,000 random subsamples of our lung tissue samples at sample sizes of 50, 100, and 150. We then performed an EWAS (described above) in each of these subsamples and determined the proportion of subsamples where we identified the smoking-associated CpG with the largest effect size magnitude (cg01584760), median effect size magnitude (cg20291548), and the smallest effect size magnitude (cg09138315) observed in lung.

Enrichment of smoking-associated CpGs in each tissue with previously reported CpGs

To assess whether smoking-associated CpGs in each of our tissues were significantly enriched for previously reported CpGs in whole blood,[39] adipose,[13] placenta,[17] or CpGs identified in analyses of GTEx lung tissue, we calculated the proportion of all smoking-associated CpGs ($p < 10^{-5}$) in each tissue that had been reported previously for each of these tissue types (and based on GTEx lung results). We performed a one-sided, two-sample z test of proportions to determine if CpGs previously reported were higher than the proportion of smoking-associated CpGs detected among all CpGs analyzed. We repeated these analyses using a $p < 10^{-3}$ threshold for classifying smoking-associated CpGs.

Association of smoking status with gene expression

The association between smoking status and expression in lung and colon for each gene was estimated using a linear model implemented in "limma," adjusting for age, sex, BMI, race/ethnicity, ischemic time, and 5 SVs (created using expression data). For lung tissue, we also adjusted for three lung-related diseases: asthma, COPD, and pneumonia (as described above).

Enrichment and pathway analyses for smoking-related CpGs

We compared the proportion of smoking-associated CpG sites (FDR 0.05) assigned to island, shore, shelf, and open sea (Illumina annotations) in lung and colon tissue to the distribution in the entire Infinium MethylationEPIC array (754,199 CpGs) using chi-square tests. We assessed enrichment of smoking-associated CpGs in chromatin segmentation features using reference data from the Roadmap Epigenomics project database[40] (primary tissue colonic mucosa and lung). We used the R package *"oddsratio"* to calculate enrichment and Fisher's exact test to compute p values. The above enrichment analyses were performed stratified by hypermethylated vs. hypomethylated smoking-associated CpG sites. Additionally, we performed a motif enrichment analysis to identify enrichment of smoking-associated CpGs in transcription factor binding sites (TFBS). We used annotations from the ENCODE version 2 and 3 chromatin immunoprecipitation sequencing experiments (1,256 experiments), representing 340 transcription factors (TFs) in 129 cell and tissue types. These annotations were obtained from the University of California, Santa Cruz (UCSC) table browser (encRegTfbsClustered, build hg38). We assessed enrichment via hypogeometric tests, using the *phyper* function in R. CpGs were additionally assigned to genes (based on annotations provided by Illumina), and genes were assigned to pathways and biologic processes using the hallmark gene set collections (n = 50 sets),[41] as well as Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations[42–44] We conducted gene set enrichment analysis (GSEA) using the *gometh* function in "missMethyl"[45] for lung and colon tissues using smoking-associated CpGs (FDR 0.05). This function accounts for the potential bias in GSEA due to the number of CpGs per gene by computing prior probabilities[46] and evaluates enrichment using a hypergeometric test. Enriched gene sets were defined as those passing an FDR of 0.05. The motif enrichment analysis and pathway analysis were performed for all smoking-associated CpGs as well as stratified by hypomethylated vs. hypermethylated smoking-associated CpGs.

Identification of mQTLs for smoking-associated CpGs that co-localize with GWAS signals for tissue-relevant diseases

For the 2,478 smoking-associated CpGs observed in lung tissue (FDR < 0.01) and the 662 CpGs observed in colon tissue, we identified the CpGs previously shown to be affected by an mQTL in GTEx lung or colon tissue.[27] For the 566 CpGs and 68 CpGs identified in lung and colon, respectively, we extracted the identifiers for the lead SNP for each corresponding mQTL (550 and 68 lead SNPs, respectively). We then searched for these lead SNPs of lung mQTLs in the genome-wide summary statistics from several large genome-wide association studies (GWASs) of lung-related diseases and phenotypes,[27,47–49] including lung cancer, asthma, chronic obstructive pulmonary disease, as well as two spirometry-based phenotypes that have been clinically used to assess lung health: FVC (forced vital capacity) and FEV1/FVC (forced expiratory volume in the first 1 s divided by the forced vital capacity) (obtained from MR-Base).[48] We

additionally searched for these lead SNPs of colon mQTLs in several large GWASs of colorectal cancer and inflammatory bowel disease.[50,51] We retained all mQTL lead SNPs showing association ($p < 5 \times 10^{-8}$) in these GWASs (corresponding to 10 mQTLs in lung). For each mQTL retained, we used mQTL results for SNPs within 250 kb of the lead mQTL SNP to test for colocalization with the corresponding GWAS association signal, using GWAS summary statistics for the same set of SNPs via the "coloc" package in R[52] with the following default prior probabilities: the prior probability of a SNP is associated with either mQTL or GWAS = 1.0×10^{-4}; the prior probability of a SNP is associated with both mQTL and GWAS = 1.0×10^{-5}. We visualized evidence of colocalization using the LocusCompare software.[53]

Analyses of cell-type proportions

For lung and colon, we estimated cell-type composition using the methylation-based EPISCORE method ("wRPC" function) using a pan-tissue DNAm atlas[54] as a reference dataset. We chose this method instead of expression-based deconvolution methods because (1) it provides cell-type estimates for all samples with DNAm data (some of which lack RNA-sequencing data) and (2) it provides cell-type estimates based on the same tissue samples used for DNAm measurement (as RNA was extracted from a different piece of tissue). The cell types estimated by EPISCORE and their distributions are shown in Figure S1. We observe inter-individual variability in cell-type proportions and correlation with DNAm-derived SVs.

For benchmarking purposes, we compared EPISCORE estimates of epithelial cell proportions to epithelial cell enrichment scores previously computed for GTEx samples using xCell, an expression-based method (Figure S2).[55] EPISCORE epithelial cell estimates and xCell epithelial scores were not correlated in lung (Spearman rho = −0.093; p = 0.29) but showed a clear correlation in colon (rho = 0.55; $p = 4.1 \times 10^{-7}$). Thus, we have more confidence that our cell-type estimates accurately represent cell-type abundances in colon samples (as compared to lung). However, we utilize EPISCORE estimates for both tissue types for smoking-by-cell-type interaction analyses. To identify effects of smoking that vary by cell types in lung tissue, we identified CpGs involved in smoking and cell-type interactions (SxCT) by performing an interaction test between smoking and a given cell type for all CpGs and for each inferred cell type. For these SxCT analyses, we transformed cell-type proportion estimates obtained from EPISCORE to standard normal distributions. We performed a linear regression testing the association of DNAm with the smoking and cell-type proportion interaction term, adjusting for age, sex, BMI, race/ethnicity, and ischemic time. For analyses of lung tissue, we also adjusted for common lung-related health conditions mentioned above.

Results

Characteristics of GTEx tissue donors

We generated DNAm data for 917 unique tissue samples, obtained from 396 unique GTEx donors, representing 9 different GTEx tissue types (Table 1). Sample sizes analyzed for each tissue type ranged from 38 (breast) to 212 (lung). The number of tissues analyzed per donor ranged from 1 to 7 (average of 2.3). Among tissue types that are not sex specific, ~70% of samples came from male donors. 85% of the 398 donors were reported to be white. Among all donors included, 70% were classified as smokers, and there were no clear differences in this percentage across age groups.

Identification of smoking-associated CpG sites across different tissue types

Analyses of smoking status in relation to genome-wide DNAm in lung (n = 212) identified 6,350 smoking-associated CpG sites at an FDR of 0.05 ($p < 0.0004$) (Table S2), including many reported in previous EWASs. However, the majority of the CpGs identified (6,209 CpGs) have not been reported in previous studies of DNAm in blood cells based on a recent review of smoking-related changes in DNAm and gene expression.[39] Analyses of smoking and DNAm in colon (n = 210) resulted in 2,735 smoking-associated CpGs at an FDR of 0.05 ($p < 0.0001$). Smoking-associated CpGs identified in lung had effect sizes with magnitudes ranging from 0.073 to 1.459, with a mean effect size of 0.286; those identified in colon had effect sizes with magnitudes ranging from 0.072 to 1.075, with a mean effect size of 0.334. For all 7 other tissue types (with sample sizes ranging from 38 to 153), no clear associations with smoking status were observed (FDR of 0.05, Table S2 and Data S1–S9).

Secondary analyses performed in lung and colon tissues using current vs. never smoking as the exposure variable produced a larger number of CpGs passing the FDR threshold (Table S2b), but the top CpGs were the same, and p values were similar (Data S10 and S11). Additionally, we observed strong correlation between association estimates from our primary EWAS in lung tissue (ever vs. never) and our secondary EWAS (current vs. never) (R = 0.87), with a slight bias toward stronger association in the secondary analysis (Figure S3); this correlation was even stronger for smoking-associated CpGs reaching an FDR of 0.05 (R = 0.99). However, our remaining analyses focus on our primary exposure variable (ever vs. never smoking).

The abundance of smoking-associated CpG sites observed for lung was clearly larger than those observed for colon (a tissue type with similar sample size). To assess the extent to which lung tissue showed more prominent effects of smoking than other tissue types, we randomly selected subsets of lung samples to produce sample sizes similar to those of the other tissue types studied (e.g., n = 111 for prostate). After this down-sampling, the number of smoking-related CpGs detectable in lung tissue was larger than the number detected in ovary (n = 153) in 1,000 of 1,000 subsamples and the number detected in prostate (n = 111) in 997 of 1,000 subsamples (Figure S4). However, for tissue types with sample sizes of ~50, we had limited power to clearly demonstrate that lung had a larger number of smoking-associated CpGs (Figure S4). To assess power to detect the effect sizes observed in lung at lower sample sizes, we performed an EWAS in 1,000 random subsamples of lung tissue samples

Table 1. Summary of GTEx tissue samples used for DNA methylation analyses

Tissue types	Lung (n = 212)	Colon (n = 209)	Ovary (n = 153)	Prostate (n = 111)	Whole blood (n = 52)	Breast (n = 38)	Testis (n = 48)	Kidney (n = 47)	Muscle (n = 46)
Age (years)	55.1 (11.1)	55.7 (11.4)	50.7 (13.5)	54 (12.4)	49.7 (12.7)	50.0 (11.9)	53.7 (12.2)	59.3 (8.3)	56.9 (10.5)
BMI (kg/m^2)	27.5 (3.9)	27 (3.9)	26.8 (4.2)	27 (3.8)	27.3 (4.2)	25.4 (3.94)	27.1 (3.8)	26.2 (3.8)	26.7 (4.4)
Sex									
Male	150 (70.8)	143 (68.4)	0	111 (100)	43 (82.7)	0	48 (100)	36 (76.6)	27 (58.7)
Female	62 (29.3)	66 (31.6)	153 (100)	0	9 (17.3)	38 (100)	0	11 (23.4)	19 (41.3)
Race									
White	180 (84.9)	183 (87.6)	124 (81.1)	101 (91)	46 (88.5)	32 (84.2)	45 (93.8)	42 (89.4)	40 (87)
Af. Americans	27 (12.7)	21 (10)	26 (17)	8 (7.2)	5 (9.6)	6 (15.8)	3 (6.3)	5 (10.6)	6 (13)
Others	5 (2.4)	5 (2.4)	3 (2)	2 (1.8)	1 (1.9)	0	0 (0)	0	0
Cigarette smoking									
Ever	150 (70.8)	152 (72.7)	96 (62.8)	73 (65.8)	40 (76.9)	27 (71.1)	36 (75)	37 (78.7)	34 (73.9)
Current	89 (42)	91 (43.5)	66 (43.1)	42 (37.8)	25 (48.1)	19 (50)	16 (33.3)	19 (40.4)	15 (32.6)
Former	61 (28.7)	61 (29.2)	30 (19.6)	31 (27.9)	15 (28.8)	8 (21.1)	20 (41.7)	18 (38.3)	19 (41.3)
Never	62 (29.3)	57 (27.3)	57 (37.3)	38 (34.2)	12 (23.1)	11 (28.9)	12 (25)	10 (21.3)	12 (26.1)

Format of metrics in table is as follows (mean [SD] or n [%]). Report of the donor's race was either reported by the donor, the donor's family/next of kin, or abstracted from medical records. The classification categories were taken from the NIH and refer to geographically based categories that humans share (i.e., common history, nationality, or geographic distribution).

(at sample sizes of 50, 100, and 150) and determined the proportion of subsamples that included the smoking-associated CpG with the largest (cg01584760), median (cg20291548), and smallest (cg09138315) effect size magnitude observed in lung. In subsets with n = 50, none of the three CpGs were identified. For sample sizes of 100 and 150, the CpG with the largest effect size was detected in 54.0% and 98.2% of subsamples, respectively, and the CpG with the median effect size was discovered in 0.4% and 34.9% of subsamples, respectively (Table S3). Together, these results demonstrate the expected power for each tissue-specific analysis to detect effect sizes similar to those observed in lung.

For each tissue type, we specifically examined the CpGs associated with smoking in lung (from analyses of GTEx reported here), blood cells (3,722 CpGs reported in a recent review of 30 studies focused on the association of active smoking with DNAm and gene expression),[39] placenta tissue (443 CpGs),[17] and adipose tissue (42 CpGs)[13] (based on prior studies) to assess the evidence that some smoking-associated CpGs are shared across tissue types. Among the tissues examined, colon, ovary, whole blood, breast, and kidney showed the strongest evidence of enrichment for smoking-associated CpGs identified in GTEx lung tissue (Figure 1; Table S4). These tissue types were also enriched for CpGs identified previously in whole blood; in addition, muscle and prostate were enriched for CpGs previously identified in adipose tissue (Figure 1; Table S4). In contrast, with the exception of lung tissue, other tissue types were not enriched for smoking-associated CpGs identified previously in placental tissue (Figure 1; Table S4). The overlap of smoking-associated CpGs between tissue types and the overlap of genes annotated to these CpGs are characterized in Figures S5 and S6.

Since the abundance of CpGs detected in each tissue is dependent on sample size, we compared association estimates of smoking-associated CpG sites in lung to estimates of the same of set CpG sites in the other 8 tissues (Figure S7). Association estimates for CpG sites identified in lung showed clear positive correlations ($p < 1.9 \times 10^{-7}$) with estimates from all other tissue types (except muscle). While correlations were generally weak, our results indicate that many smoking effects in lung are present in other tissues, suggesting the lack of signal in other tissues is due in part to limited power. Additionally, we performed stratified EWAS by sex for lung and colon tissues. We observed effect size estimates across smoking-associated CpGs identified in our primary analysis (FDR < 0.05) were strongly correlated between males and females in both lung ($R^2 = 0.70$) and colon ($R^2 = 0.56$) (Figure S8).

Regions in which smoking is associated with both DNAm and gene expression

Examining the association of smoking status with DNAm and gene expression (in lung and colon), we observe that smoking is associated with both DNAm and gene expression in some regions. In lung tissue, notable gene regions included *CYP1B1*, *AHRR*, and *CYP1A1*, with *AHRR* expression and methylation also showing clear association with

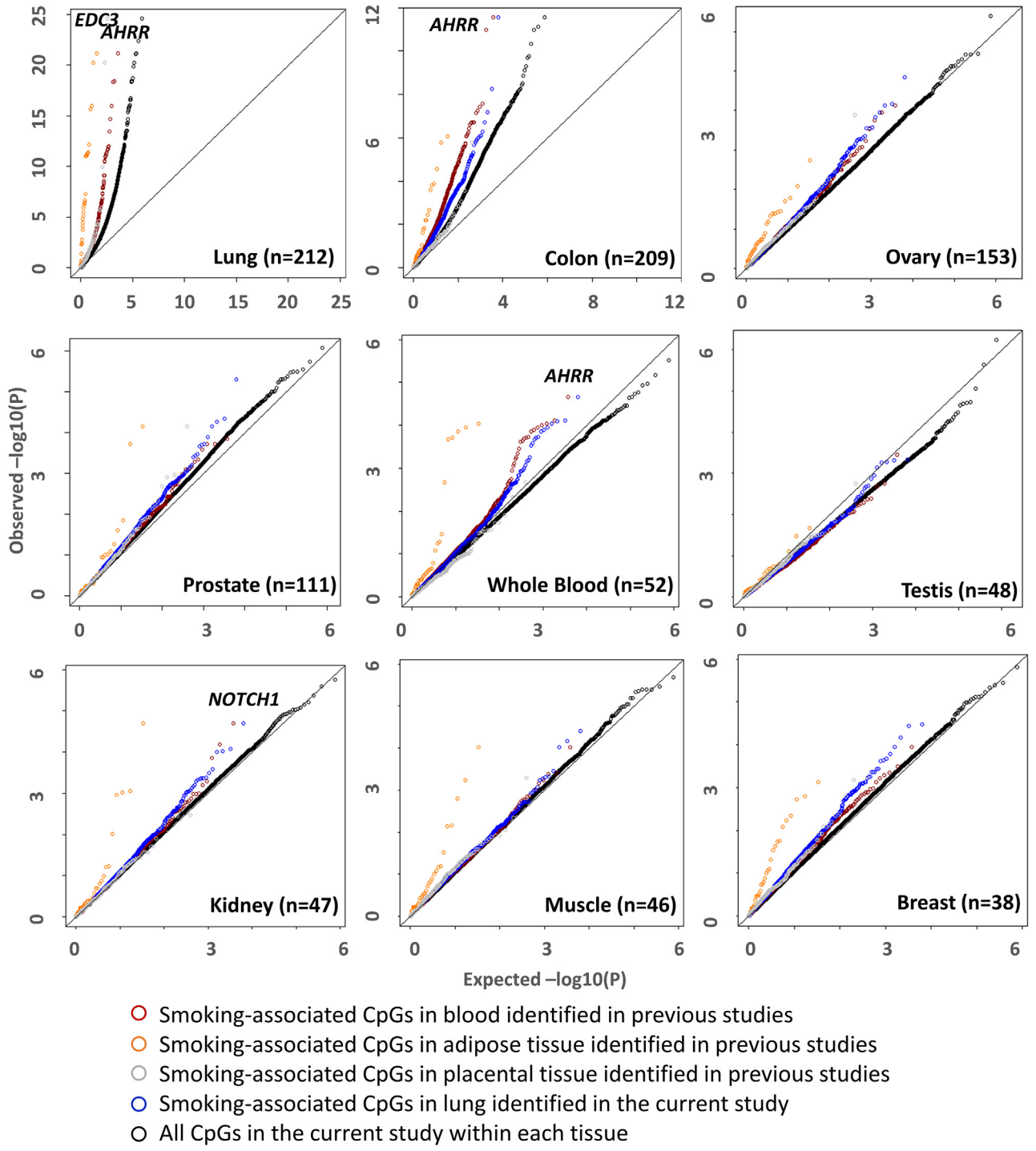

Figure 1. Quantile-quantile plots of p values showing the association between smoking status and DNAm by tissue type
Results are shown for genome-wide analyses in each tissue (black circles), for the 6,350 CpG associated with smoking based on GTEx lung samples (blue circles), and for the 3,722 CpGs associated with smoking based on prior studies of DNAm in blood samples[39] (red circles), adipose samples (orange circles),[13] and placenta samples (gray circles).[17] Several noteworthy genes with previously reported associations between smoking and DNAm are labeled including *AHRR*, *NOTCH1*, and *EDC3*.

smoking in colon tissue (Figure 2). The strongest smoking-related gene expression signal in both colon and lung was *GPR15*, a gene previously reported to be a biomarker of smoking in leukocytes (both expression and methylation).[56,57] The CpG with the strongest evidence of association in this region in colon (cg19859270, $p = 3.7 \times 10^{-8}$) did not pass QC for lung. Overall, 994 loci in lung and 3 in colon showed association of smoking with both DNAm and gene expression (Tables S5 and S6).

The top 10 smoking-associated DNAm features/regions for lung and colon are shown in Table 2. For lung, 7 of the 10 regions have been previously reported in prior

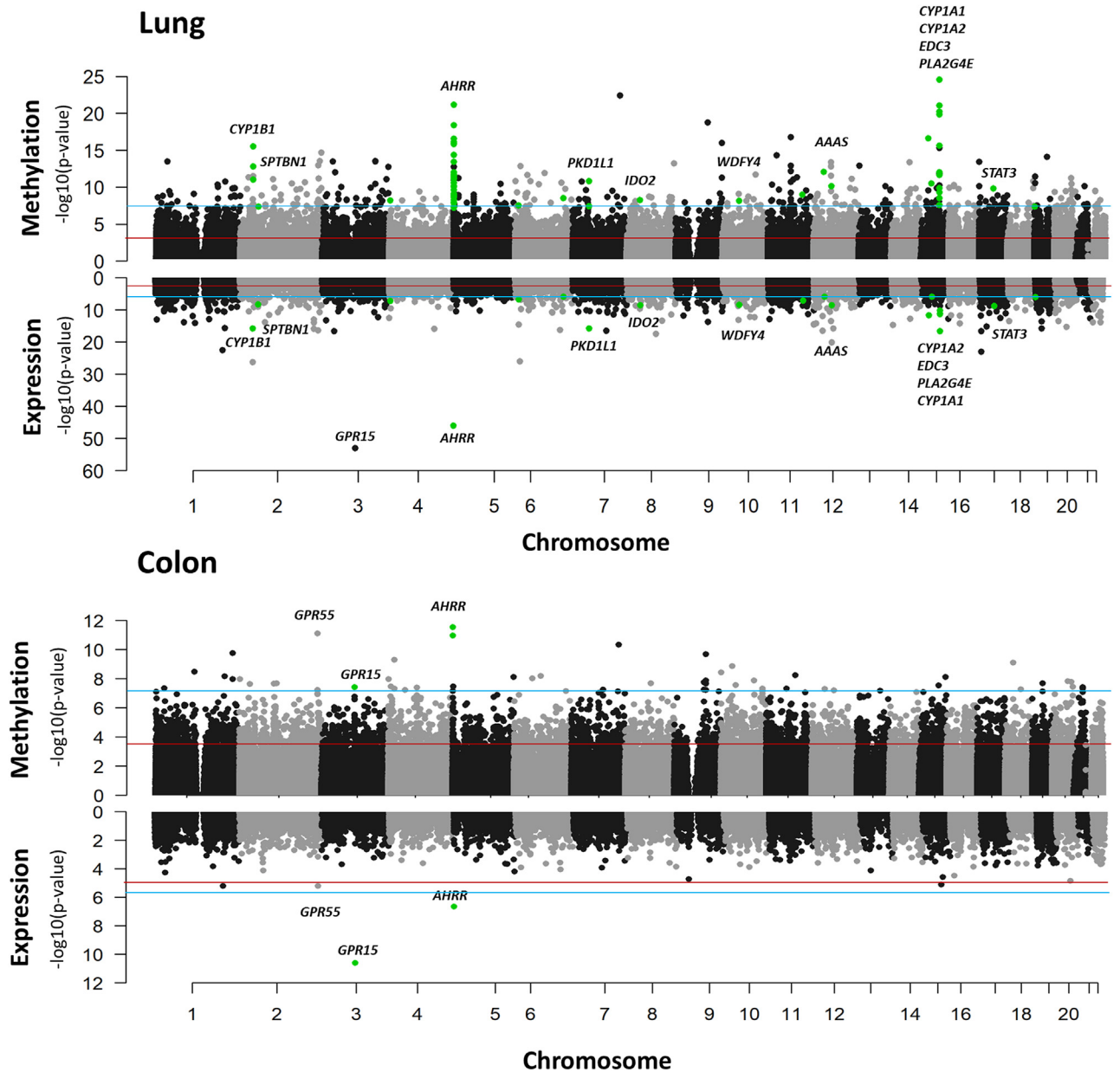

Figure 2. p values corresponding to the association of smoking status with DNAm and gene expression in lung (top) and colon (bottom) GTEx samples

Regions associated with both DNAm and gene expression based on a Bonferroni p value threshold are indicated with green dots. Gene expression analyses are based on 541 lung samples and 382 colon samples. Blue line represents the FDR threshold and red line represents the Bonferroni threshold.

studies of DNAm, including blood cells,[7,13,39,58–60] non-tumor lung tissue,[6] adipose tissue,[13] oral mucosa[61] and/or non-small-cell lung cancer.[62] Most of these regions contain multiple smoking-associated CpGs, including *AHRR* with 66 CpGs, *CYP1A1* with 18 CpGs, and *LRP5* with 15 CpGs. Six of these regions also showed evidence of association with smoking in colon (FDR 0.05): *HIPK2*, *AHRR*, *CYP1A1*, *AOPEP*, *LRP5*, and *CYP1B1*.

Among the top 10 smoking-associated regions in colon (Table 2), five have been previously reported in prior studies of non-tumor lung tissue[6] and/or blood.[25,58,63–65] However, among these 10 regions, only two (*AHRR* and *GPR55*) showed evidence of association with smoking in lung (FDR 0.05). The smoking-associated sites we identified in colon include CpGs annotated to *RHOU* and *WNK2*.

For several regions in which smoking was associated with both DNAm and gene expression (in both lung and colon), we examined the smoking and DNAm associations in detail. In the *AHRR* region (Figure 3), we observe clear differences between lung and colon with respect to the CpGs showing the clearest (based on p value) association with smoking. However, we observe some similarities

Table 2. Top smoking-associated CpGs in GTEx lung and colon samples (based on p value)

Tissue	Annotated gene[a] (or region)	Chr:Position	CpG	p value	CpGs passing FDR[b]	Region identified previously	Other tissues (FDR 0.05)	Association with expression (p value)
Lung	*EDC3*	15:74935742	cg26843110	2.61E-25	4	yes	no	up (5.63E-12)
	HIPK2	7:139420300	cg03224163	3.70E-23	4	yes	colon	down (0.05)
	AHRR	5:346695	cg04135110	6.97E-22	66	yes	colon	up (1.08E-46)
	CYP1A1[c]	15:75015502	cg23655854	8.26E-22	18	yes	colon	up (3.08E-17)
	AOPEP[d]	9:97544885	cg21081352	1.79E-19	5	no	colon	no
	5p15.33	5:1128194	cg03504128	3.76E-19	1	no	no	N/A
	LRP5	11:68079135	cg04840942	1.70E-17	15	yes	colon	no
	PLA2G4E	15:42313231	cg16167478	2.47E-17	6	no	no	up (2.61E-12)
	NOTCH1	9:139416102	cg14120703	1.04E-16	2	yes	no	down (0.04)
	CYP1B1	2:38296474	cg01584760	2.83E-16	10	yes	colon	up (1.90E-16)
Colon	*AHRR*	5:374252	cg04141806	2.85E-12	17	yes	lung	up (2.31E-7)
	GPR55[e]	2:231809610	cg08840017	7.61E-12	1	yes	no	up (6.31E-6)
	HIPK2	7:139366758	cg25748521	4.40E-11	2	yes	lung	no
	RHOU	1:228871677	cg27437294	1.73E-10	2	no	no	no
	WNK2	9:95947164	cg10281741	1.99E-10	5	no	no	no
	FAM184B	4:17783205	cg01886556	4.88E-10	4	yes	no	no
	LAMA3	18:21269793	cg25009504	7.59E-10	3	no	no	no
	NRP1	10:33624100	cg09009410	1.35E-09	4	no	lung	no
	NHLH2	1:116381475	cg24106636	3.17E-09	2	no	no	N/A
	DIP2C	10:735472	cg25488288	3.54E-09	2	yes	no	no

[a]Based on Illumina's annotation for the EPIC array. Cytoband is listed if there is no annotated gene.
[b]The number CpGs passing FDR (0.05) that are annotated to the gene listed (based on Illumina's annotation).
[c]*CYP1A1* and *EDC3* are in the same region, separated by < 25 kb.
[d]*AOPEP* is also known as *C9ORF3*.
[e]CpG cg08840017 was assigned to *GPR55* as it resides in a *GRP55* isoform.

between lung and colon with respect to patterns/clusters of increased and decreased methylation across the *AHRR* region (Figure S9). For example, decreased methylation among smokers is observed at CpG islands overlapping regulatory elements (based on ENCODE histone marks, DNase I hypersensitive sites (DHSs), and chromatin state), including the *AHRR* start site, for both lung and colon. These hypomethylated regions tend to have at least one site with a very low methylation level (beta value). In contrast, regions of increased methylation among smokers, in both lung and colon, tend to fall in the *AHRR* gene body, outside of regulatory elements coinciding with CpG islands (Figure 3). In both tissues, smoking is associated with increased expression of *AHRR* (lung $p = 9.9 \times 10^{-47}$; colon $p = 2.4 \times 10^{-7}$).

Sharper, more defined association signals (as compared to *AHRR*) were observed in both the *CYP1B1* and *CYP1A1* regions (Figure 4), for both lung and colon. While there are differences across tissues in terms of the specific CpGs showing the strongest association, these signals are located at CpG islands near the gene start site/promoter and show clear decreased methylation among smokers, with at least one smoking-associated CpG showing very low overall methylation levels (Figures 4B, S10, and S11). Smoking is associated with increased expression of *CYP1B1* (lung $p = 7.2 \times 10^{-27}$; colon $p = 0.002$) and *CYP1A1* (lung $p = 3 \times 10^{-17}$; colon $p = 8.1 \times 10^{-6}$).

Co-localization of *cis*-mQTLs and disease-related GWAS SNPs

To identify CpGs in lung and colon that may mediate the effects of smoking on lung or colon health, we first identified smoking-associated CpGs that are affected by an mQTL (using existing mQTL results from GTEx passing an FDR of 0.01).[27] For lung mQTLs, we determined if their lead SNPs were associated with lung-related phenotypes (FEV1/FVC, FVC, and lung adenocarcinoma) using GWAS summary statistics. Among our 2,478 smoking-associated CpGs in lung (FDR < 0.01), 566 are impacted by mQTLs in lung tissue. Among the 550 lead SNPs for these 566 lung mQTLs, 10 SNPs showed genome-wide significant associations with FEV1/FVC ($p < 5 \times 10^{-8}$) based on UK Biobank results[47] (Table S7). We found evidence of co-localization (between the mQTL and a FEV1/FVC GWAS signals)

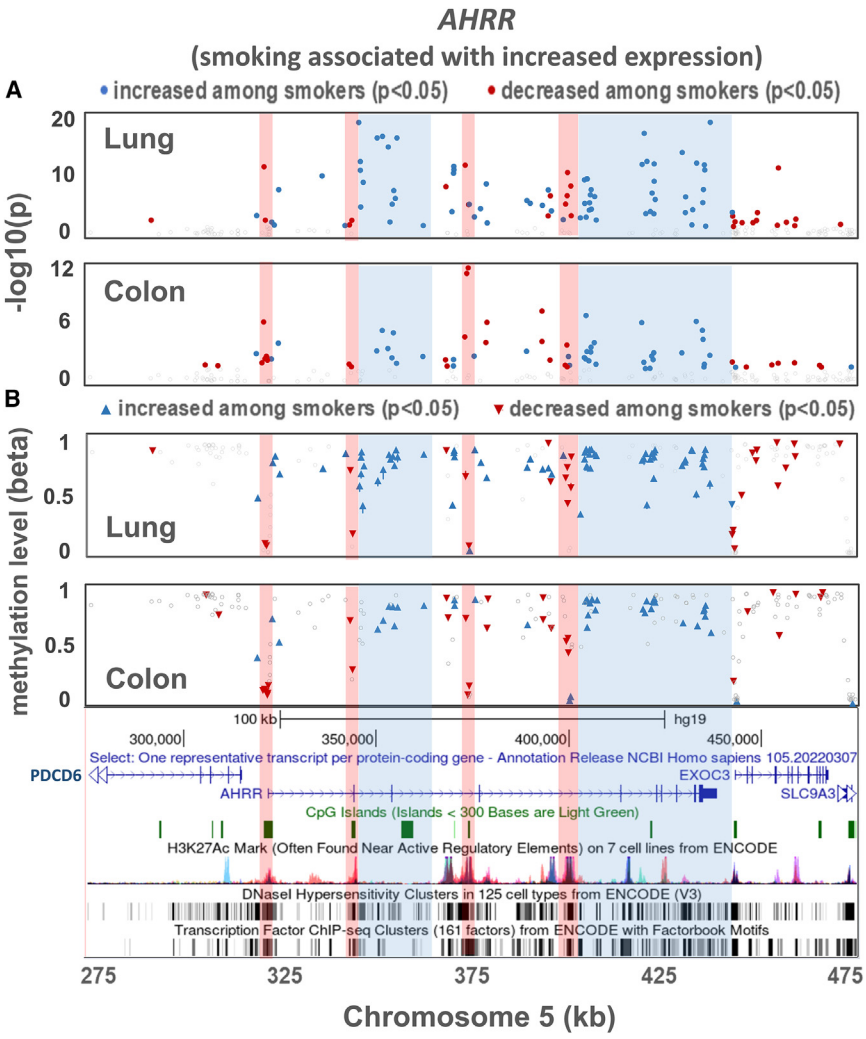

Figure 3. Association between smoking and DNA methylation for CpG sites in the *AHRR* region for lung and colon tissues
(A) p values for association.
(B) Beta values for each CpG reflecting the level of DNA methylation at the CpG. Beta values represent the average across all individuals (smokers and non-smokers). Upward arrowheads indicate DNAm values were higher in smokers for a given CpG compared to in non-smokers, while downward arrowheads indicate DNAm values were lower in smokers.

for 4 of the 10 SNPs identified (PP4 < 0.99, Figure S12). The strongest co-localization detected was for an mQTL (lead SNP rs7962469) affecting cg01996125, within the gene body of *ACVR1B* on chromosome 12 (PP4 = 0.99). Smoking was associated with decreased DNAm at cg01996125 (Figures 5A, 5D, and S13). The co-localized GWAS and mQTL signals also co-localized with an eQTL for *ACVR1B* (Figure 5B). The FEV1/FVC risk allele (G) was associated with decreased FEV1/FVC, increased DNAm at cg01996125 (and several surrounding CpGs, Figure 5C), and decreased *ACVR1B* expression (Figure 5D). Smoking was associated with decreased DNAm at cg01996125 and decreased *ACVR1B* expression (Figure 5D). Given that the risk allele (G) and smoking were both associated with decreased *ACVR1B* expression, but with opposite effects on cg01996125 methylation (Figure 5D), these results suggest the epigenetic mechanism (or response) linking smoking to repression of *ACVR1B* may be different than from the mechanism of the FEV1/FVC risk allele. We also performed an interaction analysis regressing cg01996125 methylation on an interaction term between rs7962469 and smoking status (while adjusting for other covariates in the primary lung EWAS analysis) and observed an interaction p value of 0.28, corroborating the notion that these genetic and epigenetic mechanisms leading to *ACVR1B* repression may be distinct. Despite these differences in genetic and environmental effects on cg01996125, repression of *ACVR1B* expression represents a potential mediator by which smoking (and the risk allele at rs7962469) impacts lung health. Additional examples of co-localization between FEV1/FVC GWAS signals and mQTLs include *SFTPA1* (PP4 = 0.99), *PRSS23* (PP4 = 0.96), and *MARCHF3*/*MARCH3* (PP4 = 0.96) (Figure S12). However, no eQTLs were observed for these genes.

We did not find evidence of co-localization between colon mQTLs and GWAS signals of either colorectal cancer or inflammatory bowel disease. If we relax the mQTL discovery threshold from an FDR of 0.01 to 0.1, we find evidence of co-localization between a colon *cis*-mQTL (for smoking-associated CpG cg13616097 located in the gene body of *WNT7B*) and a colorectal cancer GWAS signal (PP4 = 0.98).[50] We also find evidence of co-localization between a colon *cis*-mQTL (for smoking-associated CpG cg04048259 located downstream of *ZNF831*) and a GWAS signal for inflammatory bowel disease (PP4 = 0.99) (Table S8).[51]

Enrichment of smoking-associated CpGs within genomic features and biological pathways

Examining the distribution of hypermethylated and hypomethylated smoking-associated CpGs within genomic features, we observed that hypomethylated CpGs (FDR < 0.05) in colon were enriched in islands ($p < 10^{-5}$). In contrast, in lung, hypomethylated CpGs were depleted in islands ($p < 10^{-5}$) (Figure S14). Similarly, we observed different patterns of enrichment of smoking-associated CpGs sites in chromatin segmentation features between colon and lung. Both hypermethylated and hypomethylated lung CpGs showed enrichment in repressed polycomb states, whereas hypermethylated and hypomethylated

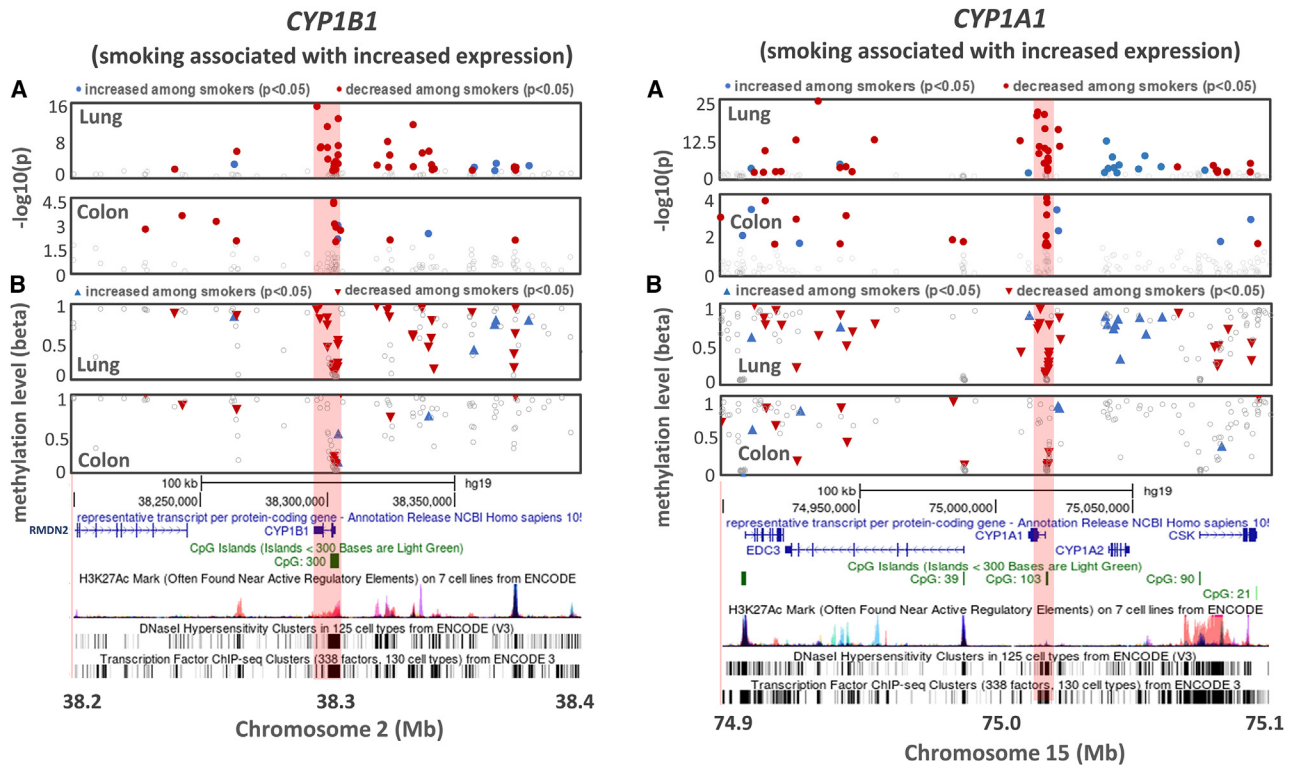

Figure 4. Association between smoking and DNA methylation for CpG sites in lung and colon for the *CYP1B1* and *CYP1A1* regions
Shown are p values for association (top) and beta values for each CpG reflecting the level of DNA methylation at the CpG (bottom). Beta values represent the average across all individuals (smokers and non-smokers). Upward arrows (blue) indicate DNAm values were higher in smokers for a given CpG compared to in non-smokers, while downward arrows (red) indicate DNAm values were lower in smokers.

colon CpGs showed enrichment in regions of active transcription (Figure S15). Finally, we observed clear enrichment of smoking-associated CpG sites across all TFBS in colon ($p = 1.48 \times 10^{-166}$) and lung ($p = 1.59 \times 10^{-105}$). The top enriched TFBS differed between the tissues and between hypermethylated and hypomethylated CpG sites within each tissue types (Figures S16 and S17).

We conducted pathway analyses of the 6,350 smoking-associated CpGs (from lung) assigned to 2,948 genes, which revealed 17 overrepresented biological pathways (FDR of 0.05). The top ten pathways identified using all 6,350 smoking-associated CpGs (Table 3) included xenobiotic metabolism ($p = 9.3 \times 10^{-4}$), a pathway that included several of our strongest signals already described (*AHRR*, *CYP1A1*, and *CYP1B1*), highlighting the response of biotransformation genes to the chemicals in cigarette smoke. Numerous cancer-related pathways also showed enrichment, including tumor necrosis factor alpha (TNF-α),[66] signaling via nuclear factor kappa beta (NFKB), apoptosis,[67] p53,[68] IL6-JAK-STAT3 signaling,[69] early estrogen response,[70] ultraviolet (UV) radiation response,[71] transforming growth factor β signaling,[72] hypoxia,[73,74] MTORC1 signaling,[75] and cholesterol homeostasis.[76] We additionally explored enrichment of gene sets related to human diseases among KEGG pathways and identified a pathway for lipids and atherosclerosis (Table 3). When examining enrichment separately for hypomethylated CpGs (n = 4,637), which comprised the majority of smoking-associated CpGs, we observed similar pathway enrichments (Table S9); the number of hypermethylated, smoking-associated CpGs was small and underpowered for pathway analysis. Similar analyses of the 2,735 smoking-associated CpGs from colon (assigned to 1,369 genes), of which 94.6% were hypomethylated, resulted in the detection of only two enriched Hallmark gene sets (FDR 0.05), epithelial to mesenchymal transition, and UV response (down regulation) (Table S10).

While gene set enrichment analysis resulted in six Hallmark pathways that were shared between primary (ever vs. never) and secondary (current vs. never) analyses, there were also several pathways specific to either analysis with more pathways unique to the secondary analysis (Tables 3 and S11; Figure S18). When exploring enrichment of human diseases among KEGG pathways in our secondary analysis, we identified enrichment of a pathway for atherosclerosis, similar to our primary analysis. The secondary analysis further identified KEGG pathways for human health conditions including circadian entrainment and insulin resistance, thereby supporting the notion that dysregulation of the epigenome impacts these previously reported, smoking-associated health conditions.[77,78] Together, these results suggest that epigenetic dysregulation, and the effects of this dysregulation on human health and disease, may be more pronounced between current vs. never smokers in comparison to ever vs. never smokers.

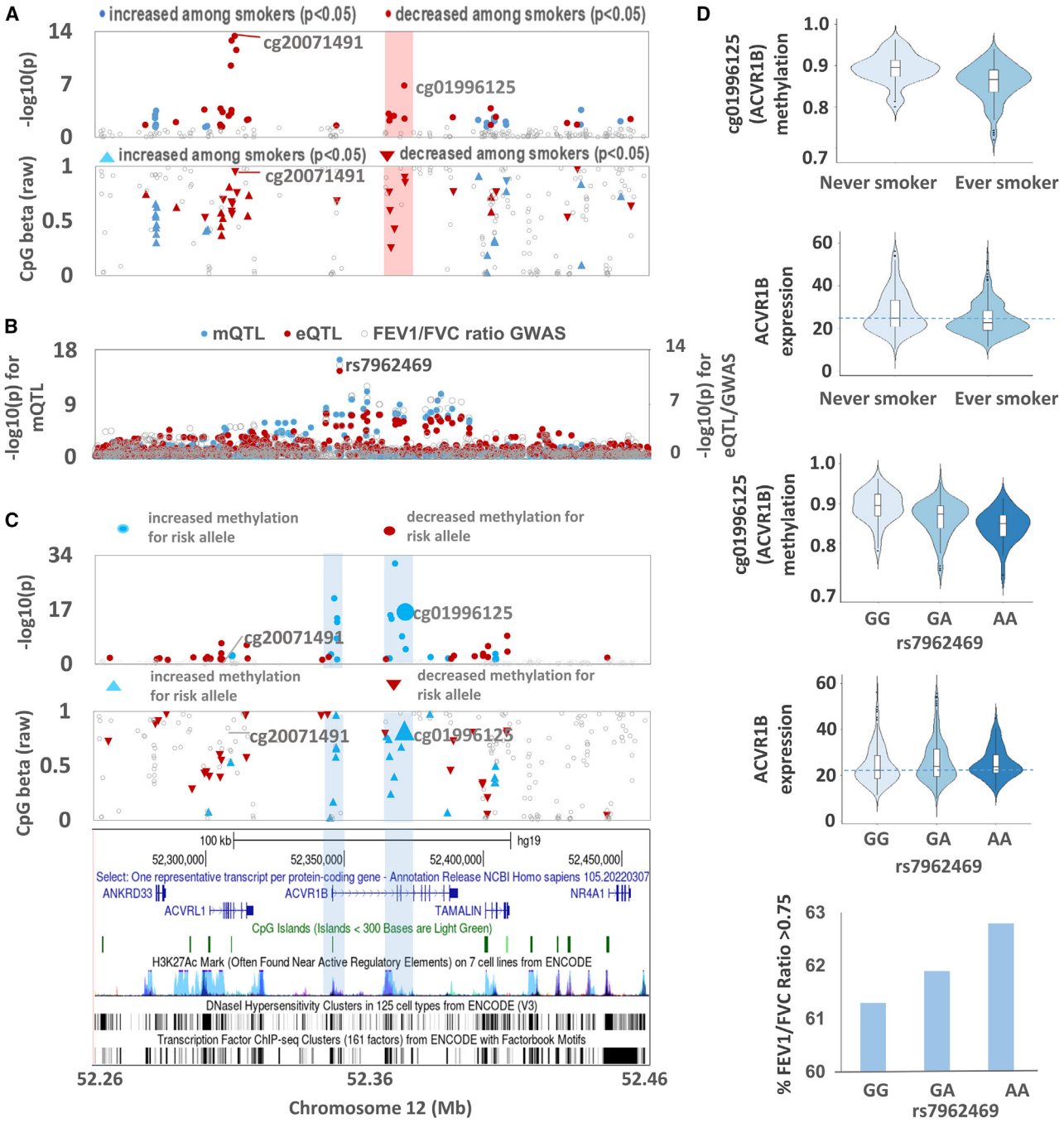

Figure 5. Co-localization of mQTL (for cg01996125), eQTL for *ACVR1B*, and lung function GWAS signal in the chromosome 12q13.13 region

(A) Plots of smoking EWAS (beta and p values) by effect direction for this region.

(B) Plot of p values for cg01996125 mQTL, *ACVR1B* eQTL, and FEV1/FVC ratio GWAS showing co-localization of all three association signals.

(C) Plot of the association of rs7962469 (FEV1/FVC ratio risk allele G) with all CpGs in the *ACVR1B* region.

(D) Distribution of cg01996125 methylation and *ACVR1B* expression by smoking status and rs7962469 genotype. Beta values represent the average across all individuals (smokers and non-smokers). Upward arrowheads (blue) indicate DNAm values were higher in smokers for a given CpG compared to in non-smokers, while downward arrowheads (red) indicate DNAm values were lower in smokers.

Cell-type-specific effects of smoking on DNAm

To search for evidence of cell-type-specific effects of smoking on DNAm, we tested the interaction between smoking and the EPISCORE-derived cell-type proportion estimates. In lung, cell types estimated, from most abundant to least, were endothelial cells, macrophages, epithelial cells, stromal cells, granulocytes, lymphocytes, and monocytes (Figure S1). In colon, cell types estimated, from most abundant to least, were lymphocytes, enterochromaffin cells, stromal cells, myeloid cells, and epithelial cells (Figure S1). The

Table 3. Pathway analysis of smoking-associated CpGs detected in lung tissue

Description	Genes in gene set	Genes with smoking-associated CpGs[a]	Enrichment P	FDR-adjusted p value
Hallmark gene sets				
TNF-alpha signaling via NFKb	199	56	8.40E-06	4.20E-04
Apoptosis	155	46	1.69E-04	4.22E-03
P53 pathway	196	53	3.35E-04	5.59E-03
Xenobiotic metabolism	197	49	9.01E-04	1.13E-02
Early response to estrogen	194	60	2.06E-03	1.69E-02
IL6-JAK-STAT3 signaling	81	23	2.21E-03	1.69E-02
UV response (down regulated)	142	52	2.36E-03	1.69E-02
TGF-beta signaling	53	21	3.59E-03	2.25E-02
Hypoxia	190	51	4.52E-03	2.51E-02
Cholesterol homeostasis	71	21	6.29E-03	3.14E-02
Myogenesis	195	55	7.08E-03	3.22E-02
IL2-STAT5 signaling	194	51	9.42E-03	3.93E-02
Adipogenesis	196	45	1.05E-02	4.02E-02
Androgen response	97	30	1.23E-02	4.17E-02
Bile acid metabolism	110	26	1.27E-02	4.17E-02
MTORC1 signaling	194	44	1.33E-02	4.17E-02
KEGG Pathways				
Parathyroid hormone synthesis, secretion, and action	105	42	9.85E-05	3.50E-02
Lipid and atherosclerosis	205	56	2.11E-04	3.75E-02

[a]Genes with CpGs (as assigned by Illumina) that are associated with smoking.

distribution of interaction p values (Figures S19 and S20) suggests that effects of smoking on methylation at certain CpG sites varies according to the abundance of the cell types present, including endothelial cells, lymphocytes, monocytes, and macrophages. The CpGs showing the strongest evidence of interaction between smoking and cell types (i.e., involved in SxCT) in lung tissue are listed in Table S12. Most CpGs involved in SxCT also show evidence of a residual "main effect" of smoking on the CpG in a direction consistent with the interaction effect. This observation suggests that a joint test of the main effect and the cell type interaction could potentially boost power for detecting environmental effects in DNAm/EWAS studies similar to methods developed for the GWAS context.[79] Interestingly, our most significant EWAS signals from lung did not show clear evidence of interaction by cell type (Table S13).

Discussion

In this work, we generated and analyzed genome-wide DNAm data for 916 human tissue samples, representing 9 unique tissue types, and characterized the association of smoking status with genome-wide measures of DNAm. We detected >6,000 smoking-associated CpGs in lung, a tissue type that showed more prominent effects of smoking compared to other tissue types. Our results show that while DNAm in some regions is impacted by smoking in multiple tissue types, the specific CpGs affected (and the magnitude of those effects) can differ between tissues. Several mQTLs impacting smoking-associated CpGs in lung tissue were found to co-localize with association signals from GWASs of lung function, suggesting smoking-related epigenetic alterations may mediate the effects of smoking on lung health. Smoking-associated CpGs were enriched in pathways related to xenobiotic metabolism and cancer.

Lung tissue had a much larger number of CpGs showing association with smoking status compared to colon and the other 7 tissue types examined. While this difference is in part due to the larger sample size for lung tissue, smoking effects in lung also appear to be more abundant after accounting for sample size differences. This is not unexpected, as lung tissue is exposed to tobacco combustion products directly via inhalation (as well as via the blood stream). In contrast, the other tissues examined are primarily exposed to tobacco combustion products via the blood stream, which carries chemicals that enter the pulmonary circulation (from the lungs) and then travel to other organs (although the colon could potentially be exposed to tobacco-derived

chemicals through the gastrointestinal tract). Furthermore, we discovered that the number of CpGs passing an FDR of 0.05 between current vs. never smokers was around 2-fold that between ever vs. never smokers in both lung and colon (Table S2). Given that the ever-smokers category includes former smokers, these results suggest that smoking cessation may lead to a reduced impact of smoking on the epigenome compared to continued smoking. This conclusion aligns with studies reporting the benefits of smoking cessation related to reduced risk of adverse health conditions, including lung cancer[80] and cardiovascular disease.[81]

Our results show that genomic regions affected by smoking can be shared across tissue types, consistent with prior studies,[3,82] as we observe enrichment for smoking-related CpGs (identified from prior studies of blood) in other tissue types, including colon, ovary, and kidney. For example, *NOTCH1* contained top CpGs for both lung and kidney (Figure 1), consistent with a prior study of adipose tissue.[13] The enrichment observed for kidney (n = 47) suggests more signals are likely present (and observable at larger sample sizes), reflecting effects of smoking that may be consistent with the strong impact of smoking on the risk of renal cell carcinoma.[83] However, for a given region, we observe that the specific CpGs associated with smoking and their relative magnitudes can vary substantially by tissue type (as observed for *AHRR*, *CYP1A1*, and *CYP1B1*). Thus, these findings suggest that while it is possible to assess exposure effects on DNAm within genes in accessible tissues to make inferences about effects in target tissues (for effects that are shared common across tissue types), it is more challenging to infer which specific CpGs may be impacted across tissues.

Our top smoking-associated regions in lung include several regions previously identified in blood, including the top three smoking-associated genes/regions involved in xenobiotic metabolism: *AHRR*, *CYP1A1*, and *CYP1B1*. Each of these regions has been identified in prior studies,[7,9,19–22] and each region shows an association of smoking with increased gene expression. Each gene has a biologically plausible response to smoking, for example, *AHRR* encodes a transcription factor with key roles in sensing xenobiotics (including aromatic hydrocarbons) and regulation of metabolizing enzymes including *CYP1A1*. Our study has also discovered smoking associated CpGs, including the *AOPEP*, *PLA2G4E*, and *PA2G4P4* gene regions in lung, as well as the *RHOU* and *WNK2* gene regions in colon. Of note, PLA2G4E is part of the secretory phospholipase A2 family, a group of enzymes secreted during inflammation and involved in the cleavage of phospholipids during synthesis of eicosanoids, which are lipid mediators released by alveolar macrophages in response to toxic elements.[84–86] Additionally, experimental evidence has shown that knockdown of *RHOU*, an atypical member of the *RHO* family, leads to higher proliferation and reduced apoptosis of colon cancer cells[87]; our discovery of smoking-associated CpGs in *RHOU* corroborates previous literature establishing smoking as a causal factor for colorectal carcinoma.[1,88]

Interestingly, we observed striking differences in the enrichment of hypermethylated and hypomethylated smoking-associated CpGs within genomic features (CpG islands and chromatin segmentation) for lung compared to colon. We additionally observed differences in the top enriched TFBS for lung and colon. Hypomethylated CpGs in colon were enriched in islands and active transcription regions, largely consistent with what we observe for our top smoking-associated regions involved in xenobiotic metabolism including *AHRR*, *CYP1A1*, and *CYP1B1* (Figures 3 and 4). However, in lung, we observe depletion in islands and enrichment in repressed transcription states for both hypermethylated and hypomethylated CpGs. We speculate that these differences are due to the difference in the nature of exposure in lung versus colon. In lung, smoking effects appear larger and more pervasive, resulting in greater power to detect more subtle effects (e.g., transcriptionally repressed regions) beyond those related to response of xenobiotic metabolism genes.

Many of the smoking-associated CpGs identified in lung are also impacted by inherited genetic variation (i.e., mQTLs), including variants impacting lung health. Analyses of co-localization between FEV1/FVC GWAS hits and mQTLs of smoking-associated CpGs identified cg01996125 in *ACVR1B* as an epigenetic feature potentially involved in mediation of the effects of smoking on lung health. *ACVR1B*, expression of which is inversely associated with both smoking and the FEV1/FVC risk allele, is a part of the transforming growth factor beta (TGFR-β) superfamily contributing to inflammation and initiation of airway remodeling.[89] Repression of *ACVR1B* (and any associated epigenetic alterations) may be a potential mediating pathway by which smoking (and the risk allele) have detrimental effects on lung health.

We estimated the proportions of individual cell types in lung and colon and observed substantial fractions of immune cells in both tissue types, including lymphocytes and myeloid cells (e.g., monocytes and macrophages). It is possible that the immune cell component of these tissues contributes to the observed overlap in smoking-associated regions between these tissues (and with regions previously reported in whole blood). To determine if effects of smoking on DNAm differ by cell type, we examined the interaction between smoking and the inferred cell-type proportions in lung tissue, identifying multiple CpGs potentially impacted by cell-type-specific effects. For example, we identified CpGs involved in SxCT, located upstream of *COPS6* (SxCT: lymphocyte) and in the second intron of *WASF2* (SxCT: monocyte and macrophage). COPS6 has been shown to promote tumor-infiltrating lymphocyte signaling in breast oncogenesis, facilitate tumor evasion,[90] and promote the growth of various lung cancer cell lines.[91] WASF2 mediates macrophage motility and phagocytosis by interacting with filamentous actin,[92,93] with *in vitro* studies demonstrating immunoreactivity of WASF2 in many lung adenocarcinomas.[94] Overall, our results

suggest that analyses of SxCT can identify CpGs missed in analyses of marginal associations (i.e., main effects) alone, implicating genes with biologically plausible roles in cancer. Additional research is needed to explore mechanisms by which the effects of smoking are mediated by genes expressed in specific cell types. While the method we use for estimating cell-type composition (EPISCORE) is well established, it has not been validated specifically on GTEx samples. Additional work to further validate and characterize DNAm-based cellular deconvolution methods across diverse tissue types will improve our understanding of the shared cell-type-specific effects across tissues. Single-cell studies of DNAm in human tissues can also be leveraged to explore such mechanisms.

While we identified many previously unreported smoking-associated regions in disease-relevant tissues, including effects that are shared across tissues and tissue specific, this study is limited by the lack of whole-genome data on DNAm as the EPIC array is only able to capture a small fraction (~2%) of all CpGs in the human genome. Additionally, we had small sample sizes for some tissues (e.g., kidney n = 48, muscle n = 46), which limited our power to detect associations. Therefore, larger studies of diverse tissues are needed to validate our results and generate additional data regarding the similarities and differences of DNAm across tissues. Overall, this work highlights the utility of using a multi-tissue approach to assess the effects of smoking on the human epigenome.

Data and code availability

Scripts to perform epigenome-wide association analyses are located at https://github.com/james-li-projects/SmokingEWAS.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.02.012.

Acknowledgments

This work was supported by grants U01 HG007601 (to B.L.P.), R35ES028379 (to B.L.P.), 2R01 GM108711 (to L.S.C), and U24 CA210993-SUB (to L.S.C) and was completed in part with computational resources provided by the Center for Research Informatics at the University of Chicago. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. We thank the donors and their families for their generous gifts of biospecimens to the GTEx research project; the Genomics Platform at the Broad Institute for data generation; F. Aguet, J. Nedzel, and K. Ardlie for sample-delivery logistics and data-release management. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

J.L.L. and N.J. performed analyses, interpreted the data, and wrote the main manuscript text. L.I.T. contributed to manuscript writing, data analysis, and interpreting results. L.T. performed analyses and prepared manuscript figures. M.G.K. and F.J. generated DNA methylation data. K.D. and M.O. performed data processing and quality control. L.S.C. advised on statistical analyses. B.L.P. conceived the project and contributed to writing/editing and data interpretation.

Declaration of interests

The authors declare no competing interests.

Received: August 14, 2023
Accepted: February 21, 2024
Published: March 14, 2024

Web resources

AnVIL, https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V9_hg38

dbGaP, GTEx data, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v9.p2

epitools: Epidemiology Tools, R package, https://cran.r-project.org/web/packages/epitools/

GEO, DNAm normalized data, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213478

GTEx Portal, https://gtexportal.org/home/

References

1. Lushniak, B.D., Samet, J.M., Pechacek, T.F., Norman, L.A., and Taylor, P.A. (2014). The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. https://stacks.cdc.gov/view/cdc/21569/cdc_21569_DS1.pdf.
2. Akhmetova, D.A., Kozlov, V.V., and Gulyaeva, L.F. (2022). New Insight into the Role of AhR in Lung Carcinogenesis. Biochemistry. *87*, 1219–1225. https://doi.org/10.1134/s0006297922110013.
3. Lee, K.W.K., and Pausova, Z. (2013). Cigarette smoking and DNA methylation. Front. Genet. *4*, 132.
4. Mingay, M., Chaturvedi, A., Bilenky, M., Cao, Q., Jackson, L., Hui, T., Moksa, M., Heravi-Moussavi, A., Humphries, R.K., Heuser, M., and Hirst, M. (2018). Vitamin C-induced epigenomic remodelling in IDH1 mutant acute myeloid leukaemia. Leukemia *32*, 11–20. https://doi.org/10.1038/leu.2017.171.
5. Terry, M.B., Delgado-Cruzata, L., Vin-Raviv, N., Wu, H.C., and Santella, R.M. (2011). DNA methylation in white blood cells. Epigenetics *6*, 828–837. https://doi.org/10.4161/epi.6.7.16500.
6. Stueve, T.R., Li, W.Q., Shi, J., Marconett, C.N., Zhang, T., Yang, C., Mullen, D., Yan, C., Wheeler, W., Hua, X., et al. (2017). Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. Hum. Mol. Genet. *26*, 3014–3027. https://doi.org/10.1093/hmg/ddx188.
7. Gao, X., Jia, M., Zhang, Y., Breitling, L.P., and Brenner, H. (2015). DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic

review of DNA methylation studies. Clin. Epigenetics *7*, 113. https://doi.org/10.1186/s13148-015-0148-3.

8. Breitling, L.P., Yang, R., Korn, B., Burwinkel, B., and Brenner, H. (2011). Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. Am. J. Hum. Genet. *88*, 450–457. https://doi.org/10.1016/j.ajhg.2011.03.003.
9. Ambatipudi, S., Cuenin, C., Hernandez-Vargas, H., Ghantous, A., Le Calvez-Kelm, F., Kaaks, R., Barrdahl, M., Boeing, H., Aleksandrova, K., Trichopoulou, A., et al. (2016). Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics *8*, 599–618.
10. Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J., Peters, A., et al. (2013). Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. PLoS One *8*, e63812. https://doi.org/10.1371/journal.pone.0063812.
11. Ringh, M.V., Hagemann-Jensen, M., Needhamsen, M., Kular, L., Breeze, C.E., Sjöholm, L.K., Slavec, L., Kullberg, S., Wahlström, J., Grunewald, J., et al. (2019). Tobacco smoking induces changes in true DNA methylation, hydroxymethylation and gene expression in bronchoalveolar lavage cells. EBioMedicine *46*, 290–304. https://doi.org/10.1016/j.ebiom.2019.07.006.
12. Siemelink, M.A., van der Laan, S.W., Haitjema, S., van Koeverden, I.D., Schaap, J., Wesseling, M., de Jager, S.C.A., Mokry, M., van Iterson, M., Dekkers, K.F., et al. (2018). Smoking is Associated to DNA Methylation in Atherosclerotic Carotid Lesions. Circ. Genom. Precis. Med. *11*, e002030. https://doi.org/10.1161/circgen.117.002030.
13. Tsai, P.-C., Glastonbury, C.A., Eliot, M.N., Bollepalli, S., Yet, I., Castillo-Fernandez, J.E., Carnero-Montoro, E., Hardiman, T., Martin, T.C., Vickers, A., et al. (2018). Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. Clin. Epigenetics *10*, 126. https://doi.org/10.1186/s13148-018-0558-0.
14. Barcelona, V., Huang, Y., Brown, K., Liu, J., Zhao, W., Yu, M., Kardia, S.L.R., Smith, J.A., Taylor, J.Y., and Sun, Y.V. (2019). Novel DNA methylation sites associated with cigarette smoking among African Americans. Epigenetics *14*, 383–391. https://doi.org/10.1080/15592294.2019.1588683.
15. Koo, H.-K., Morrow, J., Kachroo, P., Tantisira, K., Weiss, S.T., Hersh, C.P., Silverman, E.K., and DeMeo, D.L. (2021). Sex-specific associations with DNA methylation in lung tissue demonstrate smoking interactions. Epigenetics *16*, 692–703. https://doi.org/10.1080/15592294.2020.1819662.
16. Joubert, B.R., Felix, J.F., Yousefi, P., Bakulski, K.M., Just, A.C., Breton, C., Reese, S.E., Markunas, C.A., Richmond, R.C., Xu, C.J., et al. (2016). DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. Am. J. Hum. Genet. *98*, 680–696. https://doi.org/10.1016/j.ajhg.2016.02.019.
17. Everson, T.M., Vives-Usano, M., Seyve, E., Cardenas, A., Lacasaña, M., Craig, J.M., Lesseur, C., Baker, E.R., Fernandez-Jimenez, N., Heude, B., et al. (2021). Placental DNA methylation signatures of maternal smoking during pregnancy and potential impacts on fetal growth. Nat. Commun. *12*, 5095. https://doi.org/10.1038/s41467-021-24558-y.
18. Shenker, N.S., Polidoro, S., van Veldhoven, K., Sacerdote, C., Ricceri, F., Birrell, M.A., Belvisi, M.G., Brown, R., Vineis, P., and Flanagan, J.M. (2013). 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. *22*, 843–851. https://doi.org/10.1093/hmg/dds488.
19. Park, S.L., Patel, Y.M., Loo, L.W.M., Mullen, D.J., Offringa, I.A., Maunakea, A., Stram, D.O., Siegmund, K., Murphy, S.E., Tiirikainen, M., and Le Marchand, L. (2018). Association of internal smoking dose with blood DNA methylation in three racial/ethnic populations. Clin. Epigenetics *10*, 110. https://doi.org/10.1186/s13148-018-0543-7.
20. Zhang, Y., Elgizouli, M., Schöttker, B., Holleczek, B., Nieters, A., and Brenner, H. (2016). Smoking-associated DNA methylation markers predict lung cancer incidence. Clin. Epigenetics *8*, 127.
21. Haase, T., Müller, C., Krause, J., Röthemeier, C., Stenzig, J., Kunze, S., Waldenberger, M., Münzel, T., Pfeiffer, N., Wild, P.S., et al. (2018). Novel DNA Methylation Sites Influence GPR15 Expression in Relation to Smoking. Biomolecules *8*, 74. https://doi.org/10.3390/biom8030074.
22. Monick, M.M., Beach, S.R.H., Plume, J., Sears, R., Gerrard, M., Brody, G.H., and Philibert, R.A. (2012). Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. Am. J. Med. Genet. B Neuropsychiatr. Genet. *159B*, 141–151. https://doi.org/10.1002/ajmg.b.32021.
23. Fasanelli, F., Baglietto, L., Ponzi, E., Guida, F., Campanella, G., Johansson, M., Grankvist, K., Johansson, M., Assumma, M.B., Naccarati, A., et al. (2015). Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat. Commun. *6*, 10192. https://doi.org/10.1038/ncomms10192.
24. Baglietto, L., Ponzi, E., Haycock, P., Hodge, A., Bianca Assumma, M., Jung, C.H., Chung, J., Fasanelli, F., Guida, F., Campanella, G., et al. (2017). DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. Int. J. Cancer *140*, 50–61. https://doi.org/10.1002/ijc.30431.
25. Joubert, B.R., Håberg, S.E., Nilsen, R.M., Wang, X., Vollset, S.E., Murphy, S.K., Huang, Z., Hoyo, C., Midttun, Ø., Cupul-Uicab, L.A., et al. (2012). 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ. Health Perspect. *120*, 1425–1431.
26. Markunas, C.A., Xu, Z., Harlid, S., Wade, P.A., Lie, R.T., Taylor, J.A., and Wilcox, A.J. (2014). Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. Environ. Health Perspect. *122*, 1147–1153.
27. Oliva, M., Demanelis, K., Lu, Y., Chernoff, M., Jasmine, F., Ahsan, H., Kibriya, M.G., Chen, L.S., and Pierce, B.L. (2023). DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. Nat. Genet. *55*, 112–122. https://doi.org/10.1038/s41588-022-01248-z.
28. Consortium, G. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.
29. Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A., Britton, A., Buia, S.A., Compton, C.C., DeLuca, D.S., Peter-Demchok, J., Gelfand, E.T., et al. (2015). A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreserv. Biobank. *13*, 311–319.

30. Siminoff, L.A., Wilson-Genderson, M., Gardiner, H.M., Mosavel, M., and Barker, K.L. (2018). Consent to a Postmortem Tissue Procurement Study: Distinguishing Family Decision Makers' Knowledge of the Genotype-Tissue Expression Project. Biopreserv. Biobank. *16*, 200–206.
31. Heiss, J.A., and Just, A.C. (2018). Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. Clin. Epigenetics *10*, 73. https://doi.org/10.1186/s13148-018-0504-1.
32. Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., and Clark, S.J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. *17*, 208. https://doi.org/10.1186/s13059-016-1066-1.
33. GTEx Consortium; Laboratory, Data Analysis &Coordinating Center LDACC—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx eGTEx groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; and Biospecimen Collection Source Site—NDRI (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. https://doi.org/10.1038/nature24277.
34. GTEx Consortium, Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.
35. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25.
36. Smyth GK. limma: Linear Models for Microarray Data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor: Springer-Verlag. p. 397-420.
37. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.
38. Demanelis, K., Argos, M., Tong, L., Shinkle, J., Sabarinathan, M., Rakibuz-Zaman, M., Sarwar, G., Shahriar, H., Islam, T., Rahman, M., et al. (2019). Association of Arsenic Exposure with Whole Blood DNA Methylation: An Epigenome-Wide Study of Bangladeshi Adults. Environ. Health Perspect. *127*, 057011. https://doi.org/10.1289/ehp3849.
39. Silva, C.P., and Kamens, H.M. (2021). Cigarette smoke-induced alterations in blood: A review of research on DNA methylation and gene expression. Exp. Clin. Psychopharmacol *29*, 116–135.
40. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. https://doi.org/10.1038/nature14248.
41. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.
42. Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Sci. *28*, 1947–1951.
43. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.
44. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. *51*, D587–D592.
45. Phipson, B., Maksimovic, J., and Oshlack, A. (2016). missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics *32*, 286–288. https://doi.org/10.1093/bioinformatics/btv560.
46. Geeleher, P., Hartnett, L., Egan, L.J., Golden, A., Raja Ali, R.A., and Seoighe, C. (2013). Gene-set analysis is severely biased when applied to genome-wide methylation data. Bioinformatics *29*, 1851–1857.
47. Shrine, N., Guyatt, A.L., Erzurumluoglu, A.M., Jackson, V.E., Hobbs, B.D., Melbourne, C.A., Batini, C., Fawcett, K.A., Song, K., Sakornsakolpat, P., et al. (2019). New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat. Genet. *51*, 481–493. https://doi.org/10.1038/s41588-018-0321-7.
48. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. Elife *7*, e34408. https://doi.org/10.7554/elife.34408.
49. Elsworth, B., Lyon, M., Alexander, T., Liu, Y., Matthews, P., Hallett, J., Bates, P., Palmer, T., Haberland, V., Smith, G.D., et al. (2020). The MRC IEU OpenGWAS data infrastructure. Preprint at bioRxiv. https://doi.org/10.1101/2020.08.10.244293.
50. Fernandez-Rozadilla, C., Timofeeva, M., Chen, Z., Law, P., Thomas, M., Schmit, S., Díez-Obrero, V., Hsu, L., Fernandez-Tajes, J., Palles, C., et al. (2023). Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. Nat. Genet. *55*, 89–99.
51. De Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L., Rice, D.L., Gutierrez-Achury, J., Ji, S.G., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat. Genet. *49*, 256–261.
52. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. *10*, e1004383. https://doi.org/10.1371/journal.pgen.1004383.
53. Liu, B., Gloudemans, M.J., Rao, A.S., Ingelsson, E., and Montgomery, S.B. (2019). Abundant associations with gene expression complicate GWAS follow-up. Nat. Genet. *51*, 768–769. https://doi.org/10.1038/s41588-019-0404-0.
54. Zhu, T., Liu, J., Beck, S., Pan, S., Capper, D., Lechner, M., Thirlwell, C., Breeze, C.E., and Teschendorff, A.E. (2022). A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. Nat. Methods *19*, 296–306. https://doi.org/10.1038/s41592-022-01412-7.
55. Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S.E., Hamel, A.R., Viñuela, A., Roberts, A.L., et al. (2020). Cell type–specific genetic regulation of gene expression across human tissues. Science *369*, eaaz8528.

56. Bauer, M., Linsel, G., Fink, B., Offenberg, K., Hahn, A.M., Sack, U., Knaack, H., Eszlinger, M., and Herberth, G. (2015). A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. Clin. Epigenetics *7*, 81. https://doi.org/10.1186/s13148-015-0113-1.
57. Obeidat, M., Ding, X., Fishbane, N., Hollander, Z., Ng, R.T., McManus, B., Tebbutt, S.J., Miller, B.E., Rennard, S., Paré, P.D., and Sin, D.D. (2016). The Effect of Different Case Definitions of Current Smoking on the Discovery of Smoking-Related Blood Gene Expression Signatures in Chronic Obstructive Pulmonary Disease. Nicotine Tob. Res. *18*, 1903–1909. https://doi.org/10.1093/ntr/ntw129.
58. Imboden, M., Wielscher, M., Rezwan, F.I., Amaral, A.F.S., Schaffner, E., Jeong, A., Beckmeyer-Borowko, A., Harris, S.E., Starr, J.M., Deary, I.J., et al. (2019). Epigenome-wide association study of lung function level and its change. Eur. Respir. J. *54*, 1900457. https://doi.org/10.1183/13993003.00457-2019.
59. Fuemmeler, B.F., Dozmorov, M.G., Do, E.K., Zhang, J.J., Grenier, C., Huang, Z., Maguire, R.L., Kollins, S.H., Hoyo, C., and Murphy, S.K. (2021). DNA Methylation in Babies Born to Nonsmoking Mothers Exposed to Secondhand Smoke during Pregnancy: An Epigenome-Wide Association Study. Environ. Health Perspect. *129*, 057010. https://doi.org/10.1289/ehp8099.
60. Joehanes, R., Just, A.C., Marioni, R.E., Pilling, L.C., Reynolds, L.M., Mandaviya, P.R., Guan, W., Xu, T., Elks, C.E., Aslibekyan, S., et al. (2016). Epigenetic signatures of cigarette smoking. Circ. Cardiovasc. Genet. *9*, 436–447.
61. Richter, G.M., Kruppa, J., Munz, M., Wiehe, R., Häsler, R., Franke, A., Martins, O., Jockel-Schneider, Y., Bruckmann, C., Dommisch, H., and Schaefer, A.S. (2019). A combined epigenome- and transcriptome-wide association study of the oral masticatory mucosa assigns CYP1B1 a central role for epithelial health in smokers. Clin. Epigenetics *11*, 105. https://doi.org/10.1186/s13148-019-0697-y.
62. Zhang, R., Lai, L., Dong, X., He, J., You, D., Chen, C., Lin, L., Zhu, Y., Huang, H., Shen, S., et al. (2019). SIPA1L3 methylation modifies the benefit of smoking cessation on lung adenocarcinoma survival: an epigenomic–smoking interaction analysis. Mol. Oncol. *13*, 1235–1248.
63. Song, N., Sim, J.A., Dong, Q., Zheng, Y., Hou, L., Li, Z., Hsu, C.W., Pan, H., Mulder, H., Easton, J., et al. (2022). Blood DNA methylation signatures are associated with social determinants of health among survivors of childhood cancer. Epigenetics *17*, 1389–1403. https://doi.org/10.1080/15592294.2022.2030883.
64. Lee, M.K., Hong, Y., Kim, S.Y., Kim, W.J., and London, S.J. (2017). Epigenome-wide association study of chronic obstructive pulmonary disease and lung function in Koreans. Epigenomics *9*, 971–984. https://doi.org/10.2217/epi-2017-0002.
65. Cardenas, A., Ecker, S., Fadadu, R.P., Huen, K., Orozco, A., McEwen, L.M., Engelbrecht, H.R., Gladish, N., Kobor, M.S., Rosero-Bixby, L., et al. (2022). Epigenome-wide association study and epigenetic age acceleration associated with cigarette smoking among Costa Rican adults. Sci. Rep. *12*, 4277. https://doi.org/10.1038/s41598-022-08160-w.
66. Tang, D., Tao, D., Fang, Y., Deng, C., Xu, Q., and Zhou, J. (2017). TNF-Alpha Promotes Invasion and Metastasis via NF-Kappa B Pathway in Oral Squamous Cell Carcinoma. Med. Sci. Monit. Basic Res. *23*, 141–149. https://doi.org/10.12659/msmbr.903910.
67. Lowe, S.W., and Lin, A.W. (2000). Apoptosis in cancer. Carcinogenesis *21*, 485–495. https://doi.org/10.1093/carcin/21.3.485.
68. Whibley, C., Pharoah, P.D.P., and Hollstein, M. (2009). p53 polymorphisms: cancer implications. Nat. Rev. Cancer *9*, 95–107. https://doi.org/10.1038/nrc2584.
69. Johnson, D.E., O'Keefe, R.A., and Grandis, J.R. (2018). Targeting the IL-6/JAK/STAT3 signalling axis in cancer. Nat. Rev. Clin. Oncol. *15*, 234–248. https://doi.org/10.1038/nrclinonc.2018.8.
70. Oshi, M., Tokumaru, Y., Angarita, F.A., Yan, L., Matsuyama, R., Endo, I., and Takabe, K. (2020). Degree of Early Estrogen Response Predict Survival after Endocrine Therapy in Primary and Metastatic ER-Positive Breast Cancer. Cancers *12*, 3557. https://doi.org/10.3390/cancers12123557.
71. Kim, K.-H., Kim, H.J., and Lee, T.R. (2017). Epidermal long non-coding RNAs are regulated by ultraviolet irradiation. Gene *637*, 196–202.
72. Ikushima, H., and Miyazono, K. (2010). TGFβ signalling: a complex web in cancer progression. Nat. Rev. Cancer *10*, 415–424.
73. Wilson, W.R., and Hay, M.P. (2011). Targeting hypoxia in cancer therapy. Nat. Rev. Cancer *11*, 393–410. https://doi.org/10.1038/nrc3064.
74. Brahimi-Horn, M.C., Chiche, J., and Pouysségur, J. (2007). Hypoxia and cancer. J. Mol. Med. *85*, 1301–1307. https://doi.org/10.1007/s00109-007-0281-3.
75. Tian, T., Li, X., and Zhang, J. (2019). mTOR signaling in cancer and mTOR inhibitors in solid tumor targeting therapy. Int. J. Mol. Sci. *20*, 755.
76. Mok, E.H.K., and Lee, T.K.W. (2020). The Pivotal Role of the Dysregulation of Cholesterol Homeostasis in Cancer: Implications for Therapeutic Targets. Cancers *12*, 1410. https://doi.org/10.3390/cancers12061410.
77. Hwang, J.-W., Sundar, I.K., Yao, H., Sellix, M.T., and Rahman, I. (2014). Circadian clock function is disrupted by environmental tobacco/cigarette smoke, leading to lung inflammation and injury via a SIRT1-BMAL1 pathway. Faseb. J. *28*, 176–194.
78. Artese, A., Stamford, B.A., and Moffatt, R.J. (2019). Cigarette smoking: an accessory to the development of insulin resistance. Am. J. Lifestyle Med. *13*, 602–605.
79. Aschard, H., Hancock, D.B., London, S.J., and Kraft, P. (2011). Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. Human Heredity, 292–300.
80. Tindle, H.A., Stevenson Duncan, M., Greevy, R.A., Vasan, R.S., Kundu, S., Massion, P.P., and Freiberg, M.S. (2018). Lifetime Smoking History and Risk of Lung Cancer: Results From the Framingham Heart Study. J. Nat. Can. Inst. *110*, 1201–1207. https://doi.org/10.1093/jnci/djy041.
81. Duncan, M.S., Freiberg, M.S., Greevy, R.A., Jr., Kundu, S., Vasan, R.S., and Tindle, H.A. (2019). Association of Smoking Cessation With Subsequent Risk of Cardiovascular Disease. JAMA *322*, 642–650. https://doi.org/10.1001/jama.2019.10298.
82. Bakulski, K.M., Dou, J., Lin, N., London, S.J., and Colacino, J.A. (2019). DNA methylation signature of smoking in lung cancer is enriched for exposure signatures in newborn and adult blood. Sci. Rep. *9*, 4576. https://doi.org/10.1038/s41598-019-40963-2.

83. Chow, W.-H., Dong, L.M., and Devesa, S.S. (2010). Epidemiology and risk factors for kidney cancer. Nat. Rev. Urol. *7*, 245–257.
84. Granata, F., Frattini, A., Loffredo, S., Del Prete, A., Sozzani, S., Marone, G., and Triggiani, M. (2006). Signaling events involved in cytokine and chemokine production induced by secretory phospholipase A2 in human lung macrophages. Eur. J. Immunol. *36*, 1938–1950. https://doi.org/10.1002/eji.200535567.
85. Saiga, A., Uozumi, N., Ono, T., Seno, K., Ishimoto, Y., Arita, H., Shimizu, T., and Hanasaki, K. (2005). Group X secretory phospholipase A2 can induce arachidonic acid release and eicosanoid production without activation of cytosolic phospholipase A2 alpha. Prostaglandins Other Lipid Mediat. *75*, 79–89.
86. Serhan, C.N. (2007). Resolution Phase of Inflammation: Novel Endogenous Anti-Inflammatory and Proresolving Lipid Mediators and Pathways. Annu. Rev. Immunol. *25*, 101–137. https://doi.org/10.1146/annurev.immunol.25.022106.141647.
87. Slaymi, C., Vignal, E., Crès, G., Roux, P., Blangy, A., Raynaud, P., and Fort, P. (2019). The atypical RhoU/Wrch1 Rho GTPase controls cell proliferation and apoptosis in the gut epithelium. Biol. Cell *111*, 121–141. https://doi.org/10.1111/boc.201800062.
88. Habits, P., and Combustions, I. (2012). France: IARC Monographs on the Evaluation of Carcinogenic Risks to Humans.
89. Spitz, M.R., Gorlov, I.P., Amos, C.I., Dong, Q., Chen, W., Etzel, C.J., Gorlova, O.Y., Chang, D.W., Pu, X., Zhang, D., et al. (2011). Variants in Inflammation Genes Are Implicated in Risk of Lung Cancer in Never Smokers Exposed to Second-hand Smoke. Cancer Discov. *1*, 420–429. https://doi.org/10.1158/2159-8290.cd-11-0080.
90. Du, W.-Q., Zhu, Z.M., Jiang, X., Kang, M.J., and Pei, D.S. (2023). COPS6 promotes tumor progression and reduces CD8+ T cell infiltration by repressing IL-6 production to facilitate tumor immune evasion in breast cancer. Acta Pharmacol. Sin. *44*, 1890–1905. https://doi.org/10.1038/s41401-023-01085-8.
91. Mahajan, S., Majumder, A., Stewart, P.A., Chen, Y.A., Adhikari, E., Fang, B., Yang, Y., Lawrence, H., Kinose, F., Koomen, J.M., and Haura, E.B. (2022). Deubiquitinase Vulnerabilities Identified through Activity-Based Protein Profiling in Non-Small Cell Lung Cancer. ACS Chem. Biol. *17*, 776–784. https://doi.org/10.1021/acschembio.2c00018.
92. Kheir, W.A., Gevrey, J.C., Yamaguchi, H., Isaac, B., and Cox, D. (2005). A WAVE2-Abi1 complex mediates CSF-1-induced F-actin-rich membrane protrusions and migration in macrophages. J. Cell Sci. *118*, 5369–5379. https://doi.org/10.1242/jcs.02638.
93. Cui, H., Liu, Y., Zheng, Y., Li, H., Zhang, M., Wang, X., Zhao, X., Cheng, H., Xu, J., Chen, X., and Ding, Z. (2023). Intelectin enhances the phagocytosis of macrophages via CDC42-WASF2-ARPC2 signaling axis in Megalobrama amblycephala. Int. J. Biol. Macromol. *236*, 124027.
94. Semba, S., Iwaya, K., Matsubayashi, J., Serizawa, H., Kataba, H., Hirano, T., Kato, H., Matsuoka, T., and Mukai, K. (2006). Coexpression of actin-related protein 2 and Wiskott-Aldrich syndrome family verproline-homologous protein 2 in adenocarcinoma of the lung. Clin. Cancer Res. *12*, 2449–2454.

The American Journal of Human Genetics, Volume *111*

Supplemental information

The association of cigarette smoking with DNA methylation and gene expression in human tissue samples

James L. Li, Niyati Jain, Lizeth I. Tamayo, Lin Tong, Farzana Jasmine, Muhammad G. Kibriya, Kathryn Demanelis, Meritxell Oliva, Lin S. Chen, and Brandon L. Pierce

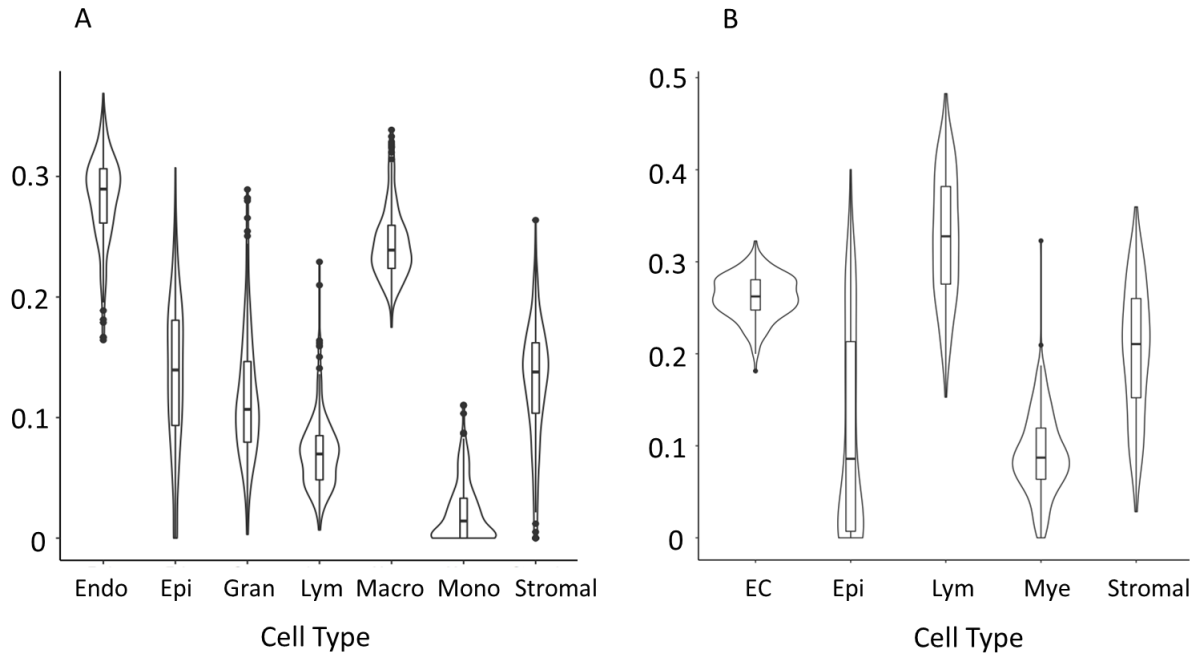

Figure S1. Distribution of cell type percentages
(A) Estimated cell type percentages for lung using the EPISCORE method with the pan-tissue DNAm atlas as a reference dataset (B) Estimated cell type percentages for colon using the EPISCORE method with the pan-tissue DNAm atlas as a reference dataset.

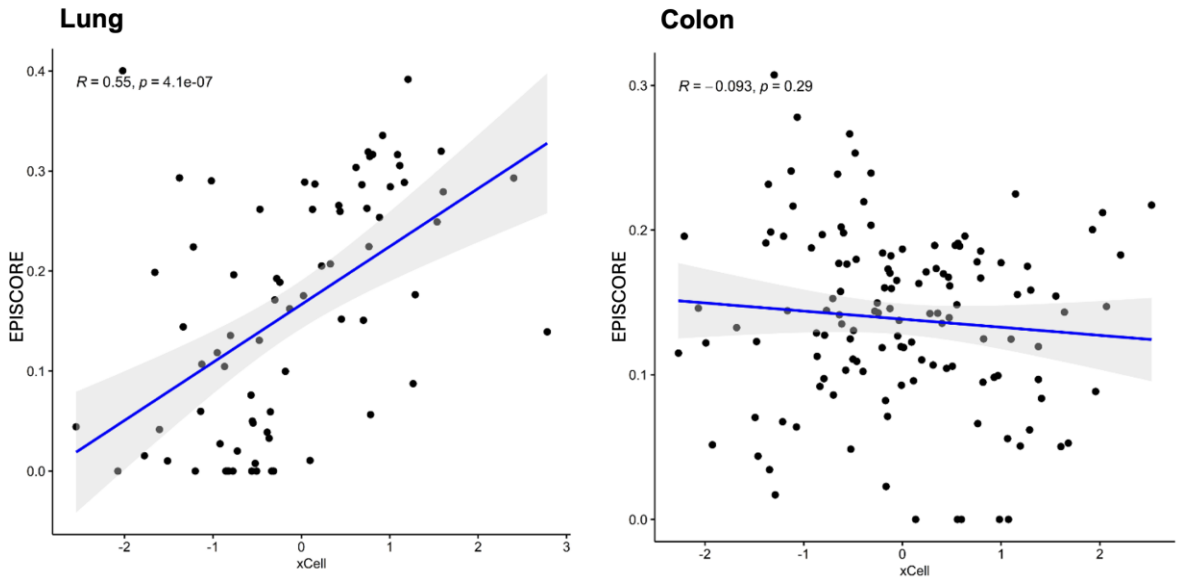

Figure S2. Correlation between EPISCORE and xCell cell type proportions
(A) Spearman correlation between EPISCORE and xCell computed epithelial cell proportions for lung. (B) Spearman correlation between EPISCORE and xCell computed epithelial cell proportions for colon.

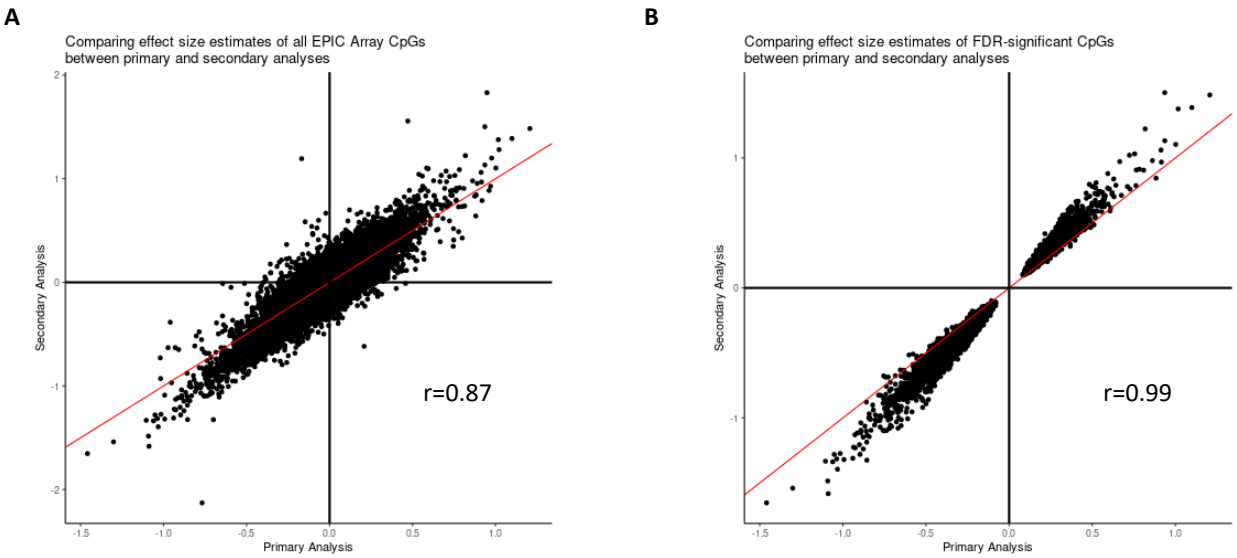

Figure S3. Comparison of effect size estimates between the primary EWAS and secondary EWAS analysis

Scatterplot comparing effect size estimates of each CpG between the EWAS of ever vs. never smokers (primary analysis) and current vs. never smokers (secondary analysis) in lung tissue for A) All CpGs on the EPIC Array and B) CpGs that were significantly associated with smoking at an FDR of 0.05.

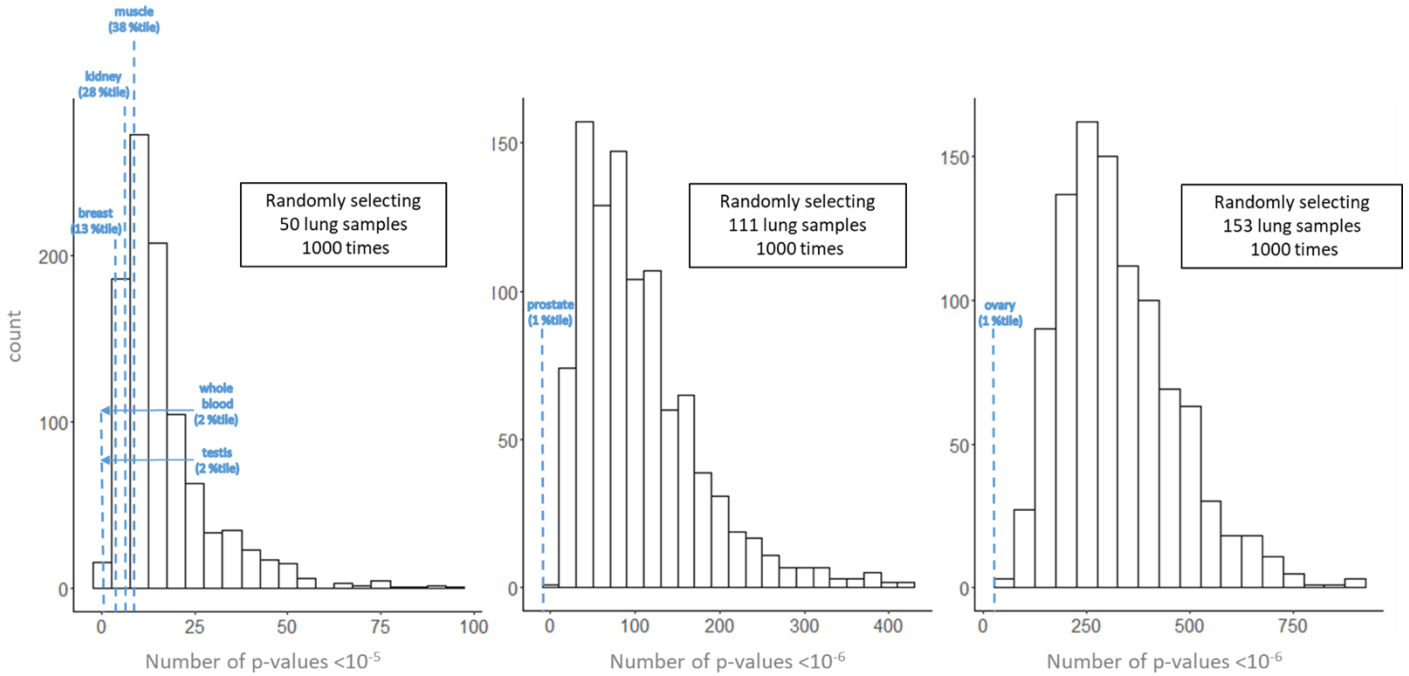

Figure S4. Lung shows more prominent effects of smoking than other tissue types. Number of smoking-associated CpGs in lung compared to the number in other tissues that pass p-value thresholds after down sampling to sample sizes of n=50, n=111, and n=153. Each distribution shown is generated by randomly selecting samples from the 212 lung samples (and conducting EWAS analyses) 1000 times.

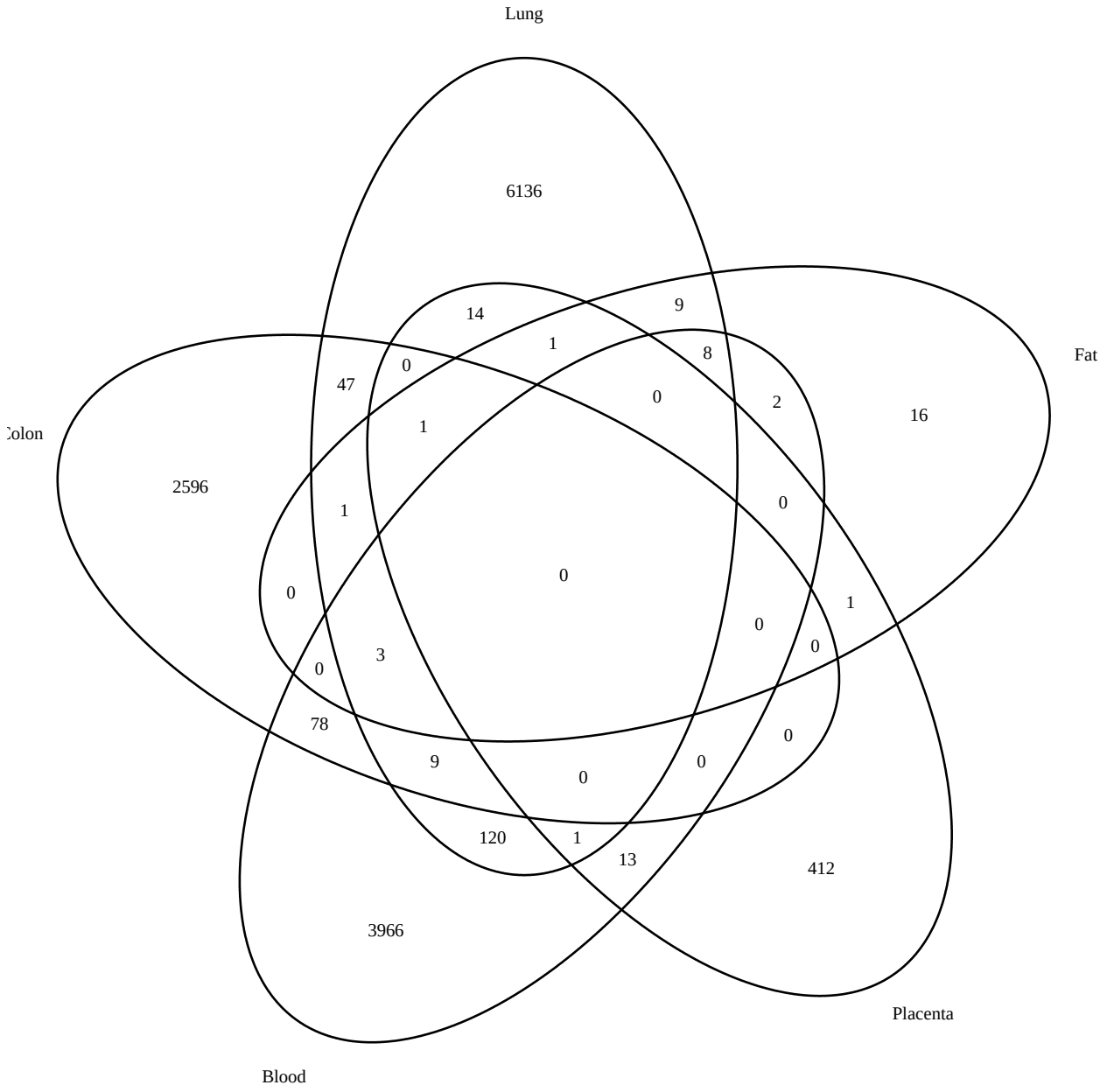

Figure S5. Venn diagram of the number of CpGs previously identified in EWAS conducted in different tissue types

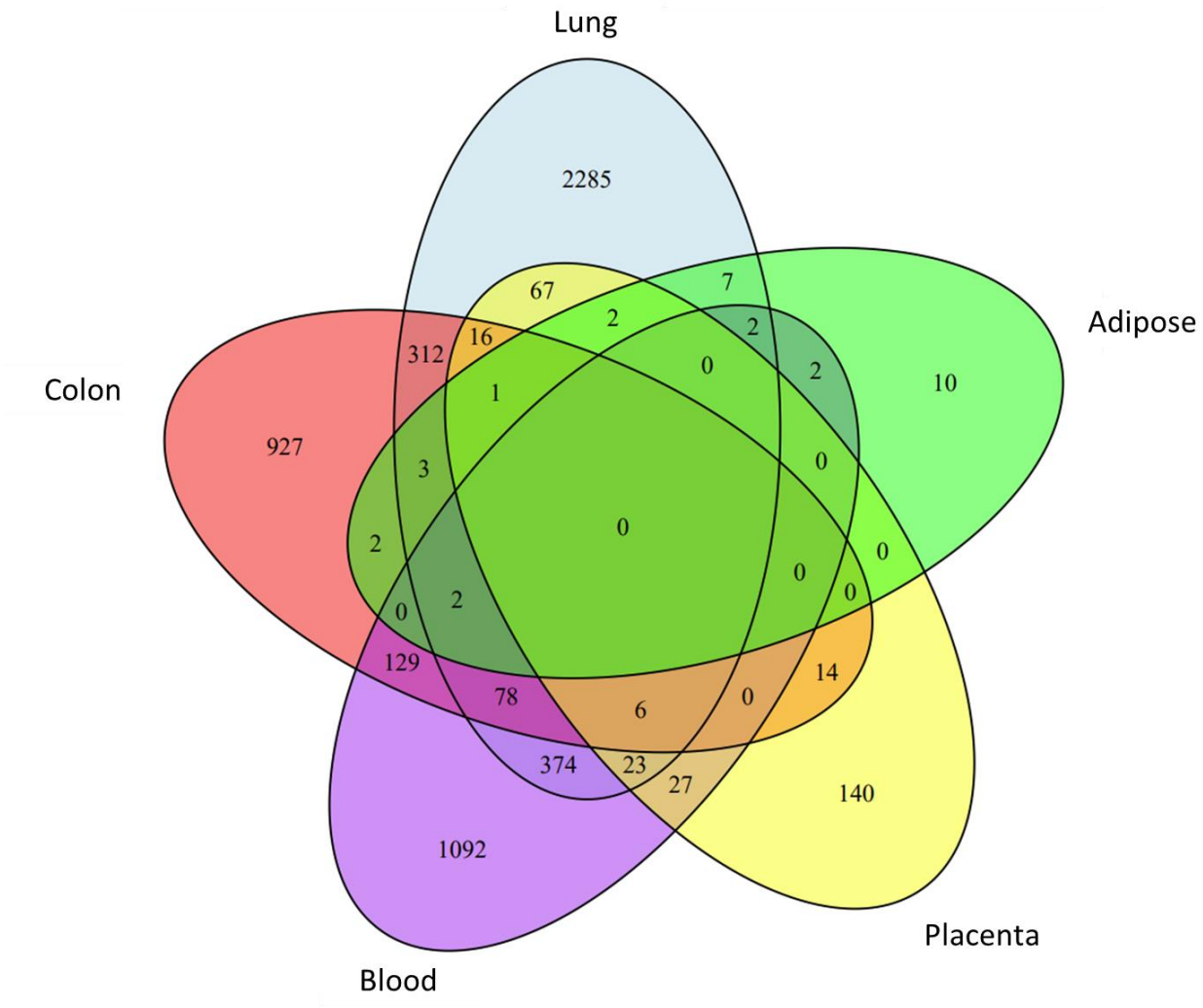

Figure S6. Venn diagram of the genes annotated to smoking-associated CpGs previously identified in EWAS conducted in different tissue types

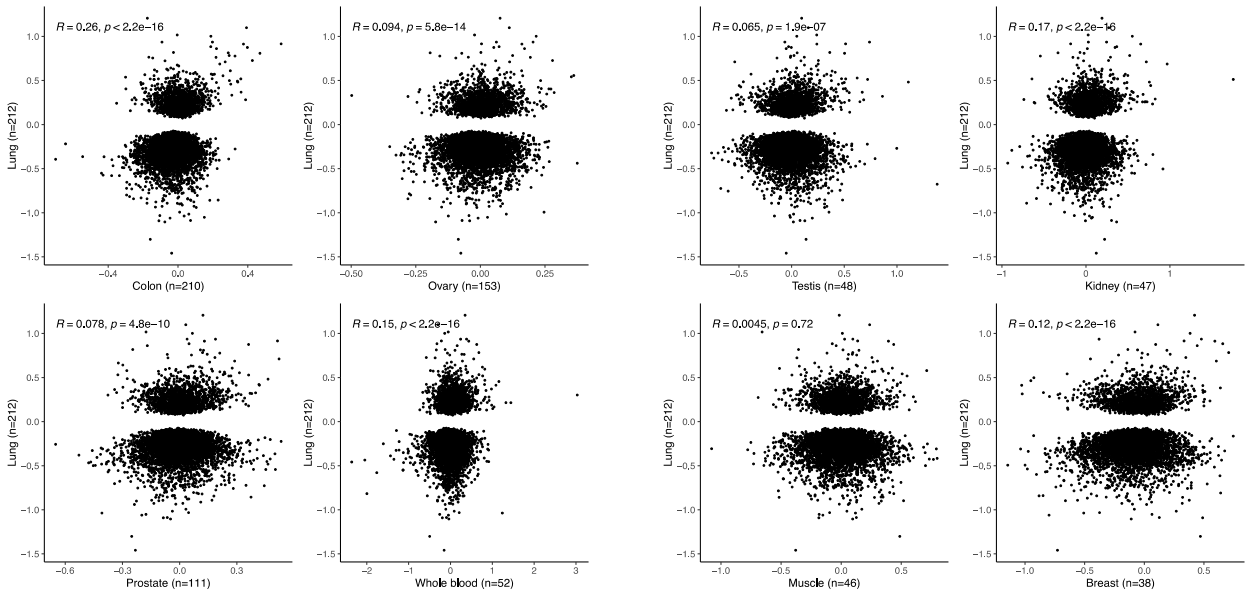

Figure S7. Comparison of effect size estimates between pairs of tissue types

Scatterplots showing the correlations between association estimates observed across pairs of tissue types. Smoking-associated CpGs that pass FDR 0.05 in lung are shown on the vertical axes, with all other tissues shown on the horizontal axes.

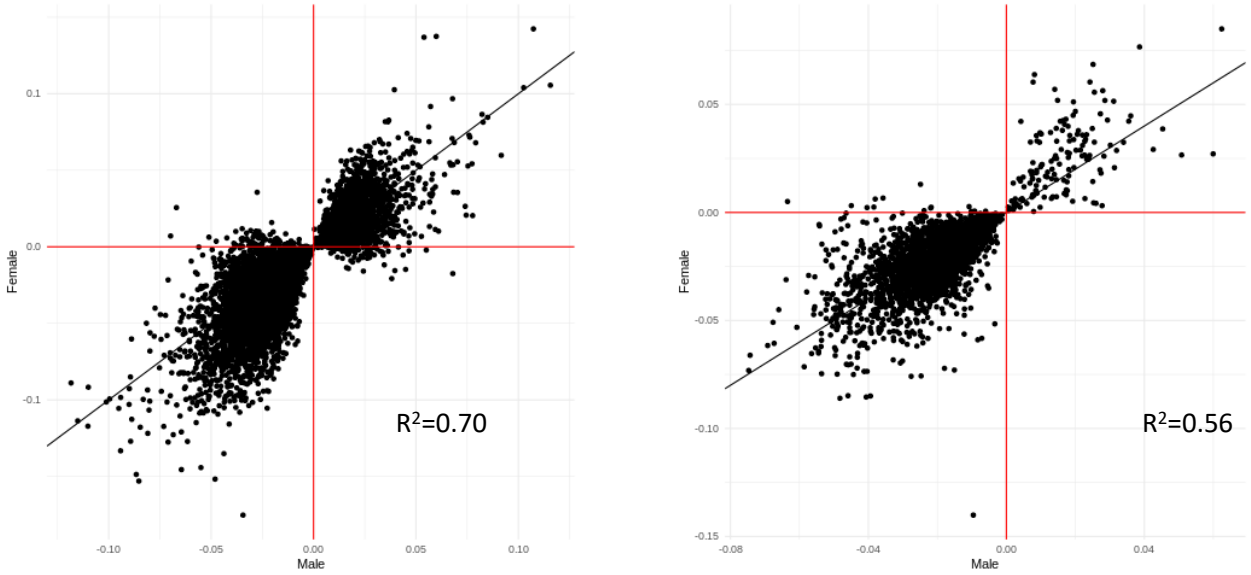

Figure S8. Comparison of effect size estimates between females and males

Scatterplot comparing effect size estimates of each CpG when stratifying by sex for all CpGs that were significantly associated with smoking at an FDR of 0.05 in lung (left) and in colon (right). Black line represents the identity line.

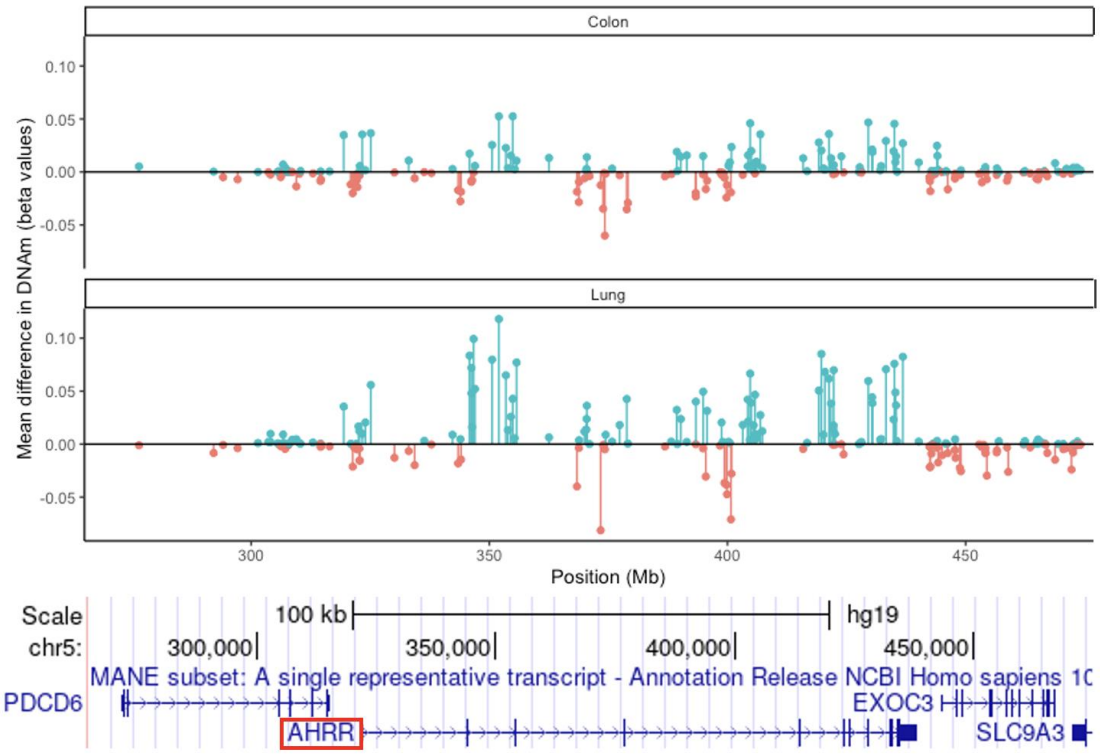

Figure S9. Difference in DNAm beta values between smokers and non-smokers across CpGs measured around the AHRR gene.

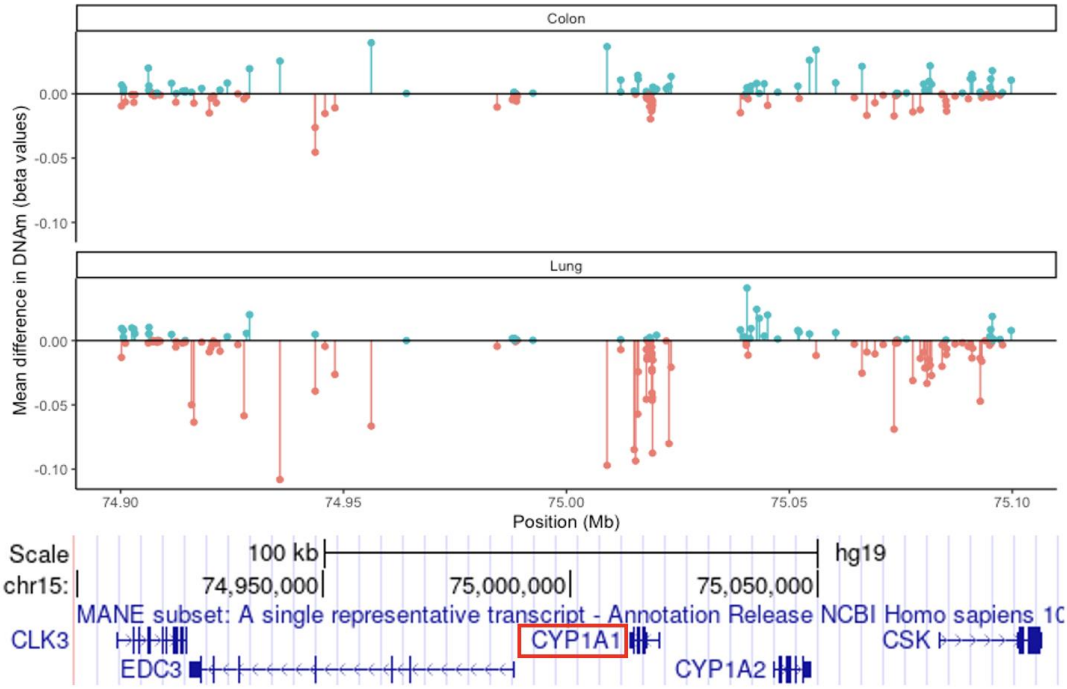

Figure S10. Difference in DNAm beta values between smokers and non-smokers across CpGs measured around the CYP1A1 gene.

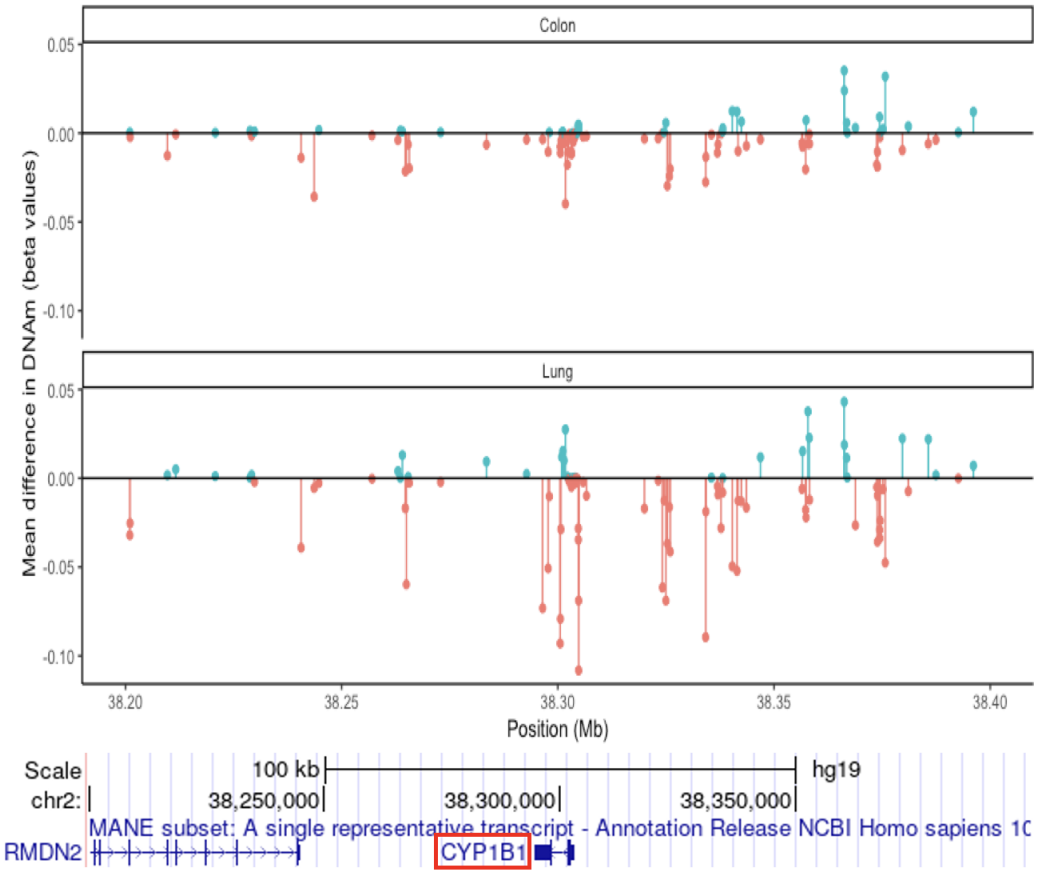

Figure S11. Difference in DNAm beta values between smokers and non-smokers across CpGs measured around the CYP1B1 gene.

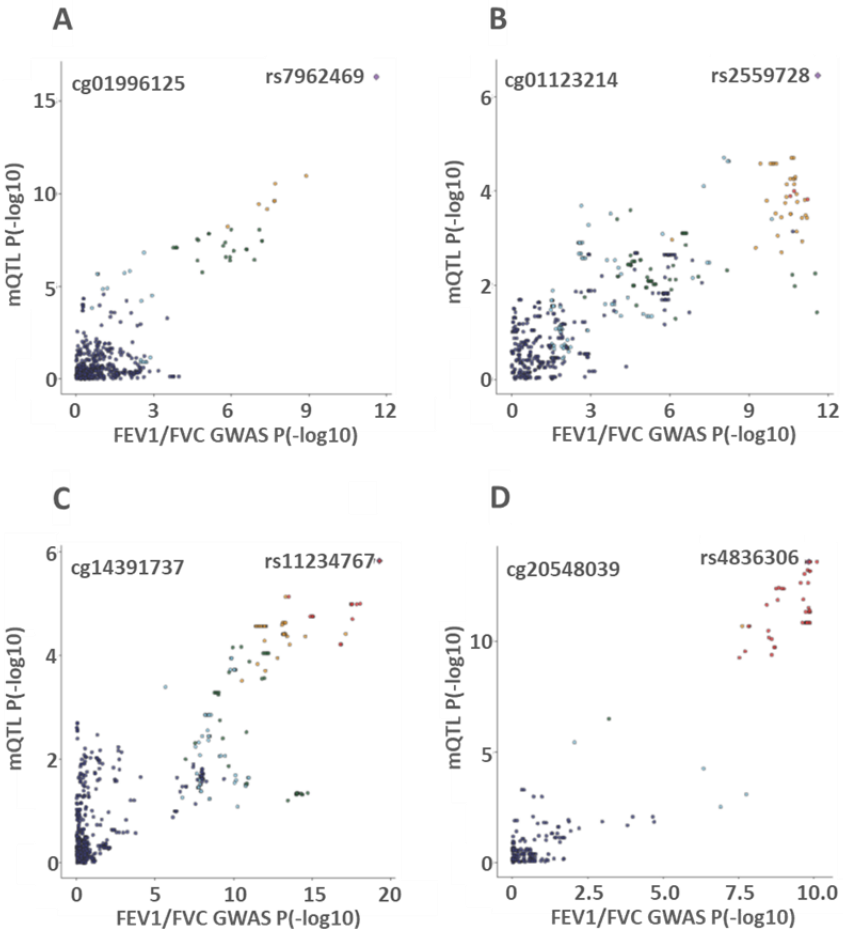

Figure S12. Comparison of P-values for mQTLs vs FEV1/FVC GWAS signals
Scatter plot of P-values for mQTL signals vs. FEV1/FVC GWAS signals for four loci with mQTL/GWAS co-localization including (A) ACVR1B, (B) SFTPA1, (C) PRSS23, and (D) MARCHF3/MARCH3

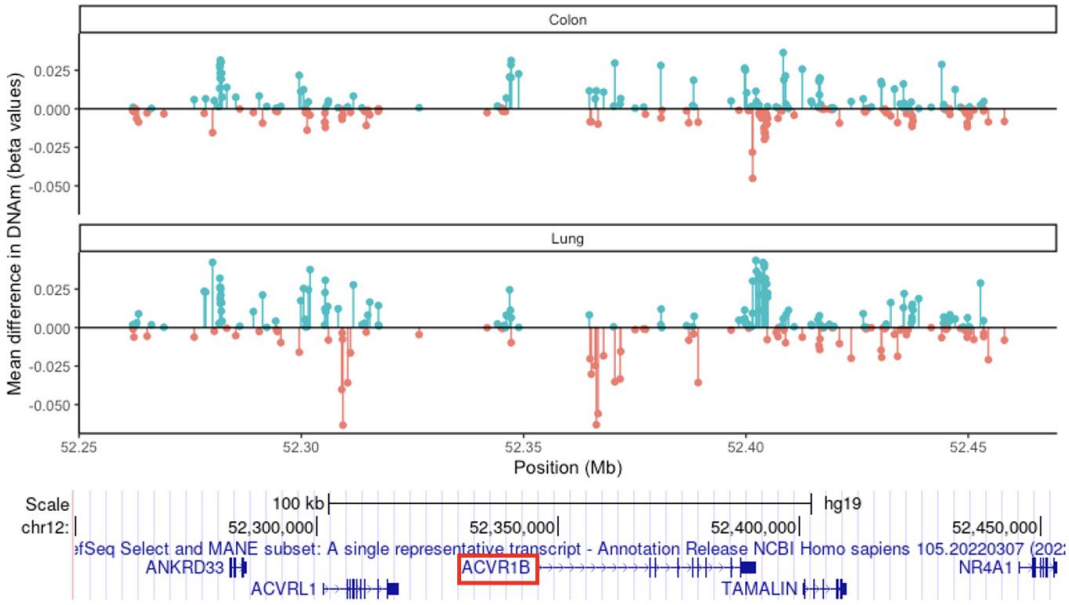

Figure S13. Difference in DNAm beta values between smokers and non-smokers across CpGs measured around the ACVR1B gene.

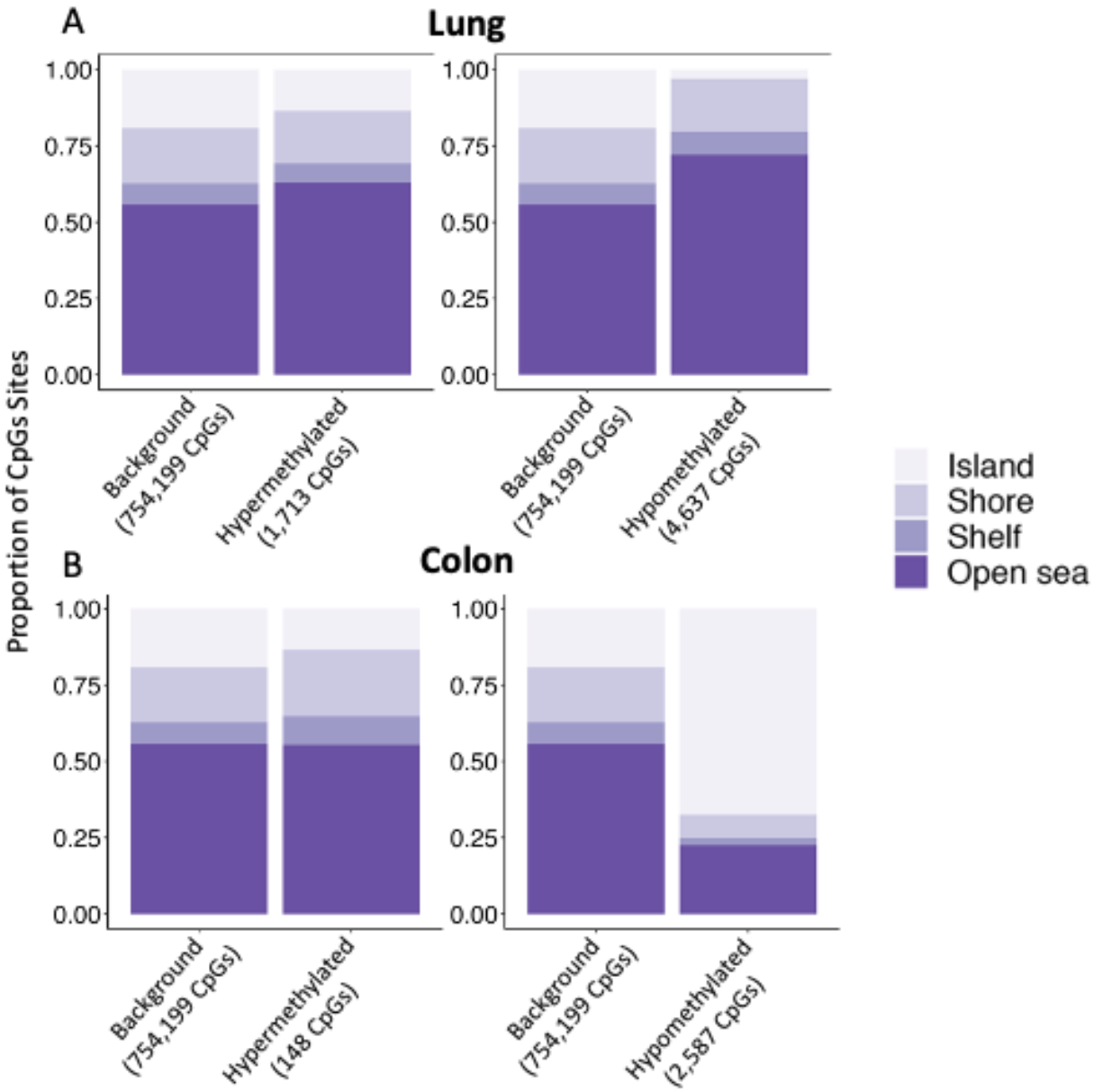

Figure S14. Enrichment of smoking-related CpGs in relation to CpG Island status

Locational distribution of significant smoking-related CpGs sites (FDR < 0.05) in relation to CpG islands in (A) lung tissue (B) colon tissue. Colors represent location of CpG site. Background: All CpGs assayed in the Infinium MethylationEPIC array included in our analyses.

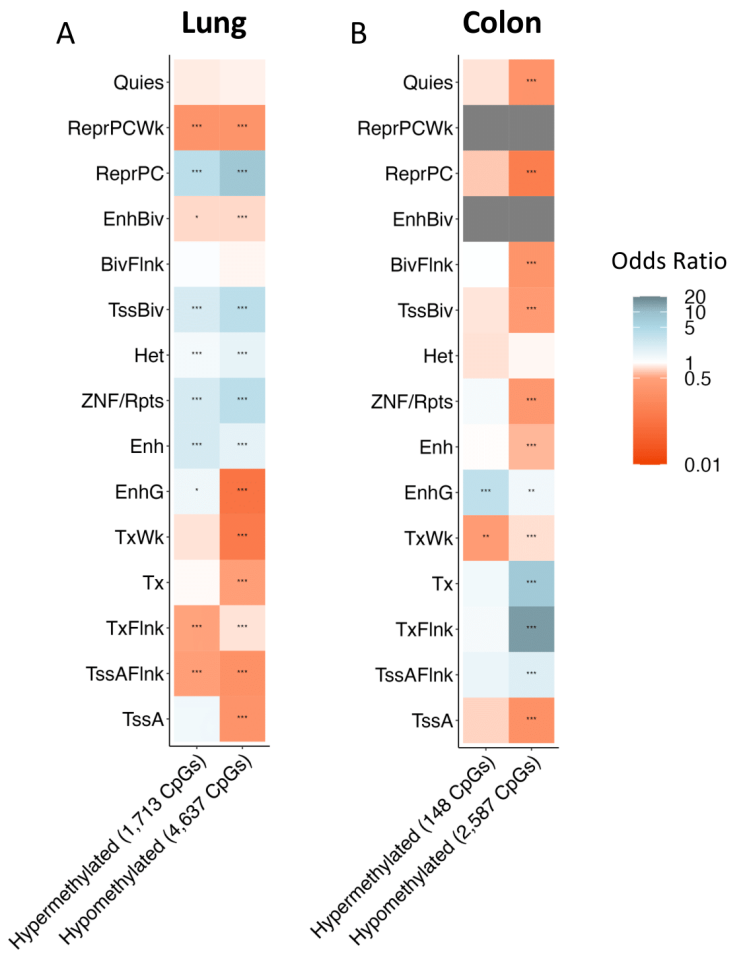

Figure S15. Enrichment of smoking-related CpG sites among chromatin segmentation features

Enrichment of smoking-related CpG sites (FDR <0.05) expressed as odd ratios in (A) lung tissue (B) colon tissue. Fisher's exact P value * < 0.05, **<0.01, ***<0.001. Active chromatin states: active transcription start site (*TssA*), flanking active TSS (*TssAFlnk*), transcription at gene 5′ and 3′ showing both promoter and enhancer (*TxFlnk*), strong transcription (Tx), weak transcription (*TxWk*), genic enhancers (*EnhG*), enhancers (*Enh*) , zinc finger protein genes and repeats (*ZNF/Rpts* ZNF*).* Inactive chromatin states: heterochromatin (*Het*), bivalent/poised TSS (*TssBiv*), flanking bivalent TSS/Enh (*BivFlnk*), bivalent enhancer (*EnhBiv*), repressed polycomb (*ReprPC*), weak repressed polycomb (*ReprPCWk*), quiescent/low (*Quies*).

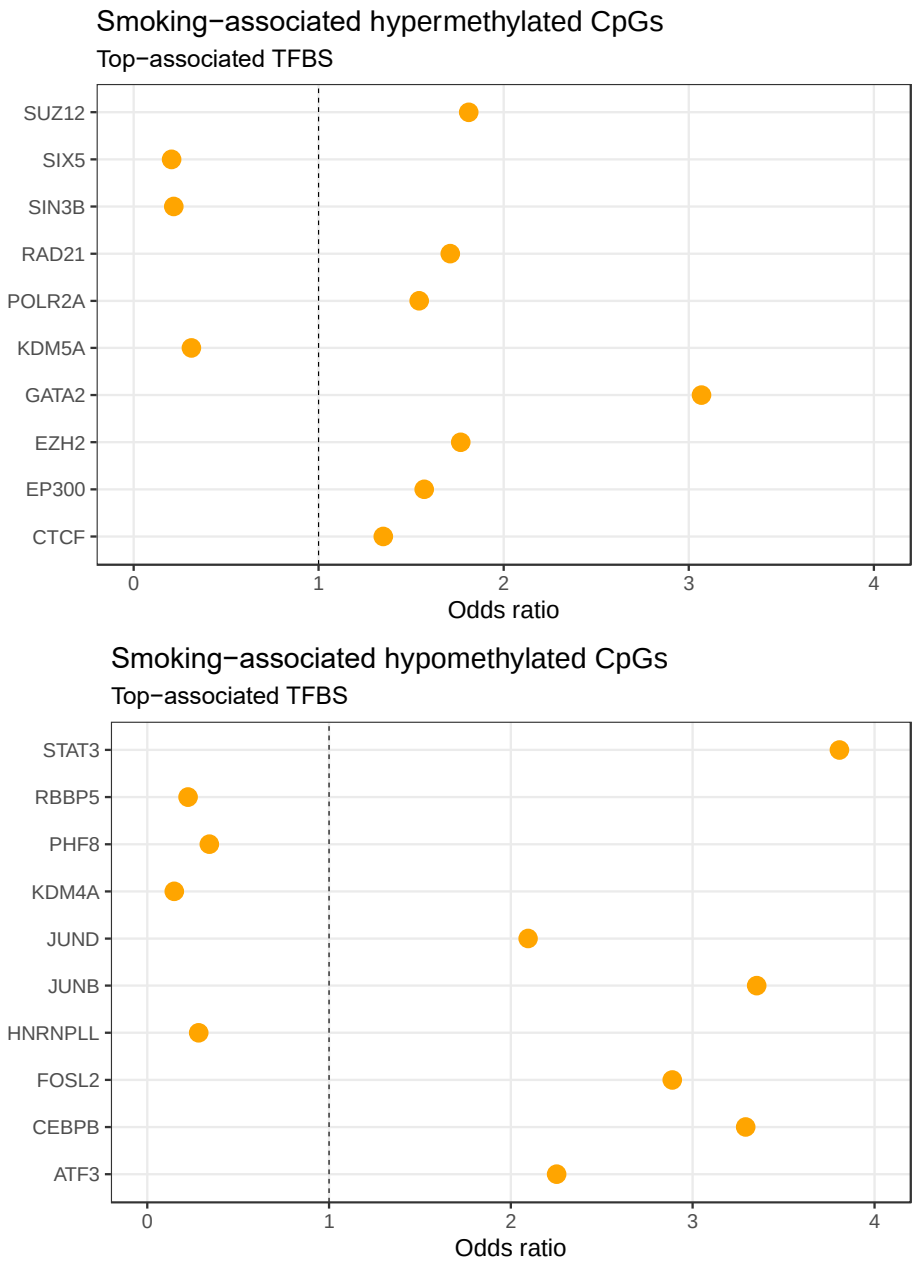

Figure S16. Enrichment of lung smoking-associated CpGs in transcription factor binding sites.

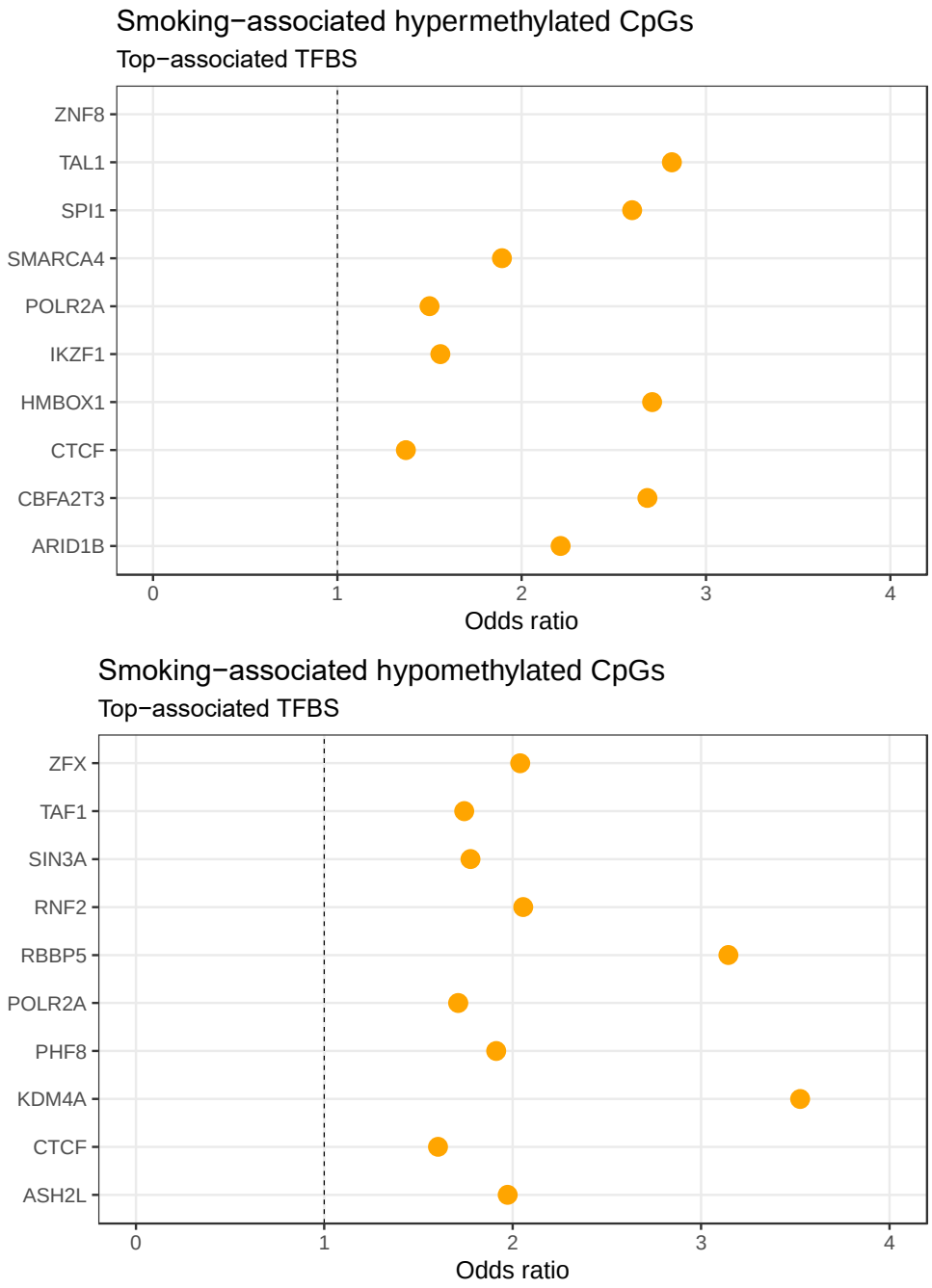

Figure S17. Enrichment of colon smoking-associated CpGs in transcription factor binding sites.

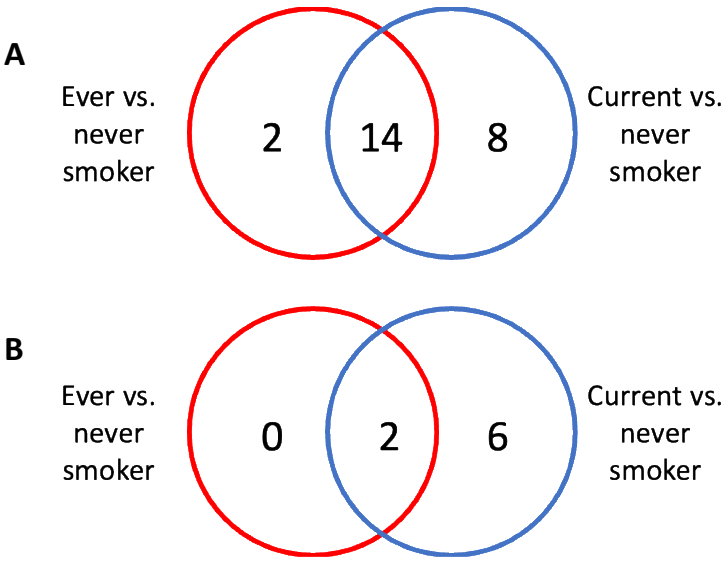

Figure S18. Comparison of the number of Hallmark gene sets and KEGG pathways

Venn diagrams comparing the number of A) Hallmark gene sets and B) KEGG pathways identified in the primary vs. secondary EWAS analysis in lung tissue.

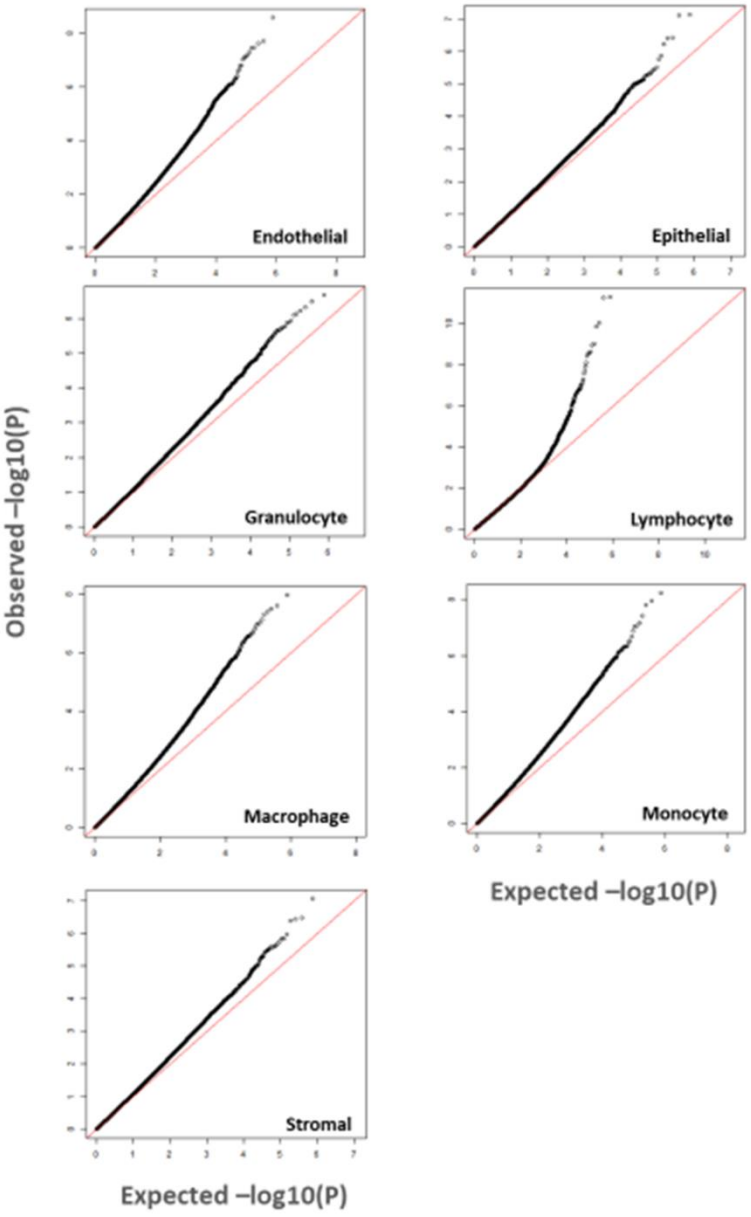

Figure S19. Q-Q plots for smoking by cell-type interactions in lung

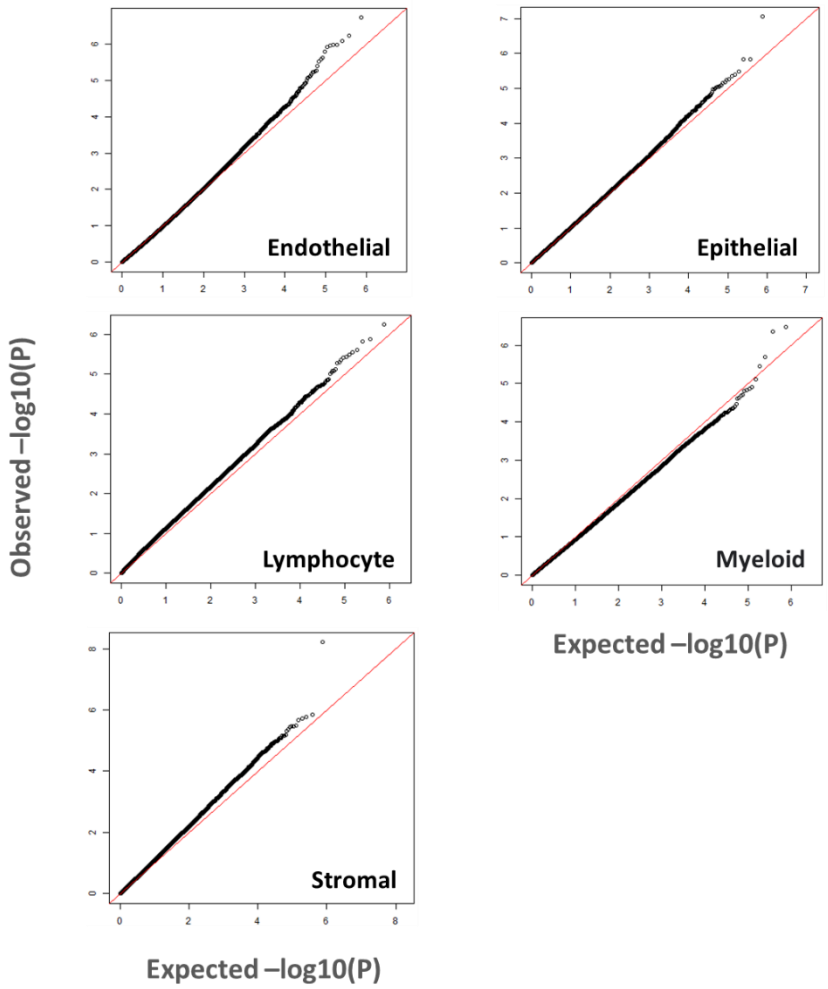

Figure S20. Q-Q plots for smoking by cell-type interactions in colon

Table S1. Correlations between EPISCORE estimates of cell type proportions with DNAm-derived SVs in lung and colon. Endo: endothelial, Epi: epithelial, Gran: granulocytes, Lym: lymphocytes, Macro: macrophages, Mono: monocytes, Stromal: stromal cells, EC: enteroendocrine cells, Mye: myeloid cells

Lung

Pearson Correlation Coefficients and P-values, n = 212										
Cell Types	**SV1**	**SV2**	**SV3**	**SV4**	**SV5**	**SV6**	**SV7**	**SV8**	**SV9**	**SV10**
Endo	-0.28	-0.4	0.47	-0.48	-0.15	-0.19	0.31	-0.14	0.003	0.12
	<.0001	<.0001	<.0001	<.0001	0.03	0.006	<.0001	0.05	0.97	0.09
Epi	-0.49	-0.65	-0.37	0.15	0.28	0.02	0.02	-0.06	0.05	-0.08
	<.0001	<.0001	<.0001	0.03	<.0001	0.73	0.73	0.41	0.44	0.25
Gran	0.59	0.65	0.27	0.02	-0.05	0.003	0.03	0.001	-0.11	-0.03
	<.0001	<.0001	<.0001	0.72	0.46	0.96	0.65	0.99	0.1	0.68
Lym	0.17	0.58	-0.34	0.28	-0.42	-0.14	0.11	0.04	-0.09	0.22
	0.01	<.0001	<.0001	<.0001	<.0001	0.05	0.11	0.59	0.21	0.001
Macro	0.74	0.48	-0.04	0.01	0.2	0.06	0.02	0.11	0.01	0.03
	<.0001	<.0001	0.59	0.94	0	0.39	0.75	0.12	0.88	0.66
Mono	0.62	0.62	0.01	0.07	0.1	0.02	-0.13	-0.04	-0.02	0.06
	<.0001	<.0001	0.9	0.29	0.14	0.75	0.06	0.56	0.78	0.37
Stromal	-0.67	-0.55	0.05	-0.07	-0.1	0.14	-0.31	0.1	0.11	-0.14
	<.0001	<.0001	0.48	0.3	0.17	0.04	<.0001	0.13	0.12	0.04

Colon

Pearson Correlation Coefficients and P-values, n = 209										
Cell Types	**SV1**	**SV2**	**SV3**	**SV4**	**SV5**	**SV6**	**SV7**	**SV8**	**SV9**	**SV10**
EC	-0.12	-0.59	0.13	-0.04	-0.09	-0.08	-0.23	0.19	-0.14	0.12
	0.08	<.0001	0.06	0.52	0.21	0.23	0.0009	0.01	0.05	0.08
Epi	<.0001	-0.6	0.14	-0.04	-0.02	-0.01	0.05	-0.005	-0.07	-0.06
	<.0001	<.0001	0.05	0.61	0.73	0.93	0.47	0.95	0.34	0.42
Lym	-0.35	0.69	-0.18	0.15	0.11	0.16	0.19	0.03	0.07	-0.15
	<.0001	<.0001	0.01	0.03	0.1	0.02	0.01	0.7	0.34	0.04
Mye	0.17	0.76	-0.18	0.14	0.23	0.26	-0.09	-0.1	0.15	0.13
	0.01	<.0001	0.01	0.05	0.0009	0.0001	0.21	0.16	0.03	0.06
Stromal	-0.86	-0.02	0.04	-0.16	-0.18	-0.28	-0.13	-0.02	-0.01	0.11
	<.0001	0.75	0.61	0.02	0.01	<.0001	0.07	0.79	0.89	0.12

Table S2a. Number of smoking-associated CpG Sites in the primary analysis of ever vs. never smokers detected in each tissue type (and the tissue-specific P-value threshold) based on false-discovery rates (FDR) of 0.01 and 0.05.

Tissue	FDR adjusted P-value threshold	
	0.01	0.05
lung (n=212)	2,478 (3.3e-5)	6,350 (0.0004)
colon (n=209)	662 (8.8e-6)	2,735 (0.0001)
ovary (n=153)	0 (N/A)	0 (N/A)
prostate (n=111)	0 (N/A)	0 (N/A)
whole blood (n=52)	0 (N/A)	0 (N/A)
breast (n=38)	0 (N/A)	0 (N/A)
testis (n=48)	0 (N/A)	0 (N/A)
kidney (n=47)	0 (N/A)	0 (N/A)
muscle (n=46)	0 (N/A)	0 (N/A)

Table S2b. Number of smoking-associated CpG Sites in the secondary analysis of current vs. never smokers.

Tissue	Smoking-associated CpG sites (Tissue-specific P-value threshold)	
	FDR 0.01	FDR 0.05
lung (n=151)	4,589 (6.11e-5)	10,495 (0.0007)
colon (n=148)	923 (1.22e-5)	4,797 (0.0003)

Table S3. Power to detect the effect sizes of smoking-associated CpGs observed in lung at different sample sizes. *Footnote*: cg01584760, cg20291548, cg09138315 are the smoking-associated CpGs with the maximum, median, and minimum effect sizes, respectively.

CpG	randomly selected n of samples		
	n=150	n=100	n=50
cg01584760	982	349	0
cg20291548	54	4	0
cg09138315	50	1	0

Table S4. Enrichment of smoking-associated CpGs based on prior studies of blood samples,[1] adipose samples,[2] placenta samples,[3] or reported in lung tissue within the current study (p-values computed through a one-sided two-proportion z-test). Red highlighted cells indicate p-values less than 0.05.

	Blood CpGs (p-value threshold: 1E-3)	Lung CpGs (p-value threshold: 1E-3)	Adipose CpGs (p-value threshold: 1E-3)	Placenta CpGs (p-value threshold: 1E-3)	Blood CpGs (p-value threshold: 1E-5)	Lung CpGs (p-value threshold: 1E-5)	Adipose CpGs (p-value threshold: 1E-5)	Placenta CpGs (p-value threshold: 1E-5)
Breast	3.30E-02	3.49E-08	1.38E-02	8.08E-02	5.00E-01	5.00E-01	5.00E-01	5.00E-01
Colon	3.09E-47	5.79E-43	4.35E-23	5.00E-01	1.23E-30	1.34E-06	9.10E-17	4.17E-01
Kidney	2.83E-02	1.92E-05	2.29E-33	5.00E-01	5.00E-01	5.00E-01	5.00E-01	5.00E-01
Muscle	5.00E-01	3.03E-01	1.06E-12	5.00E-01	5.00E-01	5.00E-01	5.00E-01	5.00E-01
Ovary	7.22E-05	4.60E-09	5.00E-01	4.26E-01	5.00E-01	5.00E-01	5.00E-01	5.00E-01
Prostate	4.65E-01	2.57E-02	1.43E-09	5.00E-01	5.00E-01	1.83E-01	5.00E-01	5.00E-01
Testis	5.00E-01	5.82E-02	5.00E-01	5.00E-01	5.00E-01	5.00E-01	5.00E-01	5.00E-01
Whole Blood	8.16E-46	2.22E-08	3.00E-196	5.00E-01	5.00E-01	5.00E-01	5.00E-01	5.00E-01
Lung	7.46E-116	0.00E+00	4.44E-265	1.48E-10	1.43E-91	0.00E+00	0.00E+00	8.14E-07

Table S5. Genes with associations between smoking and both DNAm and gene expression data at an FDR<0.05 in lung. (table in separate file)

Table S6. Genes with associations between smoking and both DNAm and gene expression data at an FDR<0.05 in colon. (table in separate file)

Table S7. Colocalization between smoking-associated CpGs in lung (FDR<0.01), mQTLs, and the 10 SNPs reaching genome-wide significance in the UK Biobank FEV1/FVC GWAS. (table in separate file)

Table S8. Colocalization between smoking-associated CpGs in colon (FDR<0.05), mQTLs, and genome-wide significant SNPs identified in genome-wide association studies of colon-related diseases. (table in separate file)

Table S9. Top hallmark gene sets detected in the primary analysis of hypomethylated smoking-associated CpGs in lung.

Description	Genes in gene set	Genes with smoking-associated CpGs*	Enrichment P	FDR-adjusted p-value
Hallmark Gene Sets				
TNF-alpha signaling via NFKb	199	49	5.01E-07	2.50E-05
P53 pathway	196	46	5.01E-05	1.25E-03
Apoptosis	155	38	1.58E-04	2.64E-03
Hypoxia	190	45	4.68E-04	5.85E-03
IL6-JAK-STAT3 signaling	81	20	9.44E-04	9.44E-03
Early response to estrogen	194	49	1.75E-03	1.35E-02
IL2-STAT5 signaling	194	44	1.90E-03	1.35E-02
MTORC1 signaling	194	38	3.02E-03	1.89E-02
Cholesterol homeostasis	71	18	3.59E-03	1.99E-02
TGF-beta signaling	53	17	6.18E-03	2.87E-02
PI3K-AKT-MTOR signaling	103	26	6.31E-03	2.87E-02
Androgen response	97	25	8.80E-03	3.41E-02
Xenobiotic metabolism	197	36	8.86E-03	3.41E-02
Genes down-regulated in response to ultraviolet (UV) radiation	142	40	1.00E-02	3.58E-02

*Genes with CpGs (as assigned by Illumina) that are associated with smoking

Table S10. Top hallmark gene sets detected in the pathway analysis of smoking-associated CpGs in colon tissue.

Hallmark gene sets	Genes in gene set	Genes with smoking-associated CpGs [1]	Enrichment P	FDR
Epithelial to mesenchymal transition	192	43	9.59e-8	4.80e-6
UV response DN	142	30	2.12e-3	0.05

[1] Genes with CpGs (as assigned by Illumina) that are associated with smoking

Table S11. Top hallmark gene sets and KEGG pathways detected in the secondary analysis of smoking-associated CpGs in lung.

Description	Genes in gene set	Genes with smoking-associated CpGs*	Enrichment P	FDR-adjusted p-value
Hallmark Gene Sets				
TNF-alpha signaling via NFKb	199	76	5.91E-07	2.96E-05
P53 pathway	196	75	1.42E-05	3.54E-04
IL2-STAT5 signaling	194	77	9.89E-05	1.65E-03
Early response to estrogen	194	82	3.07E-04	3.84E-03
Apoptosis	155	58	5.66E-04	5.57E-03
Complement	194	69	6.68E-04	5.57E-03
Allograft rejection	191	63	8.56E-04	5.98E-03
IL6-JAK-STAT3 signaling	81	31	1.06E-03	5.98E-03
Xenobiotic metabolism	197	65	1.08E-03	5.98E-03
Inflammatory response	196	62	1.30E-03	6.50E-03
Late response to estrogen	194	72	1.52E-03	6.89E-03
Adipogenesis	196	65	1.74E-03	7.04E-03
TGF-beta signaling	53	27	1.92E-03	7.04E-03
Cholesterol homeostasis	71	29	1.97E-03	7.04E-03
Genes up-regulated by KRAS activation	192	66	3.01E-03	1.00E-02
MTORC1 signaling	194	63	3.85E-03	1.20E-02
Genes down-regulated in response to ultraviolet (UV) radiation	142	65	4.41E-03	1.30E-02
Fatty acid metabolism	149	43	9.35E-03	2.60E-02
Hypoxia	190	65	1.53E-02	4.02E-02
Interferon gamma response	200	58	1.67E-02	4.16E-02
Apical junction	193	70	1.82E-02	4.34E-02
Bile acid metabolism	110	34	2.04E-02	4.63E-02
KEGG Pathways				
Lipid and atherosclerosis	205	79	6.41E-06	2.28E-03
PPAR signaling pathway	72	31	1.20E-04	2.13E-02
Pathways in cancer	510	181	2.48E-04	2.80E-02

Parathyroid hormone synthesis, secretion and action	105	51	3.15E-04	2.80E-02
Circadian entrainment	96	47	5.64E-04	4.00E-02
PI3K-Akt signaling pathway	342	124	1.00E-03	4.57E-02
Insulin resistance	105	45	1.02E-03	4.57E-02
Non-small cell lung cancer	71	36	1.03E-03	4.57E-02

*Genes with CpGs (as assigned by Illumina) that are associated with smoking

Table S12. CpGs showing strongest evidence of being impacted by the interaction between smoking and estimated cell types.

					Interaction Effect		Main Effect	
Cell-type	**Name**	**Chromosome**	**Position**	**Gene**	**log2[Fold Change]**	**P-Value**	**log2[Fold Change]**	**P-Value**
Lymphocyte	cg26075905	14	102391388	*PPP2R5C*	0.22	5.10E-12	0.06	0.08
Lymphocyte	cg17616967	2	234234525	*SAG*	0.94	5.40E-12	0.34	0.01
Lymphocyte	cg26848446	11	3382329	*ZNF195*	0.48	9.20E-11	0.23	0.002
Lymphocyte	cg12529083	12	13197360	*KIAA1467*	-0.28	1.38E-10	-0.15	0.0006
Lymphocyte	cg12722058	15	23086394	*NIPA1*	-0.28	9.10E-10	-0.22	2.01E-06
Lymphocyte	cg12901038	11	2815078	*KCNQ1*	0.39	1.15E-09	0.13	0.05
Lymphocyte	cg09197895	14	105512268	N/A	-0.22	2.36E-09	-0.14	0.0001
Lymphocyte	cg04180093	7	99686359	*COPS6*	-0.52	2.70E-09	-0.24	0.007
Lymphocyte	cg12285709	14	102975664	*ANKRD9*	-0.32	2.74E-09	-0.07	0.17
Lymphocyte	cg06509613	18	53588274	*LINC01416*	0.38	3.78E-09	0.04	0.5
Mono	cg24096902	1	27754893	*WASF2*	0.23	5.90E-09	0.16	8.50E-05
Mono	cg23606722	12	133319848	*ANKLE2*	0.41	1.10E-08	0.18	0.01
Macro	cg11538937	2	241559124	*GPR35*	-0.20	1.10E-08	-0.11	0.003
Macro	cg24096902	1	27754893	*WASF2*	0.23	2.50E-08	0.18	6.40E-05
Endo	cg06232680	21	39608414	*KCNJ15*	0.346	2.60E-09	-0.049	0.367
Endo	cg14800014	5	125800764	*GRAMD3*	0.212	1.90E-08	-0.131	0.0003

Footnote: *Abbreviations*: Mono, Monocyte; Macro, Macrophage; Endo, Endothelial cell.

Table S13. Smoking-by-cell type interaction results for the CpGs in lung showing the strongest evidence of association with smoking in the primary EWAS analysis. (table in separate file)

References:

1. Silva CP, Kamens HM. Cigarette smoke-induced alterations in blood: A review of research on DNA methylation and gene expression. *Exp Clin Psychopharmacol*. 2021;29(1):116-135. doi:10.1037/pha0000382

2. Tsai PC, Glastonbury CA, Eliot MN, et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. *Clin Epigenetics*. 2018;10(1):126. doi:10.1186/s13148-018-0558-0

3. Everson TM, Vives-Usano M, Seyve E, et al. Placental DNA methylation signatures of maternal smoking during pregnancy and potential impacts on fetal growth. *Nat Commun*. 2021;12(1):5095. doi:10.1038/s41467-021-24558-y