Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking.

Natalie S. Shenker,[1] Silvia Polidoro,[2] Karin van Veldhoven,[2,3] Carlotta Sacerdote,[2] Fulvio Ricceri,[2] Mark A. Birrell,[4] Maria G. Belvisi,[4] Robert Brown,[1] Paolo Vineis,[2,3] and James M. Flanagan[1 *]

1. Epigenetics Unit, Department of Surgery and Cancer, Imperial College London, W12 0NN, UK.
2. HuGeF Foundation, 52, Via Nizza, Torino,10126, Italy.
3. MRC-HPA Centre for Environment and Health, School of Public Health, Imperial College London, W2 1PG, UK.
4. Respiratory Pharmacology Group, National Heart and Lung Institute, South Kensington Campus, Exhibition Road, Imperial College London. London SW7 2AZ, UK.

The authors wish it to be known that the first two authors should be regarded as joint First Authors.

***Address for correspondence:**

Dr. James M. Flanagan
Epigenetics Unit, Division of Cancer, Department of Surgery and Cancer,
Faculty of Medicine, Imperial College London
4th Floor IRDB, Hammersmith Campus, Du Cane Road, London W12 0NN
Tel: 020 75942127; fax: 020 75942129; email: j.flanagan@imperial.ac.uk

Abstract

A single CpG site within *F2RL3* was recently found to be hypomethylated in peripheral blood genomic DNA from smokers compared to former and non-smokers. We performed two epigenome-wide association studies (EWAS) nested in a prospective healthy cohort using the Illumina 450K Methylation Beadchip. The two populations consisted of matched pairs of healthy individuals (n=374), of which half went on to develop breast or colon cancer. The association was analysed between methylation and smoking status, as well as cancer risk. In addition to the same locus in *F2RL3*, we report several loci that are hypomethylated in smokers compared to former and non-smokers, including an intragenic region of the aryl hydrocarbon receptor repressor gene (*AHRR*; cg05575921, $p=2.31x10^{-15}$; effect size = 14%-17%), an intergenic CpG island on 2q37.1 (cg21566642, $p=3.73x10^{-13}$; effect size = 12%), and a further intergenic region at 6p21.33 (cg06126421, $p=4.96x10^{-11}$, effect size = 7%-8%). Bisulphite pyrosequencing validated six loci in a further independent population of healthy individuals (n=180). Methylation levels in *AHRR* were also significantly decreased ($p<0.001$) and expression increased ($p=0.0047$) in the lung tissue of current smokers compared to non-smokers. This was further validated in a mouse model of smoke exposure. We observed an association with breast cancer risk for the 2q37.1 locus ($p=0.003$, adjusted for smoking status), but not for the other loci associated with smoking. These data show that smoking has a direct effect on the epigenome in lung tissue, which is also detectable in peripheral blood DNA and may contribute to cancer risk.

Introduction

Epigenetic profiles, including methylation of the 5-carbon of cytosines, are helping to unravel the pathogenesis of numerous complex diseases, in particular for diseases that have an environmental component to their etiology. Since the first studies by Doll in the 1950s that linked smoking with lung cancer risk, smoking has been identified as a major risk factor for numerous cancers. However, the molecular changes that occur as a consequence of the neoplastic process itself, in addition to changes caused by chemo- and radiotherapy, mean that any investigation into the epigenetic alterations induced by cancer risk factors must be performed prospectively in pre-diagnostic samples.

Improvements in array-based technology have enabled the molecular effects of altered environments, such as cigarette smoke inhalation, to be assessed in larger populations. The only study to show a significant link between smoking and the epigenome used the Illumina 27K platform and identified a single locus in the *F2RL3* gene out of 27,578 loci tested that was less methylated in current smokers (n = 65) compared to former (n = 56) and non-smokers (n = 56).(1) This gene is associated with platelet activation and coagulation, but cannot be clearly connected to the carcinogenic processes that are induced by an individual's exposure to tobacco smoke.

The introduction of the Illumina 450K methylation bead array has enabled the analysis of the DNA methylation pattern across the genome with additional coverage of promoters, 5'UTRs, first exons, gene bodies and 3'UTRs. This study aimed to assess the impact of current and former smoking on DNA methylation in an epigenome-wide approach using prospectively collected blood samples from healthy individuals who subsequently developed breast or colon cancer compared to matched controls. We hypothesised that smoking induces gene specific methylation,

detectable in peripheral blood DNA and that these exposure induced changes may impact on cancer risk.

Results

Epigenome-wide association studies

In this study, we performed two epigenome-wide association studies (EWAS) on genomic DNA from peripheral white blood cells (WBC) that were prospectively collected from two nested case-control studies within a large cohort from the general population. All individuals were healthy at the time of blood collection, but cases were selected from individuals who subsequently developed either breast or colon cancer (average lag-time to diagnosis = 4.6 and 7 years, respectively). These samples were being analysed to identify markers associated with breast cancer and colon cancer risk and survival; however, these analyses are under-powered to detect genome-wide significant individual markers and will need to be validated in larger sample sizes currently underway (see Supplementary Figures 1-2). We have used this data in order to further examine the association between methylation in peripheral blood genomic DNA from smokers compared to former and non-smokers.(1) For the present analysis, we used multivariate linear regression to investigate the association between DNA methylation levels and smoking status, adjusting for age and batch. Using a cutoff of $p < 1x10^{-5}$, we identified 17 and 19 loci in the breast cancer and colon cancer EWAS, respectively, that were differentially methylated between smokers, former smokers and those who had never smoked (Figure 1). Eight of these loci were shared by both studies (Table 1). The top hits for both studies ($p < 1x10^{-5}$) are shown in Supplemental Table 1. In all instances, the degree of methylation was lower in smokers than non-smokers, and less difference was found between former smokers and non-smokers. Of note, the

F2RL3 locus previously identified was in this list. There was no strong association ($p < 1x10^{-5}$) between methylation levels at any of these loci and disease status, despite smoking being a weak risk factor for colon cancer (Supplementary Table 2).(2, 3) We observed no evidence for an association between smoking and risk of colon cancer in this study (p = 0.857), which was likely to be a consequence of the small numbers of participants in the study. One of the smoking-associated loci (cg01940273) at 2q37.1 showed an association with developing breast cancer, after adjustment for smoking (p = 0.003) and estrogen receptor status (p = 0.035). This association showed significant heterogeneity by smoking status as it associates with breast cancer case-control status in a logistic regression model (Interaction P=0.039). This region, therefore, warrants further investigation in larger studies as a cancer risk marker and a mechanism for smoking-induced carcinogenesis.

Bisulphite pyrosequencing validation

We used bisulphite pyrosequencing on an additional set of healthy subjects (n = 180) to validate the methylation association with smoking using an alternative method. We validated the methylation of six CpG sites identified by the 450K array (2 CpG sites in *AHRR,* 2 CpG sites in 2q37, 1 CpG site in *F2RL3* and 1 CpG site in 6p21.33). The direction of methylation change and effect sizes were of the same magnitude in the validation groups, and again, a significant association with smoking was observed (Table 1). There was also remarkable evidence for an association between methylation levels and smoking intensity for the *AHRR*, 2q37 and 6p21 loci (p < 6.09e-05), but not *F2RL3*, with individuals smoking ≥4 cigarettes per day having significantly lower methylation levels at these genomic loci (Supplementary Figure 4, Supplementary Table 3). In former smokers, methylation levels at these genomic loci returned to

the levels of non-smokers with increasing time from cessation and those who had smoked more intensively had methylation levels that were closer to that of current smokers (Supplementary Figure 5).

Regional association plots showing the intragenic CpG island in *AHRR* and the intergenic CpG island at 2q37 are shown in Figure 2 with examples and pyrosequencing validation shown in Supplementary Figure 6. The CpG site in one of the *AHRR* regions of interest (AHRR_p1, cg23576855) was also the site of a CG→CA SNP (rs6869832), with an A allele frequency of approximately 10% as confirmed by pyrosequencing (Figure 3). Minor allele (A) carriers (n = 31) were excluded from the pyrosequencing statistical analysis, as the cytosine could not be methylated in the CpA dinucleotide. Interestingly, three of the current smokers in the validation set were heterozygous carriers; their methylation levels (mean, 37.7%) were approximately half of the value found in smokers who were homozygous for the G allele (mean, 66.9%; Table 1). This indicated that the CpG site on the G allele was methylated to a similar degree as homozygous G alleles. We did not have genotyping data on the individuals in the EWAS sample sets, but predict that the individuals with ~40% and 5% β-values are likely to be heterozygous and homozygous for this SNP, respectively (Supplementary Figure 6).

Methylation and expression in human and mouse lung tissue

Pyrosequencing assays were conducted on bisulphite-converted genomic DNA extracted from 27 human lung samples. We have investigated lung tissue in this case to assess the most relevant tissue to the initial exposure. A marked association was present between smoking status and methylation levels in human lung tissue. As in peripheral circulating mononuclear cell DNA, methylation levels at cg23576855 and cg21161138 in the *AHRR* gene were significantly decreased in

current smokers (p = 0.00327 and 0.00143, respectively) (Figure 4A). The methylation values in lung tissue were also identical to those shown in PBMCs, which suggests that blood sampling could offer a useful surrogate for future biomarker studies of lung tissue methylation for this gene.

Expression analyses using qRT-PCR for *AHRR* in human lung samples from smokers vs. non-smokers (n = 5 for each group) was inversely correlated with methylation levels (R^2=0.157) and showed increased expression by 5.7-fold (p = 0.0047) in the current smokers compared to the non-smokers (Figure 4B). In a mouse model of smoke exposure (4), smoking significantly increased the levels of *Ahr* and *Cyp1a1* in a time-dependent manner compared to controls (Figure 4C), while *Ahrr* expression was initially reduced 2.6-fold after the initial exposure to smoke (3 d; $p = 4.56 \times 10^{-6}$), but increased by 1.7-fold after 28 d (p = 0.003). Methylation of the *Ahrr* locus was not performed as this intragenic CpG island region is not conserved in the mouse genome.

Discussion

The findings from these three groups of individuals give strong evidence for the role of smoking in inducing changes in DNA methylation levels. In this study, we have validated previously identified associations with methylation and smoking for the *F2RL3* and *AHRR* probes and have identified further *AHRR* probes that were significantly associated with smoking.(1, 5) We have identified these in two cancer case-control studies using the Illumina 450K methylation beadchip and validated them by bisulphite pyrosequencing in an additional validation cohort. Our study has identified two novel loci, at 2q37.1 and 6p21, which are also strongly associated with smoking status. Importantly, we show for the first time that one of these loci, 2q37.1, is also associated with breast cancer risk. Lastly, we show that these associations between smoking and methylation are unlikely to be cell type specific differences due to different blood cell proportions.

Previous evidence has suggested that intragenic methylation levels are correlated with expression levels, with highly expressed genes having high levels of intragenic methylation, and vice versa.(6-8) According to this rationale, we predicted that decreased levels of methylation in the *AHRR* gene in smokers would indicate lower levels of expression. In the smoke-exposed mice, this was indeed the case with short-term exposure (3 d), but the consequent increase in Ahrr expression after 28 d of smoke exposure suggest that other compensatory mechanisms of gene induction override the short-term decrease in expression marked by lower intragenic DNA methylation levels. This supports data that shows an increase in Ahrr expression following exposure to benzo(a)pyrene, a chemical found in cigarette smoke, in a mouse model.(9) In humans, the expression of *AHRR* in lung tissue mirrored the long term exposure in mice, however, the observed hypomethylation of this locus in lung tissue may be an indicator of past expression changes in this differentiated tissue type.

A recent study of 165 individuals on the 450K methylation array identified a single CpG site within the *AHRR* gene associated with smoking status in EBV-transformed lymphoblastoid cell lines.(5) With a larger study size of 554 individuals, we have validated the association and identified additional intragenic CpG sites in the *AHRR* gene associated with smoking in white blood cell DNA and lung tissue. The aryl hydrocarbon receptor (AHR) is a crucial receptor in the pathway that metabolizes a range of biological compounds and synthetic environmental pollutants. Benzopyrene and dioxin-like compounds are highly toxic organic molecules that are released from various components of cigarettes during smoking and enter the circulation via the pulmonary vasculature. These compounds are metabolised by AhR,(10, 11) releasing further carcinogenic metabolites that have been implicated in lung cancer development.(12). However, the AHR pathway may also mediate carcinogenesis through other pathways such as oxidative

stress.(13) *AHRR* (aryl hydrocarbon receptor repressor; chromosome position 5p15.33) encodes for a class E basic helix-loop-helix protein,(14) which is mainly found in the cytoplasm. It inhibits the translocation of the AHR-ligand complex into the nucleus by disrupting the binding of AHR to the AHR-nuclear translocator (*AHRNT*).(15) The AHR-AHRNT heterodimer also inhibits the transcription of *AHRR* in a feedback loop. The knockdown of *AHRR* has been shown to produce increased tumour cell invasiveness in a range of tissue types, including breast, colon, lung, and ovarian. (16, 17) An increase of the repressor of the AHR pathway should lead to a decrease in the activity of the pathway, and therefore, a decrease in AHR pathway-mediated carcinogenesis. However, the data from our study shows an increase in *AHRR* expression and decreased intragenic methylation due to smoking, with no evidence of an association with breast or colon cancer risk. Therefore, more work is needed to understand the complex mechanisms of smoking induced carcinogenesis via the AHR pathway.

A novel finding of this study was that a genomic locus, comprising four consecutive probes at 2q37.1, was differentially methylated between smokers and former or non-smokers. The four probes are located within 824 bp of an intergenic CpG island (chr2: 233,284,112 – 233,284,935; http://genome.ucsc.edu/, hg19). This region maps to a DNase hypersensitivity site within a CpG island, indicating a possible regulatory region (Supplemental Figure 7),(18) and is a potential pseudogene of *ECEL1* (endothelin-converting enzyme-like 1; chr2: 233,344,537 to 233,351,464 bp, 60,425 bp downstream of 2q37), as it possesses a high level of sequence homology (>95%). We observed an association with breast cancer risk at this locus with significant heterogeneity in the effect in smokers compared to non-smokers. While the estimate of breast cancer risk associated with smoking is still unclear (19), future work on larger case control studies will be needed to validate this cancer association. We hypothesize that molecular markers of smoking

exposure, such as methylation markers described here, will provide more accurate measurement of the exposure than questionnaire based data and allow a more robust assessment of any associated cancer risk. Whether this 2q37 locus is causally linked to breast cancer risk due to smoking is not yet known, however, if verified we hypothesize that the mechanism may involve *in cis* regulation of the developmentally regulated homologue, *ECEL1*. There are several limitations in establishing causality between an epigenetic trait, an exposure and cancer risk. These include the reversible nature of epigenetic modifications and the need for an appropriate tissue type in which to investigate gene expression – methylation associations (for example a large epidemiology sized sample set of normal breast tissue prior to disease onset).

A second novel intergenic locus with lower methylation levels in smokers was at 6p21.33 (chr 6: 307,020,080). This locus is in a gene desert which maps onto a DNase I hypersensitivity site and transcription factor binding site, associated with an H3K27 acetylated chromatin site, which is often associated with active regulatory elements. It is not associated with any gene or SNP. Further studies will be required to assess the functional nature of these two novel intergenic genomic loci.

Toxic components of cigarette smoke enter the bloodstream via the alveolar capillary system, after which they could directly affect the epigenetic profile of circulating WBC. For the loci that correlated with smoking status, we showed a strong correlation with time since quitting and duration of smoking, with methylation levels eventually returning to those of non-smoker levels. This is an observation that has been seen for methylation at the *F2RL3* locus using the 27K methylation chip,(20) and in the gene expression signature of smoking for some genes.(21) Interestingly, the expression of some genes, including *AHRR*, do not return to previous levels, which indicates a long-term gene expression consequence of prior smoking history that may be

locked in by DNA methylation changes.(21) However, we have to consider the possibility that other confounding factors may influence this association. We and others have not adjusted for alcohol and body mass index, which may be smoking-associated or methylation-associated confounding factors. Furthermore, we have assessed methylation levels in the genomic DNA of all circulating WBC. Smoking increases the circulating WBC count;(22) although the effect of cigarette smoke on specific WBC types have not been assessed in large-scale populations, it may be that subsets with different methylation levels are clonally expanded due to the exposure, which might affect the overall methylation result. Using fractionation of blood cell types into T-cells, B-cells, monocytes, granulocytes and buffy coat (all leukocytes) compared to whole blood cell DNA from the same individuals we have shown no evidence that any of these blood cell types have significantly different methylation levels that would confound the association with smoking (Supplementary Table 6). Lastly, genetic haplotype differences between individuals tagging allele specific methylation may confound associations with smoking status if such haplotypes are also associated with smoking (23). We have performed a preliminary analysis of 25 individuals for which we have both GWAS data and EWAS 450K methylation data (unpublished data). We found evidence for potential allele specific methylation at the F2RL3 locus (general logistic regression p=0.01), but not for the other smoking associated loci (Supplementary Table 7). While the AHRR CG>CA SNP (cg23576855, presented in Figure 3) does indeed influence the methylation of this site and is in linkage disequilibrium with the haplotype tag SNP, it does not influence allele specific methylation of the haplotype given that the two other nearby CpG sites (79bp and 26,061 bp apart from the first site) show no evidence of allele specific methylation (p=0.798; p=0.916, respectively). Therefore, for the majority of loci it is unlikely that the methylation association with smoking reported in this study is due to genetic polymorphism.

Taken together, these studies report a strong link between tobacco use and a direct biological result on the epigenome, which is detectable in blood and lung tissue, and may impact on the cancer risk associated with smoking. The level of methylation was directly associated with smoking intensity and duration, however, the biological consequences of these epigenetic alterations will require further investigation, particularly at these novel intergenic loci. The results of our study and others regarding a key gene in the AHR pathway may have relevance beyond tobacco smoke exposure given the key role played by this pathway in the metabolism of many environmental carcinogens.

Materials and Methods

Subject recruitment

Study participants were drawn from the Italian component of the European Prospective Investigation into Cancer and Nutrition cohort, a large general population cohort consisting of ~520,000 individuals with standardised lifestyle and personal history questionnaires, anthropometric data and blood samples collected for DNA extraction.(24, 25) In this study, we used the Torino sub-cohort (EPIC-Turin) which consists of ~10,000 subjects. For the microarray component of this study we included 92 incident female breast cancer cases and 92 matched controls and 95 incident colon cancer cases and 95 matched controls including 68 male and 27 female pairs. Controls were individually matched on age (±5 years), sex, seasonality of blood collection and duration of follow-up. Blood samples from cancer cases were taken 55 months before diagnosis on average (range, 24-108 months) for breast cancer cases and 84 months before diagnosis (range, 0.2-173 months) for colon cancer cases (Supplementary Table 5). Blood cell

type fractionation was performed on six healthy volunteers using magnetic bead separation (EasySep, StemCell Technologies, France).

Microarray protocol

For the microarray, DNA samples were extracted using the QIAsymphony DNA Midi Kit (Qiagen, Crawley, UK). Bisulphite conversion of 500 ng of each sample was performed using the EZ-96 DNA Methylation-Gold™ Kit (Zymo Research, Orange, CA). Bisulphite-converted DNA was used for hybridisation on the Infinium HumanMethylation 450 BeadChip, following the Illumina Infinium HD Methylation protocol. The methylation score for each CpG was represented as a β-value according to the fluorescent intensity ratio representing any value between 0 (unmethylated) and 1 (completely methylated). Raw microarray data and processed normalised data will be available from Gene Expression Omnibus (GEO) (accession TBA).

Validation of array-based methylation results by pyrosequencing

For the validation component of this study, 180 healthy control individuals were randomly sampled from females enrolled in the EPIC-Turin cohort. This group comprised 33 current smokers, 45 former smokers and 102 individuals who had never smoked. Genomic DNA (250 ng) from each subject was bisulphite converted as above. For pyrosequencing, specific primers were designed for six CpG loci using PyroMark software (Qiagen, Hilden Germany), and PCR conditions were as described previously.(26) (Supplemental Table 6). Methylation values were calculated as an average of all high quality CpG sites, which were determined as "passed" by the quality control thresholds within the Pyro Q-CpG software (Qiagen). Pyrosequencing assays for three loci (cg05951221, cg01940273 and cg05575921) could not be designed.

Human and murine expression analyses

Human lung samples (n=27, 14 smokers, 13 non-smokers) were obtained from either lung transplants performed at The Royal Brompton or Harefield Hospital or purchased from IIAM (International Institute for the Advancement of Medicine, Edison, NJ). In all cases, the tissue was consented for use in scientific research and ethics approval obtained from the Royal Brompton & Harefield Trust. DNA was extraction and pyrosequencing was performed as described above. RNA was extracted from lung tissue (n=5 smokers, n=5 non-smokers), and quantitative RT-PCR (qRT-PCR) was performed on a Bio-Rad PCR machine with SyBR green (Sigma), for *AHRR* normalised against *GAPDH*, using standard protocols. Lung tissue was collected from ten mice exposed to air (n=5) or cigarette smoke (n=5), as described previously.(7) RNA was extracted and qRT-PCR was performed for *Ahrr*, *Ahr*, *Cyp1a1* and *F2rl3* as described above, using a ribosomal gene, *Rpl7*, as the internal reference standard.

Statistical analysis

For the statistical analysis, raw data was exported from GenomeStudio (Illumina) as background subtracted β-values with corresponding detection p-values. Following quality control the resulting data sets included 86 breast cancer cases with 87 matched controls (86 matched pairs) and 95 colon cancer cases and 95 matched controls. Overall, 484,804 probes were analysed in the breast cancer data set and 485,152 in the colon cancer data set. There were no significant genome-wide methylation differences between smokers, former and never smokers in raw β-values (Supplementary Figure 6); therefore, the data was normalised using quantile normalisation. Multivariate linear regression was used to identify associations between methylation β-values as

the outcome and coded smoking status as the exposure (0, 1 and 2 for 'Never', 'Former' and 'Current' smokers, respectively), adjusting for age and batch. A secondary analysis was performed that also adjusted for case-control status, which did not significantly alter the results (Supplementary Table 2). P-values less than $1x10^{-7}$ were considered to be significant at the level of epigenome-wide significance. We found no evidence for a bias towards probes containing SNPs or type I or type II probes in probes associated with smoking. For association with smoking intensity, coded intensity categories were used as the exposure, Time to quitting and duration of smoking in former smokers were analysed as continuous variables. Attributable risk for the 2q37 locus was calculated as the difference in rate of hypomethylation between an exposed population (cancer cases) and an unexposed population (healthy controls), stratified by smoking status. All analyses were performed in R, v2.13.1. Gene expression levels were compared between the control and smoke-exposed human or mouse lung tissue using Student's t-tests with $P < 0.05$ considered significant.

Acknowledgements

JMF is funded by a Breast Cancer Campaign Fellowship. JMF and RB acknowledge funding from Cancer Research UK (A13086) and the Imperial Biomedical Research Centre. PV is funded by the HuGeF Foundation, Torino, Italy. NS is funded by a Medical Research Council UK graduate scholarship. The human tissue experiments in this study were undertaken with the support of the NIHR Respiratory Disease Biomedical Research Unit at the Royal Brompton and Harefield NHS Foundation Trust and Imperial College London.

Statement of competing financial interests

None declared.

Supplemental Data

Supplemental data include supplementary methods, seven figures and seven tables (see attached files).

References

1. 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.

2. Sasco, A.J., Secretan, M.B. and Straif, K. (2004) Tobacco smoking and cancer: a brief review of recent epidemiological evidence. *Lung Cancer ,* **45** Suppl 2, S3-S9.

3. Hannan, L.M., Jacobs, E.J. and Thun, M.J. (2009) The association between cigarette smoking and risk of colorectal cancer in a large prospective cohort from the United States. *Cancer Epidemiol. Biomarkers Prev.,* **18**, 3362-3367.

4. Eltom, S., Stevenson, C.S., Rastrick, J., Dale, N., Raemdonck, K., Wong, S., Catley, M.C., Belvisi, M.G. and Birrell, M.A. (2011) P2X7 receptor and caspase 1 activation are central to airway inflammation observed after exposure to tobacco smoke. *PloS One,* **6**, e24097.

5. Monick, M.M., Beach, S.R., 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.

6. Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S. (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.,* **39**, 61-69.

7. Aran, D., Toperoff, G., Rosenberg, M. and Hellman, A. (2011) Replication timing-related and gene body-specific methylation of active human genes. *Hum. Mol. Genet.,* **20**, 670-680.

8. Shenker, N. and Flanagan, J.M. (2012) Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research. *Br. J. Cancer,* **106**, 248-253.

9. Bernshausen, T., Jux, B., Esser, C., Abel, J. and Fritsche, E. (2006) Tissue distribution and function of the Aryl hydrocarbon receptor repressor (AhRR) in C57BL/6 and Aryl hydrocarbon receptor deficient mice. *Arch. Toxicol.,* **80**, 206-211.

10. Moennikes, O., Loeppen, S., Buchmann, A., Andersson, P., Ittrich, C., Poellinger, L. and Schwarz, M. (2004) A constitutively active dioxin/aryl hydrocarbon receptor promotes hepatocarcinogenesis in mice. *Cancer Res.,* **64**, 4707-4710.

11. Shimizu, Y., Nakatsuru, Y., Ichinose, M., Takahashi, Y., Kume, H., Mimura, J., Fujii-Kuriyama, Y. and Ishikawa, T. (2000) Benzo[a]pyrene carcinogenicity is lost in mice lacking the aryl hydrocarbon receptor. *Proc. Natl. Acad. Sci. USA.,* **97**, 779-782.

12. Chiba, T., Uchi, H., Yasukawa, F. and Furue, M. (2012) Role of the arylhydrocarbon receptor in lung disease. *Int. Arch. Allergy. Immunol.,* **155 Suppl 1**, 129-134.

13. Cheng, Y.H., Huang, S.C., Lin, C.J., Cheng, L.C. and Li, L.A. (2012) Aryl hydrocarbon receptor protects lung adenocarcinoma cells against cigarette sidestream smoke particulates-induced oxidative stress. *Toxicol. Appl. Pharmacol.,* **259**, 293-301.

14. Baba, T., Mimura, J., Gradin, K., Kuroiwa, A., Watanabe, T., Matsuda, Y., Inazawa, J., Sogawa, K. and Fujii-Kuriyama, Y. (2001) Structure and expression of the Ah receptor repressor gene. *J. Biol. Chem.,* **276**, 33101-33110.

15. Evans, B.R., Karchner, S.I., Allan, L.L., Pollenz, R.S., Tanguay, R.L., Jenny, M.J., Sherr, D.H. and Hahn, M.E. (2008) Repression of aryl hydrocarbon receptor (AHR) signaling by AHR repressor: role of DNA binding and competition for AHR nuclear translocator. *Mol. Pharmacol.,* **73**, 387-398.

16. Zudaire, E., Cuesta, N., Murty, V., Woodson, K., Adams, L., Gonzalez, N., Martinez, A., Narayan, G., Kirsch, I., Franklin, W. *et al.* (2008) The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers. *J. Clin. Invest.,* **118**, 640-650.

17. Kanno, Y., Takane, Y., Izawa, T., Nakahama, T. and Inouye, Y. (2006) The inhibitory effect of aryl hydrocarbon receptor repressor (AhRR) on the growth of human breast cancer MCF-7 cells. *Biol. Pharm. Bull.,* **29**, 1254-1257.

18. Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D. *et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.,* **16**, 123-131.

19. Cox, D.G., Dostal, L., Hunter, D.J., Le Marchand, L., Hoover, R., Ziegler, R.G. and Thun, M.J. (2011) N-acetyltransferase 2 polymorphisms, tobacco smoking, and breast cancer risk in the breast and prostate cancer cohort consortium. *Am. J. Epidemiol.,* **174**, 1316-1322.

20. Wan, E.S., Qiu, W., Baccarelli, A., Carey, V.J., Bacherman, H., Rennard, S.I., Agusti, A., Anderson, W., Lomas, D.A. and Demeo, D.L. (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum. Mol. Genet.,* **21**, 3073-3082.

21. Bosse, Y., Postma, D.S., Sin, D.D., Lamontagne, M., Couture, C., Gaudreault, N., Joubert, P., Wong, V., Elliott, M., van den Berge, M. *et al.* (2012) Molecular signature of smoking in human lung tissues. *Cancer Res.,* **72**, 3753-3763.

22. Wannamethee, S.G., Lowe, G.D., Shaper, A.G., Rumley, A., Lennon, L. and Whincup, P.H. (2005) Associations between cigarette smoking, pipe/cigar smoking, and smoking cessation, and haemostatic and inflammatory markers for cardiovascular disease. *Eur. Heart J.,* **26**, 1765-1773.

23. Munafo, M.R., Timofeeva, M.N., Morris, R.W., Prieto-Merino, D., Sattar, N., Brennan, P., Johnstone, E.C., Relton, C., Johnson, P.C., Walther, D. *et al.* (2012) Association between genetic variants on chromosome 15q25 locus and objective measures of tobacco exposure. *J. Natl. Cancer Inst.,* **104**, 740-748.

24. Riboli, E. and Kaaks, R. (1997) The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.,* **26 Suppl 1**, S6-S14.

25. Riboli, E. (2001) The European Prospective Investigation into Cancer and Nutrition (EPIC): plans and progress. *J. Nutr.,* **131**, 170S-175S.

26. Flanagan, J.M., Cocciardi, S., Waddell, N., Johnstone, C.N., Marsh, A., Henderson, S., Simpson, P., da Silva, L., Khanna, K., Lakhani, S. *et al.* (2010) DNA methylome of familial breast cancer identifies distinct profiles defined by mutation status. *Am. J. Hum.Genet.,* **86**, 420-433.

Figure Legends

Figure 1: Manhattan plot and quantile-quantile plot for EWAS results for smoking status in two case-control studies.

A) Breast cancer case-control study. B) Colon cancer case-control study. In the Manhattan plot, the vertical axis indicates ($-\log_{10}$ transformed) observed p-values, and the horizontal thresholds indicate the significance levels of $p = 1 \times 10^{-5}$ and $p = 1 \times 10^{-7}$. In the quantile-quantile (QQ) plot, the horizontal axis shows ($-\log_{10}$ transformed) expected p-values, and the vertical axis indicates ($-\log_{10}$ transformed) observed p-values. The lambda inflation factor (median[obs]/median[exp]) is shown.

Figure 2: Regional association plots showing two regions associated with smoking status AHRR (A) and 2q37.1 intergenic CpG island (B). Regional association plots are shown with the gene map (from UCSC genome browser, hg19) with a graph of $-\log_{10}$ p-values on the y-axis and nucleotide position on the x-axis for the breast cancer EWAS (blue diamonds), colon cancer EWAS (red circles) and pyrosequencing validation (black asterisks).

Figure 3: Pyrosequencing validation of AHRR locus cg23576855 reveals a CG>CA single nucleotide polymorphism (rs6869832). Boxplots represent pyrosquencing-based methylation levels of individuals with the homozygous G allele in the validation control population. Heterozygotes and homozygotes for the minor A allele are marked as a cross and show the same relationship to smoking status.

Figure 4: Smoking induces methylation and expression changes in human and mouse lung tissue. A) Pyrosequencing data from two regions of interest, presented as boxplots, in human lung tissue samples from smokers (n=14) vs. never smokers (n=13). For cg23576855 individuals heterozygous for the CG>CA polymorphism are marked as crosses. B) AHRR mRNA expression data from human lung tissue from smokers (n=5) versus non-smokers (n=5), showing a 5.7-fold increase in expression in smokers compared to non-smokers (upper panel) and correlation between methylation and expression in the same 5 non-smokers (grey circles) and 5 smokers (black triangles) (lower panel). C) Mouse lung expression changes in AHRR and F2RL3, in addition to two other genes from the aryl hydrocarbon receptor pathway, AHR and CYP1A1. As predicted, AHR and CYP1A1 increase in response to cigarette smoke, and the expression of AHRR initially decreases after 3 days, before increasing after 28 days. There was no significant change in the expression of F2RL3 after exposure to cigarette smoke.

Table 1: Leading differentially methylated genomic loci between smokers and former or non-smokers in a breast and colonic case-control cohort. This list contains the top eight overlapping CG sites between the breast cancer and colon cancer EWAS studies, in addition to an additional site within the 2q37.1 locus that was also validated by pyrosequencing. Rows in bold indicate the data that was independently validated by bisulphite pyrosequencing of that locus in a separate EPIC cohort of healthy individuals (n = 180) with known smoking status.

Chr, chromosome; FDR, false discovery rate; nd, not done; n/a, not applicable.

				B-values in breast cancer case-control cohort					**B-values in colon cancer case-control cohort**					**Pyrosequencing methylation in the validation cohort**				
Target ID	**Chr**	**MAPINFO (bp)**	**Symbol**	Never	Former	Current	**P-value ***	**Effect size (%) ****	Never	Former	Current	**P-value ***	**Effect size (%) ****	Never	Former	Current	**P-value (FDR)**	**Effect size (%) ****
cg06644428	**2**	**233284112**	**2q37.1**	**0.07**	**0.05**	**0.05**	**6.17E-07**	**2**	**0.11**	**0.09**	**0.09**	**3.38E-04**	**3**	**12.11**	**11.02**	**8.11**	**1.48E-05**	**4**
cg05951221	2	233284402	2q37.1	0.39	0.33	0.28	1.80E-13	11	0.41	0.37	0.34	1.83E-07	7					
cg21566642	**2**	**233284661**	**2q37.1**	**0.44**	**0.37**	**0.32**	**3.73E-13**	**12**	**0.51**	**0.47**	**0.39**	**4.41E-08**	**12**	**58.86**	**53.04**	**40.11**	**2.22E-08**	**18.75**
cg01940273	2	233284934	2q37.1	0.58	0.56	0.49	1.47E-10	9	0.62	0.59	0.54	5.96E-09	8					
cg23576855	**5**	**373299**	**AHRR**	**0.66**	**0.64**	**0.5**	**3.46E-06**	**16**	**0.73**	**0.68**	**0.53**	**9.66E-06**	**20**	**82.68**	**80.17**	**66.91**	**1.51E-12**	**15.77**
cg05575921	5	373378	AHRR	0.84	0.79	0.68	2.31E-15	17	0.84	0.81	0.7	1.73E-09	14					
cg21161138	**5**	**399360**	**AHRR**	**0.66**	**0.65**	**0.6**	**5.44E-08**	**5**	**0.72**	**0.71**	**0.68**	**8.29E-06**	**4**	**75.47**	**74.67**	**68.3**	**3.22E-05**	**7.17**
cg03636183	**19**	**17000585**	**F2RL3**	**0.64**	**0.61**	**0.56**	**8.38E-11**	**8**	**0.68**	**0.65**	**0.61**	**3.84E-07**	**7**	**62.19**	**60.92**	**54.32**	**0.00172**	**7.87**
cg06126421	**6**	**30720080**	**6p21.33**	**0.65**	**0.61**	**0.57**	**4.96E-11**	**8**	**0.71**	**0.7**	**0.64**	**2.46E-06**	**7**	**67.4**	**66.8**	**59.5**	**0.00095**	**7.88**

* Logistic regression adjusting for age and batch effect; ** effect size represents percentage methylation difference between current smokers and individuals who have never smoked. Pyrosequencing assays could not be designed for three of the nine CpG sites.

Abbreviations

AHR, aryl hydrocarbon receptor

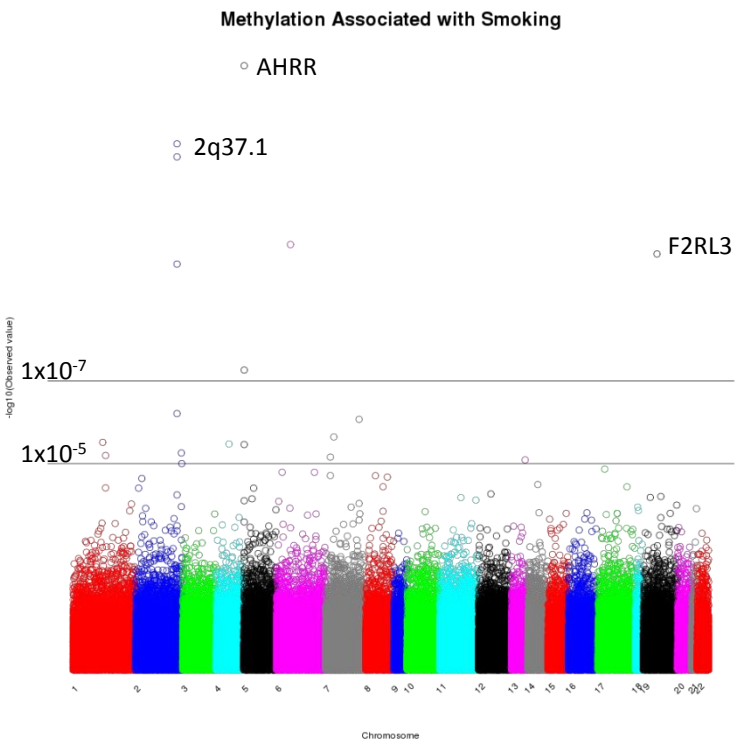
AHRR, aryl hydrocarbon receptor repressor

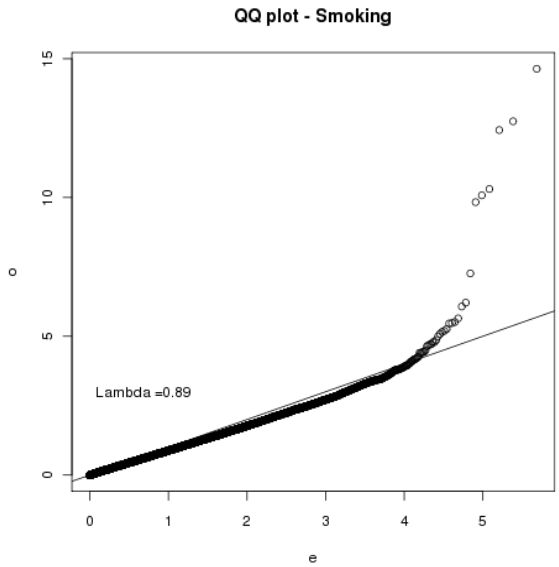
CpG, cytosine-guanine dinucleotide

EWAS, epigenome-wide association study

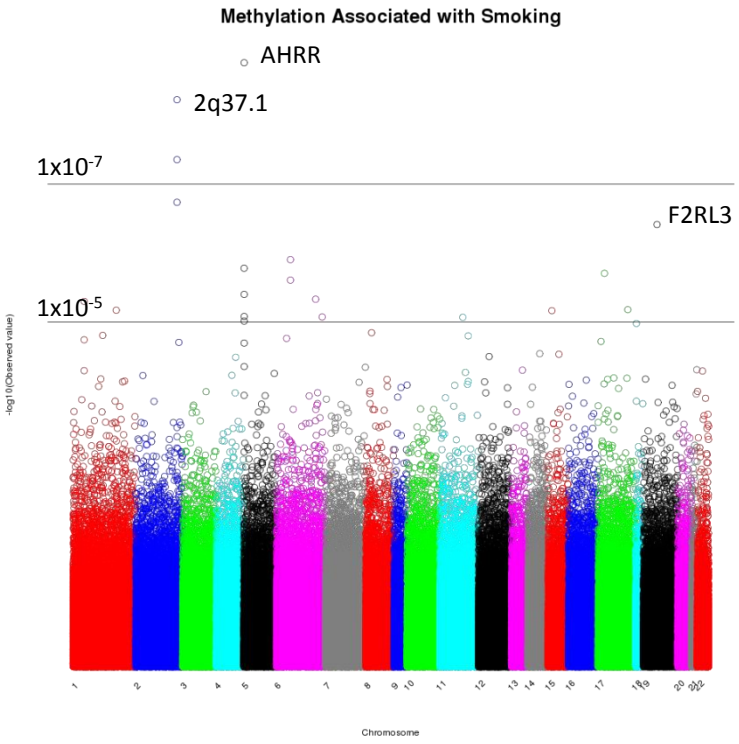
F2RL3, coagulation factor II (thrombin) receptor-like 3

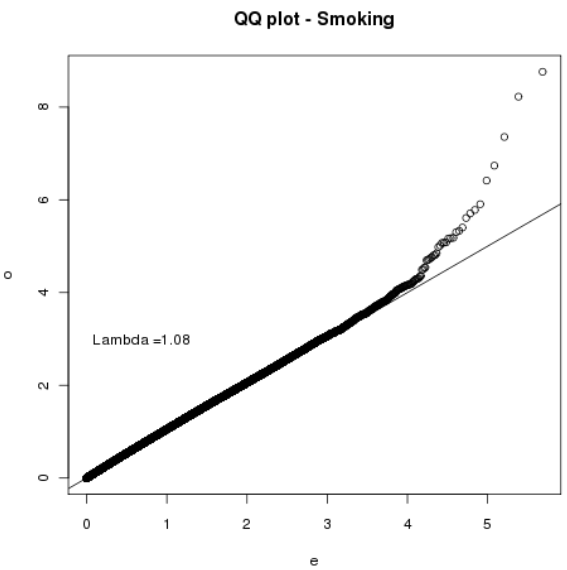
A

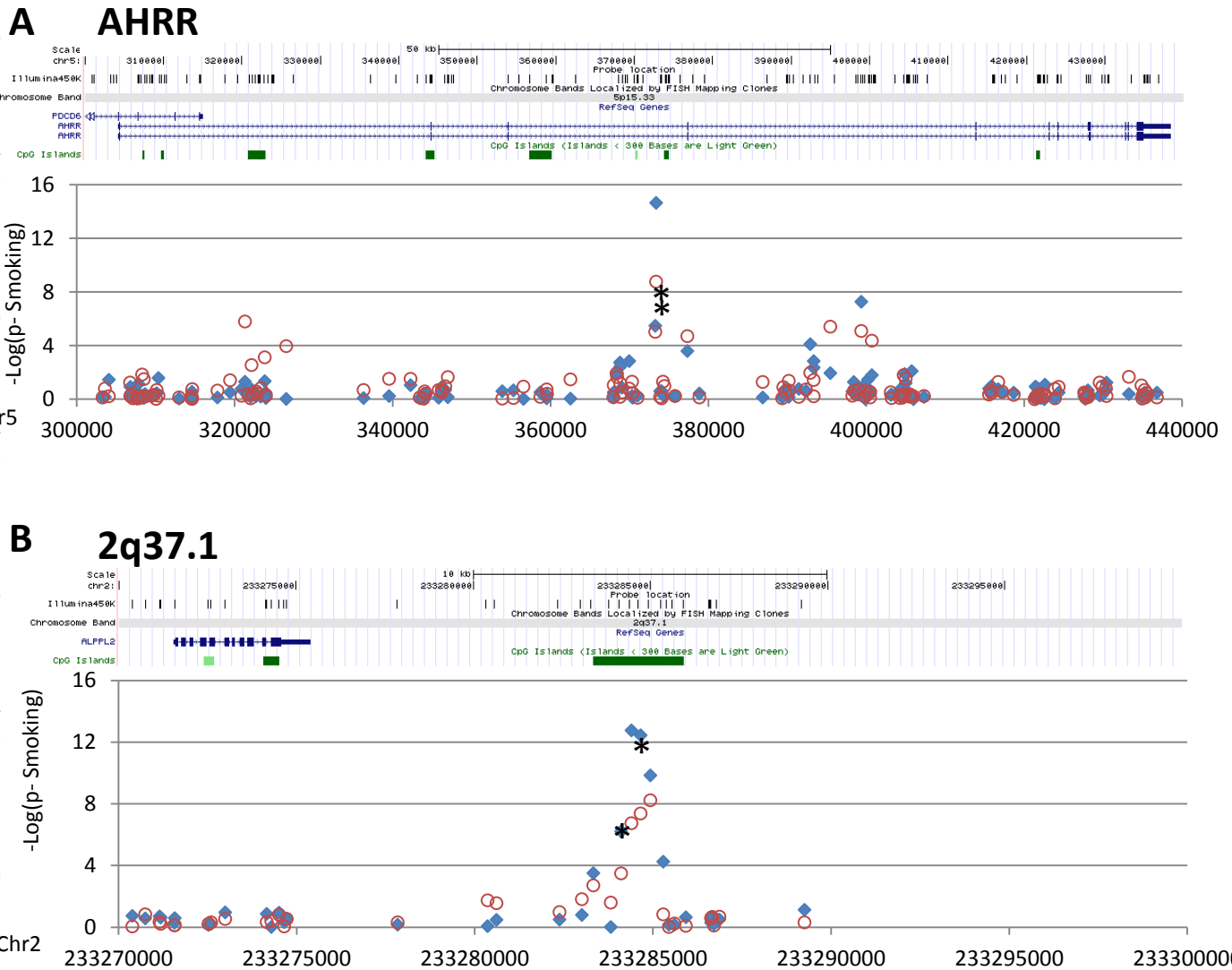

Methylation Associated with Smoking
AHRR
2q37.1
F2RL3
1x10-7
1x10-5
-log10(Observed value)
Chromosome

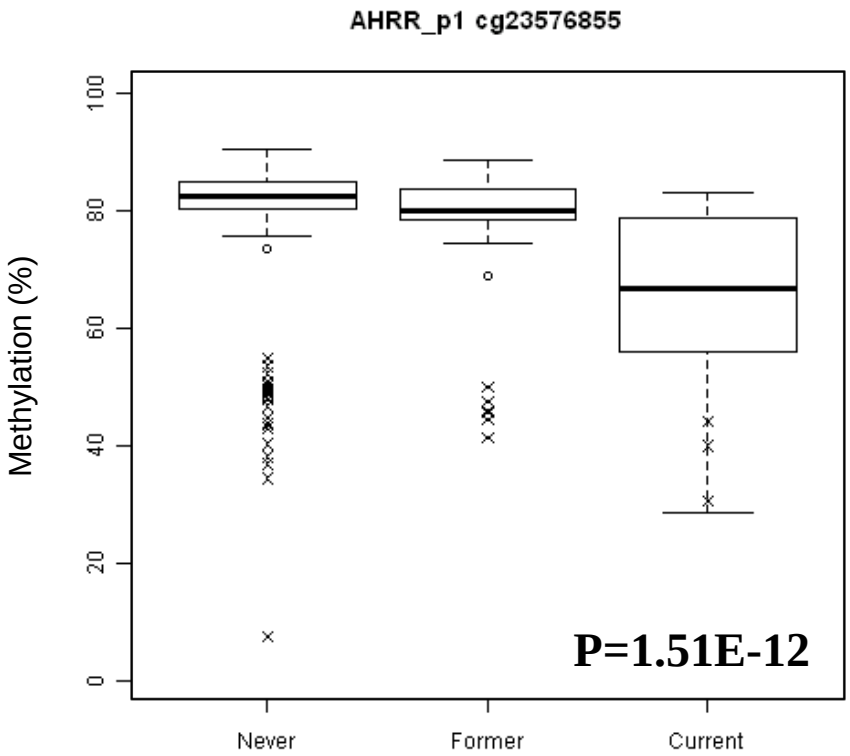

QQ plot - Smoking
Lambda =0.89
o
e

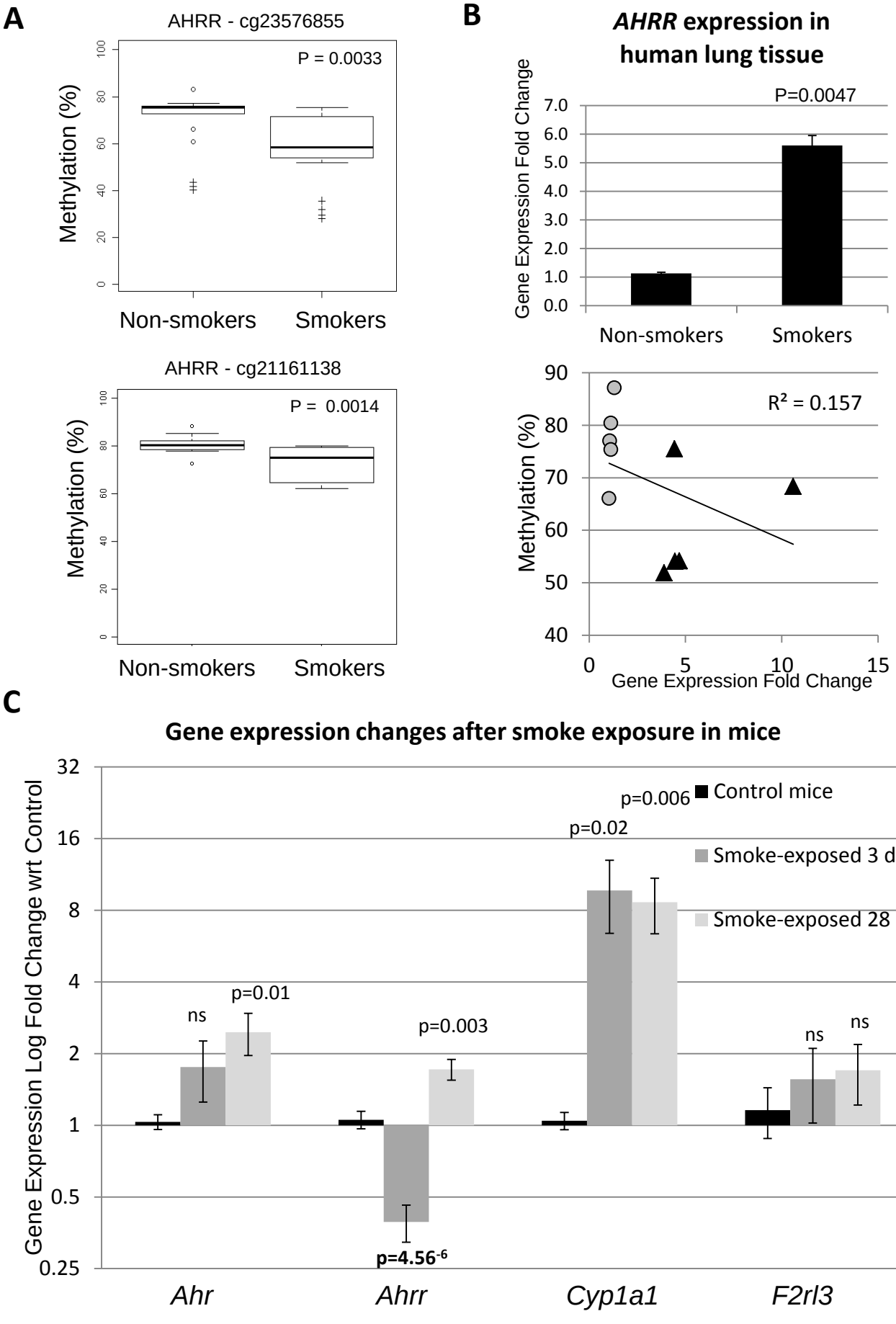
B


Methylation Associated with Smoking
AHRR
2q37.1
1x10-7
F2RL3
1x10-5
-log10(Observed value)
Chromosome


QQ plot - Smoking
Lambda =1.08
o
e


A
AHRR
Scale
chr5:
50 kb
Illumina450K
Probe location
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones
5p15.33
RefSeq Genes
PDCD6
AHRR
AHRR
CpG Islands
CpG Islands (Islands < 300 Bases are Light Green)
-Log(p- Smoking)
Chr5
B
2q37.1
chr2:
10 kb
Chromosome Band
2q37.1
ALPPL2
-Log(p- Smoking)
Chr2


AHRR_p1 cg23576855
Methylation (%)
0
20
40
60
80
100
Never
Former
Current
P=1.51E-12


A
AHRR - cg23576855
P = 0.0033
Methylation (%)
Non-smokers
Smokers
AHRR - cg21161138
P = 0.0014
Methylation (%)
Non-smokers
Smokers
B
AHRR expression in human lung tissue
P=0.0047
Gene Expression Fold Change
Non-smokers
Smokers
R² = 0.157
Methylation (%)
Gene Expression Fold Change
C
Gene expression changes after smoke exposure in mice
Gene Expression Log Fold Change wrt Control
Control mice
Smoke-exposed 3 d
Smoke-exposed 28 d
p=0.006
p=0.02
p=0.01
ns
p=0.003
p=4.56-6
ns
ns
Ahr
Ahrr
Cyp1a1
F2rl3