



# Epigenetic mechanisms of lung carcinogenesis involve differentially methylated CpG sites beyond those associated with smoking

Dusan Petrovic<sup>1,2,3</sup> · Barbara Bodinier<sup>1</sup> · Sonia Dagnino<sup>1</sup> · Matthew Whitaker<sup>1</sup> · Maryam Karimi<sup>1,4,5</sup> · Gianluca Campanella<sup>1</sup> · Therese Haugdahl Nøst<sup>6</sup> · Silvia Polidoro<sup>7</sup> · Domenico Palli<sup>8</sup> · Vittorio Krogh<sup>9</sup> · Rosario Tumino<sup>10</sup> · Carlotta Sacerdote<sup>11</sup> · Salvatore Panico<sup>12</sup> · Eiliv Lund<sup>6,13</sup> · Pierre-Antoine Dugue<sup>14,15,16</sup> · Graham G. Giles<sup>14,15,16</sup> · Gianluca Severi<sup>17</sup> · Melissa Southey<sup>14,16,18</sup> · Paolo Vineis<sup>1</sup> · Silvia Stringhini<sup>2,3</sup> · Murielle Bochud<sup>2</sup> · Torkjel M. Sandanger<sup>6</sup> · Roel C. H. Vermeulen<sup>1,19,20</sup> · Florence Guida<sup>1,21</sup> · Marc Chadeau-Hyam<sup>1,19</sup>

Received: 28 September 2021 / Accepted: 25 April 2022 / Published online: 20 May 2022  
© The Author(s) 2022

## Abstract

Smoking-related epigenetic changes have been linked to lung cancer, but the contribution of epigenetic alterations unrelated to smoking remains unclear. We sought for a sparse set of CpG sites predicting lung cancer and explored the role of smoking in these associations. We analysed CpGs in relation to lung cancer in participants from two nested case–control studies, using (LASSO)-penalised regression. We accounted for the effects of smoking using known smoking-related CpGs, and through conditional-independence network. We identified 29 CpGs (8 smoking-related, 21 smoking-unrelated) associated with lung cancer. Models additionally adjusted for Comprehensive Smoking Index-(CSI) selected 1 smoking-related and 49 smoking-unrelated CpGs. Selected CpGs yielded excellent discriminatory performances, outperforming information provided by CSI only. Of the 8 selected smoking-related CpGs, two captured lung cancer-relevant effects of smoking that were missed by CSI. Further, the 50 CpGs identified in the CSI-adjusted model complementarily explained lung cancer risk. These markers may provide further insight into lung cancer carcinogenesis and help improving early identification of high-risk patients.

**Keywords** Lung cancer · Smoking · DNA methylation · Variable selection · Partial correlation network

## Abbreviations

AUC	Area under the curve
CSI	Comprehensive smoking index
EPIC	European prospective investigation into cancer and nutrition
LASSO	Least absolute shrinkage and selection operator
NOWAC	Norwegian women and cancer study
ROC	Receiver operating characteristics

## Background

Although tobacco smoking has been identified as the leading cause of lung cancer since the early 1950's, and despite considerable progress in the diagnosis and treatment of the disease, lung cancer is still the leading cause of cancer-related deaths worldwide. Lung cancer survival remains dismal (five-year survival rate ranging from 10 to 20%) mainly due to patients being diagnosed at later stages of the disease [1–4]. The lack of effective early diagnostic tools is in part resulting from the incomplete understanding of the molecular mechanisms involved in the carcinogenesis process. Moreover, in populations where the prevalence of smoking is low, an increasing proportion of lung cancer occurs in never-smokers and former smokers [1] highlighting a pressing need for identifying specific molecular mechanisms involved in lung carcinogenesis unrelated to smoking.

While the smoking-induced genetic changes affecting oncogenes and tumour suppressor genes have now been well-established [5], recent findings have shown that

---

Dusan Petrovic and Barbara Bodinier have contributed equally to this work.

---

Dusan Petrovic and Barbara Bodinier are joint first authors.

---

Florence Guida and Marc Chadeau-Hyam are joint last authors.

---

✉ Marc Chadeau-Hyam  
m.chadeau@imperial.ac.uk

Extended author information available on the last page of the article

tobacco use also leads to *epigenetic* modifications through DNA methylation at Cytosine-phosphate-Guanine (CpG) sites, potentially resulting in modified gene expression [2, 5]. Previous studies have shown that the lung epithelium, and other tissues such as blood, buccal cells, and sputum, present relatively conserved methylation profiles in relation to smoking and/or lung cancer [6, 7]. Such epigenetic signatures include methylation changes in CpG sites related to inflammation, detoxification of xenobiotics, or cell proliferation [2, 6, 8–10].

Several studies have investigated the role and potentially mediating molecular pathways affected by smoking-related epigenetic changes in relation to lung carcinogenesis [11–15], but so far, the potential contribution of smoking-independent epigenetic modifications to lung cancer development remains understudied.

In the present study, we used full-resolution DNA methylation data from two European cohorts, the Italian component of the European Prospective Investigation into Cancer and Nutrition (EPIC-Italy) and the Norwegian Women and Cancer Study cohort (NOWAC). We used (LASSO) penalised logistic regression models calibrated via stability selection to identify robust sets CpG sites that are jointly explanatory of future risk of lung cancer. We subsequently investigated to what extent these associations were driven by exposure to tobacco smoke, and further assessed the performance of selected (sets of) CpG sites in predicting lung cancer in comparison to established metrics of tobacco smoke exposure. Finally, we investigated the complex correlation across selected CpG sites and their relationship with smoking via conditional independence network inference.

## Methods

### Study population

The present work uses data from two European population-based prospective cohorts: the Italian component of the European Prospective Investigation into Cancer and Nutrition (EPIC-Italy, N = 697 men and women aged 35–72) and the Norwegian Women and Cancer Study Cohort (NOWAC, N = 442 women aged 35–65) [6].

EPIC-Italy included over 47,000 participants between 1993 and 1998, who provided anthropometric measurements, blood samples, and information on medical history and lifestyle factors collected through a self-administered questionnaire [6, 11]. Upon sampling, blood was transported to local laboratories and prepared for DNA extraction according to standard laboratory protocols [6, 11]. Within EPIC-Italy, we used data from a lung cancer nested case–control (CC) study including 192 incident lung cancer cases and 192 healthy controls matched to cases by sex, date

of birth, date of inclusion, and study centre. Additionally, we included 322 healthy control individuals from a breast and colon cancer study nested within EPIC-Italy [11] (Flow chart: Supplementary Fig. 1).

The NOWAC study recruited over 172,000 participants between 1991 and 2007 and collected anthropometric measurements, and self-reported information on medical history and lifestyle factors [16]. In a subset of (N = 50,000) participants, recruited between 2003 and 2006, a blood sample was also available. Upon collection, blood samples were sent out to the Department of Community Medicine at the University of Tromsø and subsequently prepared for DNA extraction according to previously described laboratory protocols [6, 17]. During the follow-up and to the end of 2011, 132 lung cancer cases with blood samples were identified and were used for the DNA methylation profiling [11]. For each case, one control with an available blood sample was selected and matched on time since blood sampling and year of birth. We also included data from 190 healthy control individuals from a breast and colon cancer study nested within NOWAC (Supplementary Fig. 1) [11].

We excluded seven and four lung cancer cases from EPIC-Italy and NOWAC, respectively, as well as two and eight controls, due to blood samples not passing the quality control checks, or to missing data for one or more covariates. This left 313 lung cancer cases (185 in EPIC-Italy, and 128 in NOWAC) and 826 controls (512 in EPIC-Italy, and 314 in NOWAC) for statistical analyses (Supplementary Fig. 1). Both EPIC-Italy and NOWAC studies were approved by relevant international, national, and local ethics committees and all participants provided signed informed consent.

### CpG methylation measurement and data pre-processing

For both cohorts, epigenome-wide analyses were carried out from whole blood cells DNA using the IlluminaInfiniumHumanMethylation450 array, with all laboratory procedures performed at the Human Genetics Foundation (Turin, Italy) according to the manufacturer's protocols [6]. In EPIC-Italy, data pre-processing was performed using in-house scripts as previously described [6]. Briefly, for each sample and each probe, measurements were set to missing if obtained by averaging intensities over less than three beads, or if averaged intensities were below detection thresholds estimated from negative control probes. Background subtraction and dye bias correction (probes using Infinium II design) were also performed. This procedure resulted in a subset of 473,929 CpG sites, of which probes detected in < 20% of the samples were excluded, yielding 465,886 CpG markers in the lung cancer nested study, and 443,150 in the breast and colon cancer controls samples (Supplementary Fig. 1). The same pre-processing procedure was performed for NOWAC, yielding

485,512 CpGs suitable for analyses. The final, merged study sample included 443,150 CpGs available in both cohorts, for a total of 1139 participants.

Further data processing included imputation of missing methylation values according to the *k*-nearest neighbour procedure ( $k = 10$ ), followed by the *M*-transformation of imputed CpG data, expressed as  $\log_2$ -transformed ratios of intensities arising from methylated cytosines over those arising from unmethylated cytosines.

To account for technically induced noise, we adopted a two stage ‘denoising’ strategy fitting a linear mixed model for the methylation level at each CpG site (as outcome variable) as a function of age at blood collection, sex, BMI, and case control status as fixed effect (predictors), and including the technical confounders: chip ID (177 modalities), position of the sample on the chip (12 modalities), and recruitment centre (six modalities) as random intercepts. Denoised methylation levels were then obtained by subtracting for the observed levels the estimated random effects [18, 19].

## Statistical analyses

### CpG site classification

We categorized the selected 443,150 CpG sites into two groups based on their reported association with smoking, using data from the largest meta-analysis investigating the effects of smoking on epigenome-wide CpG markers [20]. We define ‘smoking-related’ CpG sites as those ( $N = 2623$ ) found associated with smoking in that meta-analysis at a Bonferroni-corrected significance level ensuring a control of the Family Wise Error Rate below 0.05. Conversely, we define ‘smoking-unrelated’ CpG sites ( $N = 440,527$ ) as those not associated with any of the smoking metrics.

### Multivariate regression

Penalised logistic regression using Least Absolute Shrinkage and Selection Operator regularization (LASSO) was used to perform variable selection and identify a sparse set of CpG sites complementarily contributing to the risk of lung cancer. Models were adjusted for sex and age at blood collection (*base model*). To account for potential confounding by exposure to tobacco smoking, we also considered models further adjusted for the Comprehensive Smoking Index (CSI), a continuous score accounting for the duration and the intensity of smoking across the life-course (*CSI-adjusted model*) [21]. Adjustment was achieved by not applying any penalisation to the regression coefficients of the adjustment variables [22, 23]. To ensure reproducibility of the findings, LASSO regression was calibrated via stability by means of resampling [24]. Selection proportions for each predictor were computed over 1000 subsamples of 80% of the study

participants. The proportion of cases and controls in each subsample was controlled to be representative of that in the full population. Stably selected CpG sites were defined as those with selection proportions, computed over models fitted with a given penalty parameter, above a threshold. The penalty parameter and threshold in selection proportions are jointly calibrated by maximising a stability score derived from the likelihood of uniform (i.e. uninformative) feature selection [25]. The average beta-coefficients, conditionally on selection, estimated over the 1000 LASSO models with calibrated penalty are reported. To assess potential confounding by blood cell composition differentials, we ran our variable selection model adjusting (i.e. by non penalising) for estimated blood cell type proportions. We estimated blood cell type composition from the sentinel CpG sites proposed by Houseman [26] and of the six estimated proportions we adjusted for proportions of Monocytes, B cells, CD4+ T cells, Natural Killers, CD8+ T cells, and Neutrophil. We report and compare the recalibrated effect size estimates for the model without and with adjustment for blood cell composition.

### ROC analysis

Logistic regression models including the stably selected CpG sites were recalibrated on 1000 training sets (80% of the samples), and the discriminatory ability of each model was assessed in the out-of-bag test set (remaining 20% of the samples) by estimating the sensitivity, specificity, and Area under the Receiver Operating Characteristic (ROC) curve. Performances are reported in terms of average, 5th and 95th percentiles for these metrics computed across the 1000 test sets. In addition, to quantify the amount of information brought about by each of the stably selected CpG sites, a series of models sequentially adding the CpG sites by order of selection proportion were evaluated and their AUC was reported.

### Conditional independence networks

To better characterize DNA methylation changes associated with the future risk of lung cancer, we constructed a conditional independence network of lung cancer related CpG sites from the stability selection LASSO and also included CSI as a node in the network. The partial correlation network was estimated using stability selection applied to the graphical LASSO [24]. Selection proportions of the edges were estimated on ( $N = 1000$ ) subsamples of 50% of the population and our calibration jointly defined the penalty of the graphical LASSO and the threshold in selection proportion for an edge to be considered stable by maximizing the stability score, while ensuring that the expected number of False Positive selected edges is below 10 [25].

All statistical analyses were carried out using the R statistical software (version 4.0.3) using *glmnet* package and in-house script for stability selection and conditional independence networks. These available upon request to the corresponding author.

## Results

### Descriptive analyses

The characteristics of the study population are presented in Table 1. In EPIC-Italy, the mean age of participants at blood sampling was 54 years for controls, and 54.5 for lung cancer cases, whereas lung cancer cases were older in NOWAC (56 vs. 51.1 years). In both EPIC-Italy and NOWAC, there was a higher proportion of current and former smokers in lung cancer cases, as well as a higher smoking duration, smoking intensity and CSI.

### Associations with future lung cancer risk

A total of  $N=29$  CpG sites (including  $N=8$  smoking-related and  $N=21$  smoking-unrelated CpG sites) were stably selected in the base model (Fig. 1A, Supplementary Table 1). In the CSI-adjusted model,  $N=50$  CpG sites (1 smoking-related and 49 smoking unrelated sites) were selected. Of these,  $N=1$  smoking-related (*CIRBP-ASI-cg00073090*), and  $N=19$  smoking-unrelated CpG sites were selected across both models (Fig. 1B, Supplementary Table 2). We

observed lower selection proportions and effect sizes for the  $N=7$  smoking-related sites selected in the base model only (Fig. 1C, D). Despite stable inclusion in both models, *CIRBP-ASI-cg00073090* had a lower effect size in the CSI-adjusted model (average  $\beta$  coefficient of  $-1.35$  in the base model and  $-0.33$  in the adjusted model). Selection proportions and effect size estimates remained generally unchanged upon adjustment for CSI for the ( $N=19$ ) smoking-unrelated CpG sites selected in both models. An additional set of  $N=30$  smoking-unrelated CpG sites was stably selected in the CSI-adjusted model (out of a total of 50 selected CpG sites). Although these were not stably selected in the base model, they showed selection proportions above 0.46 and comparable average  $\beta$  coefficients (results not shown).

Effect sizes obtained from our recalibrated models with and without adjustment for estimated white blood cell proportions showed very good consistency in both models unadjusted (Supplementary Fig. 2A) and adjusted for CSI (Supplementary Fig. 2B), hence suggesting limited confounding by cell type differentials.

### ROC analyses: quantifying the disease relevant information

ROC analyses performed in the 1000 testing sets each including 20% of the full population yielded a mean AUC of 0.87 (5th–95th percentiles 0.87–0.88) for the model including age, sex and the  $N=29$  stably selected CpG sites from the base model, yielding an increase in average AUC of 0.26

**Table 1** Characteristics of study participants stratified by cohort and future lung cancer status. The mean (standard deviation) and counts (percentage) are reported for continuous and categorical variables respectively

	EPIC-Italy		NOWAC		Full population
	Controls (N=512)	Cases (N=185)	Controls (N=314)	Cases (N=128)	
Sex (women)	331 (65%)	81 (44%)	314 (100%)	128 (100%)	854 (75%)
Age (years)	54 (6.8)	54.5 (6.3)	51.1 (6.9)	56 (4.2)	53.5 (6.7)
<i>Smoking status</i>					
Never	257 (50%)	26 (14%)	136 (43%)	14 (11%)	433 (38%)
Former	143 (28%)	59 (32%)	97 (31%)	34 (27%)	333 (29%)
Current	112 (22%)	100 (54%)	81 (26%)	80 (62%)	373 (33%)
Smoking duration (years)	12.1 (14.2)	27.3 (14.3)	15.9 (16.5)	31.6 (14.8)	17.8 (16.6)
Smoking intensity (cig./day)	6 (8.9)	14.4 (9.4)	5.5 (5.8)	10.3 (5.5)	7.7 (8.6)
Comprehensive Smoking Index (CSI)	0.5 (0.7)	1.4 (0.8)	0.7 (0.8)	1.4 (0.7)	0.8 (0.8)
Time to diagnosis (years)		7.2 (3.7)		3.9 (2.0)	5.9 (3.5)
<i>Centre (EPIC-Italy)</i>					
Florence	92 (18%)	63 (34%)			
Naples	11 (2%)	3 (2%)			
Ragusa	29 (6%)	14 (8%)			
Turin	246 (48%)	60 (32%)			
Varese	134 (26%)	45 (24%)			

over models including age and sex only and an increase of 0.09 over the model including age, sex and CSI (Fig. 2A), which had an AUC of 0.78 (5th–95th percentiles 0.77–0.78). The addition of each stably selected CpG site only incrementally improved the AUC, with the largest contributions from smoking-related CpGs, namely *CIRBP-AS1-cg00073090* and *F2RL3-cg03636183* (Fig. 2B:  $\Delta$ AUC > 4%). The best performing model was that including CSI and the 50 CpG sites selected in the CSI-adjusted model. It yielded an AUC of 0.94 (5th–95th percentiles 0.94–0.95), representing an increase in mean AUC of 0.16 compared to that of the model including age, sex and CSI, and an AUC increase of 0.07 compared to that of the model including age, sex and all (N = 29) stably selected CpG sites in the base model (Fig. 2A). For both models, we observed a limited increase in AUC by including CpG sites with selection proportion below the calibrated threshold (Fig. 2B, C).

### Conditional Independence Network

The stability-enhanced conditional independence network of CSI and the 29 CpG sites selected in the base model (Fig. 3) included (N = 20) edges between a set of (N = 9) inter-connected nodes (module), mostly constituted of smoking-related CpG sites (N = 7). Of these, four were directly related to CSI (*AHRR-cg05575921*, *ALPP2-cg01940273*, *F2RL3-cg03636183* and *GNG12-AS1-cg25189904*). The remaining stably selected CpG sites (1 smoking-related, 19 smoking-unrelated) were not connected to any other node, suggesting their independence and weaker links to smoking.

### Discussion

We performed variable selection from full-resolution DNA-methylation profiles using stability selection approaches. Of the 443,150 assayed CpG sites, we identified a sparse subset of 29 CpG sites in the base LASSO model, jointly associated with the future risk of developing lung cancer. These included (N = 8) sites which were previously reported to associate with smoking and (N = 21) CpG sites, which were not.

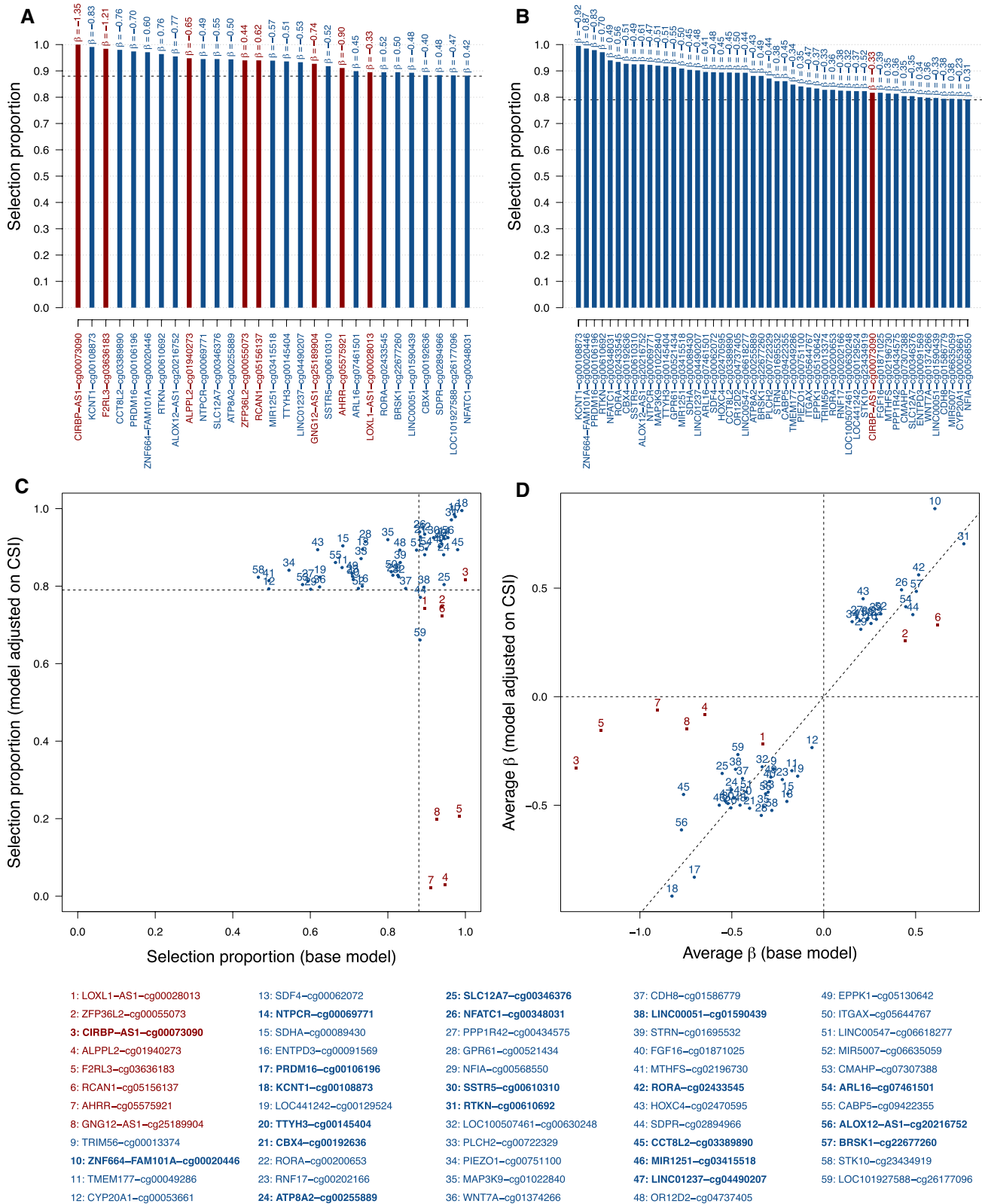
Our conditional independence network shows that 7 of the 8 smoking-related CpG sites are inter-connected and (directly or indirectly) linked to CSI, while the (N = 21) CpG sites not reported to be associated with smoking appeared more independent and not related to smoking in our data. This lends plausibility to our CpG sites classification and suggests that the smoking-unrelated CpG sites we found associated with lung cancer risk may each capture non-redundant and complementary disease-relevant information.

Among the selected smoking-related CpG sites, three were previously identified as being associated with lung

cancer risk (*AHRR-cg0557921*, *F2RL3-cg03636183* and *ALPPL2-cg01940273*) [11, 27, 28]. Conversely, *CIRBP-AS1-cg00073090* was not previously associated to lung cancer risk, although recent work identified *CIRBP* as hypomethylated in cases developing occult lymph node metastases among non-small cell lung cancer cases [29]. A genome wide analysis conducted with multivariate analysis also identified *CIRBP* as an important prognostic gene for non-small cell carcinoma [30].

Models adjusted for CSI selected only one CpG found to be associated with smoking (*CIRBP-AS1-cg00073090*), possibly indicating that this site may reflect smoking exposure that is not fully captured by CSI. In a subset of our study population, we found that methylation levels at *CIRBP-AS1-cg00073090* were only nominally associated with childhood exposure to tobacco smoke and that this association did not survive correction for multiple testing and adjustment for CSI. This suggests that early-life exposure cannot fully explain the association linking methylation level at *CIRBP-AS1-cg00073090* and future risk of lung cancer. However, short-term, environmental/occupational, second-hand exposure to tobacco smoke can potentially confound this association.

The remaining lung cancer-relevant information is captured by a set of (N = 49) smoking-unrelated CpG sites, potentially reflecting a large range of biological pathways. The fact that we highlighted novel, smoking-unrelated CpG sites upon accounting for the effect of CSI, tends to be in line with former investigations, which also capture weaker effect-size and more heterogeneous signals when compared to the generally conserved smoking-related epigenetic signature (i.e. *AHRR-cg05575921*, *F2RL3-cg03636183*) [15]. Among these 49 smoking-unrelated markers, 33 were found hypomethylated and 17 hypermethylated in lung cancer cases, and our discussion will be restricted to those presenting stronger effects and a high selection proportion ( $|\beta| > 0.5$ , selection proportion > 0.9, N = 14). Following adjustment for CSI, the most frequently selected CpG site was *KCNT1-cg00108873* which was hypomethylated in lung cancer cases. *KCNT1* is a protein coding gene involved in the intracellular potassium activated channel activity. It has, to our knowledge never been associated to lung cancer risk, although recent work on gene expression identified low expression of *KCNT1* consistently in four cancers, including lung [31]. Additional smoking-unrelated CpG sites selected in the CSI-adjusted model included *PRDM16-cg00106196*, *CBX4-cg00192636*, *SSTR5-cg00610310*, *ALOX12-AS1-cg20216752*, *NTPCR-cg00069771*, *MAP3K9-cg01022840*, *TTYH3-cg00145404*, *GPR61-cg00521434*, and *MIR1251-cg03415518*, which were all hypomethylated ( $\beta < -0.5$ ), whereas *ZNF664-FAM101A-cg0020446*, *RTKN-cg00610692*, *NFATC1-cg00348031*, and *RORA-cg02433545* were hypermethylated ( $\beta > 0.5$ ). We



found that PRDM16 encodes for a zinc transcription factor which controls the development of brown fat cells. While its role in lung cancer has been poorly investigated, recent

work has shown that it could potentially be an interesting therapeutic target for lung adenocarcinomas [32]. Further, we found that CBX4 has been specifically related with lung

**Fig. 1** Stability selection models exploring the joint associations between CpG sites and the future risk of lung cancer. Selection proportions of stably selected CpG sites are derived from LASSO-penalised logistic models for the risk of lung cancer including all  $N=443,150$  CpG sites as predictors and adjusted for age, sex (**A**) and CSI (**B**). Comparison of selection proportions (**C**) or  $\beta$ -coefficients (**D**) from the base versus CSI-adjusted models for CpG sites that are stably selected in at least one of these two models. The list of stably selected CpG sites is reported at the bottom, with overlapping signals in bold. CpG sites related to smoking at a Bonferroni-corrected significance level ensuring a family-wise error rate below 0.05 are presented in red, and sites unrelated to smoking in blue

cancer development, acting as an oncogene that enhances cell proliferation and promotes cancer cell migration [33]. *SSTR5* encodes for the somatostatin receptor type 5, which is involved in signalling to alter hormone secretion, increase apoptosis, and decrease cellular proliferation. In concordance with our findings, differential expression of *SSTR5* has been observed in neuroendocrine lung cancer cases in respect to controls [34]. *TTHY3* is a gene coding for an intracellular calcium activated chloride channel activity. No observations linking *TTHY3* to lung cancer risk were found, although its differential expression has been linked to poor clinical outcomes in cases of gastric cancer [35]. We have also found that *ALOX12* (arachidonate 12-lipoxygenase) was associated with lung cancer risk. This gene has not been specifically related to lung cancer risk, but former studies have reported (i) that hypomethylation at one of its CpG site was associated with inflammation of the airways, and (ii) that *ALOX12* expression was upregulated in breast cancer tissue [36–38]. We could not find evidence for an association between *ZNF664-FAM101A* and *NTPCR* genes and lung cancer, although *ZNF664-FAM101* expression was found to be up-regulated in gastric cancer tissue [39]. Among the CpG sites we identified, some were located in intragenic regions of the *RTKN1*, *GPR61*, *MAP3K9*, and *CABP5* genes, whose functions include cell signalling, cell cycle control, and cell growth regulation [40]. Although only *MAP3K9* has been specifically related with lung cancer, acting as a cell proliferation promoting factor (oncogene) [41], previous research has reported that an increased expression of *RTKN1* was observed in multiple cancer types, including lung cancer [42, 43], whereas *CABP5* and *GPR61* were previously related with increased RNA expression in human gastric cancer tissue, and aberrant methylation patterns resulting from air pollution, respectively [44, 45]. Exploring the roles of *NFATC1*, *RORA*, and *MIR1251*, we found that all three genes were implicated in lung cancer, with *NFATC1* being an oncogene, *RORA* being a key circadian clock regulator in non-small cell lung carcinoma, and *MIR1251* encoding for non-coding RNA which promotes cell migration and invasion by lung cancer cells [46–48].

Finally, we explored the role of the 36 remaining markers by applying a gene ontology approach, in order to highlight

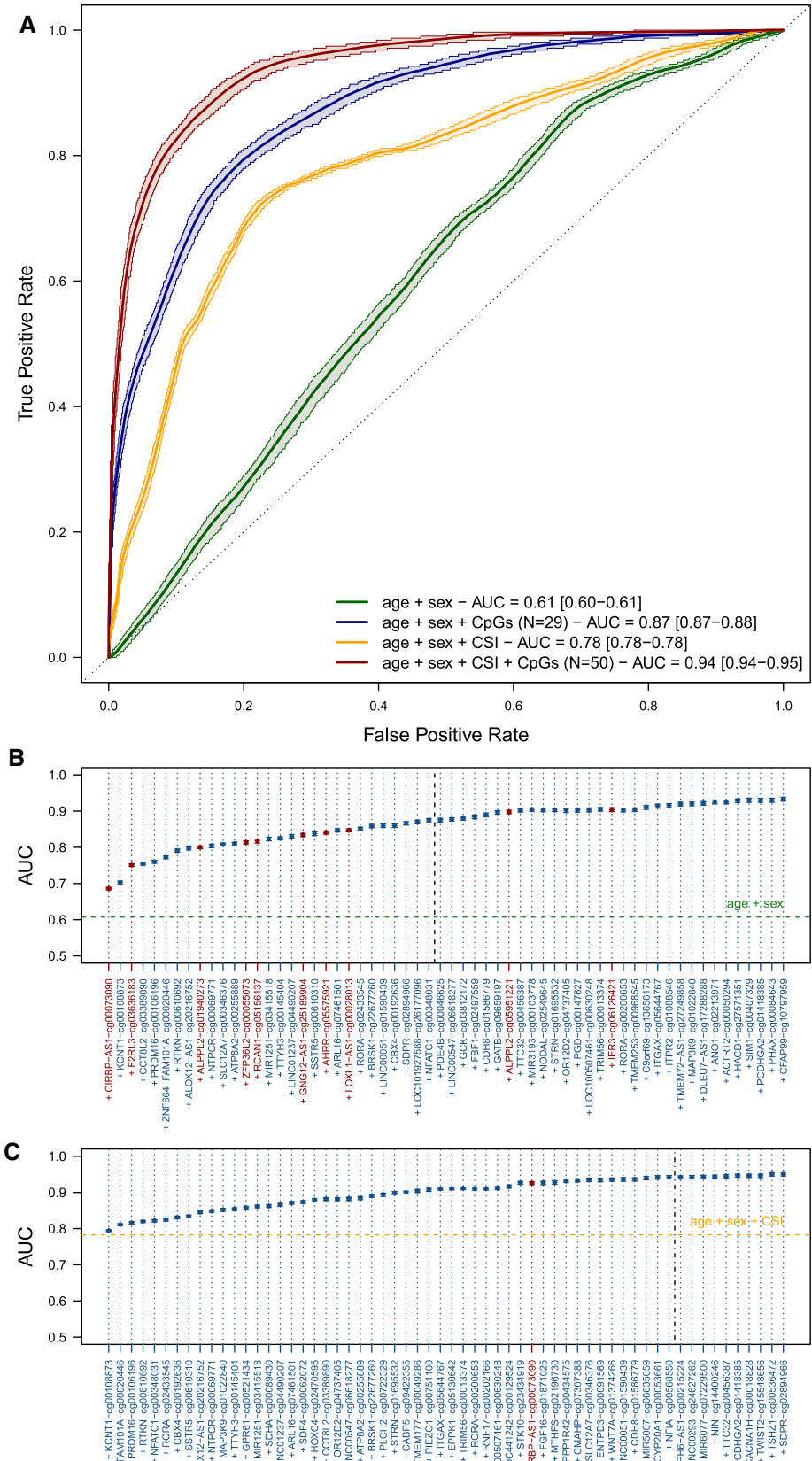
the general processes in which the identified genes may be involved [49]. Seven major pathways were identified, including FGF, Integrin, Wnt, and Cadherin signalling pathways, as well as Angiogenesis, Gonadotropin-releasing hormone receptor, and the Alzheimer disease-presenilin pathway, with all seven pathways being implicated in lung cancer [50–56].

## Strengths and limitations

To our knowledge, this is the first study to investigate associations between epigenome-wide methylation profiles and lung cancer risk by applying stability selection. This approach has allowed us to highlight two sets of CpG sites that were explanatory of lung cancer: those related and those unrelated to smoking.

Our study also has several limitations. First, we examined DNA methylation from blood, and not from more proximal tissues such as the lung epithelium, buccal cells, or sputum, which may represent a challenge regarding the interpretability of the identified CpG sites from the pathophysiological perspective. Second, we used a classification of CpGs according to their associations with smoking in previous large meta-analyses [20]. Our classification relies on marginal associations, which may result in potential misclassification of CpG sites and may overlook potential interactions between smoking-related CpG sites and those initially classified as smoking-unrelated [20]. Nevertheless, our conditional independence network seems to support this classification. In addition, exposure to passive smoking has not been taken into account in our analyses, hence, some of the CpG sites we refer to as being smoking unrelated CpG sites may still be related to other routes of exposure to tobacco smoke (e.g. second-hand smoking, environmental exposure, early-life exposure). Further, even though we observed a high consistency across smoking-unrelated CpGs found in the base and CSI-adjusted models ( $N=19$  smoking-unrelated CpG sites selected across both models), an additional ( $N=30$ ) smoking-unrelated CpG sites were selected in the CSI-adjusted model. These consist of several (uncorrelated) CpG sites capturing complex (and inherently multidimensional) molecular pathways, other than those related to smoking, involved in lung carcinogenesis. Another limitation of our work is that women are largely over-represented in our study. Although sex-stratified analyses showed similar predictive performances of our selected CpG sites in men and women separately (results not shown), additional work would be warranted to assess the potential sex bias that could have been introduced as well as the effect of sex-specific smoking patterns, by replicating our results in other studies. Although our results are based on the comparison of prospective cases to controls, reverse causation, whereby some of the

**Fig. 2** Receiver Operating Characteristic curve for lung cancer prediction. Mean and 5th–95th percentiles of the Area Under the Curve (AUC) were calculated across the 1000 recalibrated models including (i) age and sex (green), (ii) age, sex and stably selected CpG sites from the base model (dark blue), (iii) age, sex and CSI (orange), and (iv) age, sex, CSI and the stably selected CpG sites from the adjusted model (dark red) (A). Mean and 5th–95th percentiles of the AUC are reported for models sequentially including the first 50 CpG sites by order of selection proportion in the base (B) and adjusted (C) models. Calibrated stability selection models are indicated by a black dashed vertical line. CpG sites related to smoking at a Bonferroni-corrected significance level ensuring a family-wise error rate below 0.05 are presented in red, and sites unrelated to smoking are presented in blue







explanatory of lung cancer risk. If validated in independent data, these may prove instrumental in understanding biological mechanisms involved in lung carcinogenesis that are not directly linked to smoking. Overall, our findings identify sets of differentially methylated sites that are jointly explanatory of future lung cancer risk, these may provide leads into the mechanisms involved in lung cancer development (with and without the implication of smoking) and may prove useful for better early identification of patients at higher risk of lung cancer.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10654-022-00877-2>.

**Acknowledgements** We thank Prof. Benoît Liquet and Julien Chiquet for their insightful comments on stability selection.

**Author contributions** DP and BB are joined first authors. MC-H, FG are joined last authors. MC-H, FG, BB, DP conceived the study. MC-H, FG, DP, BB, SD interpreted the results and drafted the manuscript. BB, DP, FG performed the statistical analyses. MC-H and FG, supervised the analyses and together with RV drafted the analytical plan. MW, MK, GC, P-AD contributed to the development of the analytical pipeline, and to the analyses. The EPIC Italy data was provided by PV, SP, VK, DP, CS, SP, RT. NOWAC data was provided and curated by EL, THN and TMS. RM, GG, GS, MS, SS, MB contributed to the study design and results interpretation and to the paper drafting. All authors revised the manuscript for important intellectual content and approved the submission of the manuscript.

**Funding** This work was supported by Cancer Research UK Population Research Committee ‘Mechanomics’ project Grant (Grant #22184 to MC-H). The NOWAC post-genome cohort study was funded by the ERC advanced Grant; Transcriptomics in Cancer Epidemiology (ERC-2008-AdG-232997). MC-H, FG, KS-B, THN, MJ, and TS acknowledge support from the Research Council of Norway (Id-Lung project FRIPRO 262111 to TS). MC-H, RV, acknowledge support from the H2020-EXPANSE project (Horizon 2020 Grant No 874627 to RV). SD acknowledges support to Horizon 2020 Marie Skłodowska-Curie fellowship EXACT Identifying biomarkers of EXposure leading to Lung Cancer with AdducTomics (Grant # 708392 to SD). DP was supported by a Grant from the Swiss National Science Foundation (SNSF Doc-Mobility Grant P1LAP3\_178061). BB received a PhD studentship from the MRC Centre for Environment and Health.

**Data availability** The EPIC and NOWAC data cannot be shared publicly because of local and national ethical and security policy. Data access for researchers will be conditional on adherence to both the data access procedures of the Norwegian Women and Cancer Cohort and the UiT The Arctic University of Norway (contact via Torkjel Sandanger, [torkjel.sandanger@uit.no](mailto:torkjel.sandanger@uit.no), Tonje Braaten [tonje.braaten@uit.no](mailto:tonje.braaten@uit.no), and Arne Bastian Wiik, [arne.b.wiik@uit.no](mailto:arne.b.wiik@uit.no)) for NOWAC in addition to the local ethical committee. Access to EPIC data can be request upon application and acceptance from the project steering committee ([epicadmin@imperial.ac.uk](mailto:epicadmin@imperial.ac.uk)), data sharing will also be conditional to ethic approval and compliance to GDPR regulations.

**Code availability** The computer code used to support the findings of this study are available on GitHub, at [https://github.com/barbarabodinier/Epigenetic\\_mechanisms\\_lung\\_cancer](https://github.com/barbarabodinier/Epigenetic_mechanisms_lung_cancer)

## Declarations

**Conflict of interest** MC-H holds shares in the O-SMOSE company and has no conflict of interest to disclose. Consulting activities conducted by the company are independent of the present work. The authors have no conflict of interest to disclose. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

**Ethical approval and consent to participate** All study participants gave written informed consent for the study. For EPIC Italy, the research was approved by the Ethics Committees at the Italian Institute of Genomic Medicine (IIGM, Turin, Italy). For NOWAC, the study was approved by the Regional Committee for Medical and Health Research Ethics in North Norway.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References


1. Wild C, Weiderpass E, Stewart BW (Eds.) (2020) World cancer report: cancer research for cancer prevention. IARC Press.
2. Balgkouranidou I, Liloglou T, Lianidou ES. Lung cancer epigenetics: emerging biomarkers. *Biomark Med.* 2013;7(1):49–58.
3. Begum S, Brait M, Dasgupta S, et al. An epigenetic marker panel for detection of lung cancer using cell-free serum DNA. *Clin Cancer Res.* 2011;17(13):4494–503.
4. Doll R, Hill AB. Lung cancer and other causes of death in relation to smoking. *BMJ.* 1956;2(5001):1071.
5. Hecht SS. Tobacco smoke carcinogens and lung cancer. *JNCI J Natl Cancer Inst.* 1999;91(14):1194–210.
6. Guida F, Sandanger TM, Castagné R, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Genet.* 2015;24(8):2349–59. <https://doi.org/10.1093/hmg/ddu751>.
7. Maas SC, Vidaki A, Wilson R, et al. Validated inference of smoking habits from blood with a finite DNA methylation marker set. *Eur J Epidemiol.* 2019;34(11):1055–74.
8. Brothers JF, Hijazi K, Mascaux C, El-Zein RA, Spitz MR, Spira A. Bridging the clinical gaps: genetic, epigenetic and transcriptomic biomarkers for the early detection of lung cancer in the post-national lung screening trial era. *BMC Med.* 2013;11(1):168.
9. Spira A, Beane J, Shah V, et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci.* 2004;101(27):10143–8.
10. Teschendorff AE, Yang Z, Wong A, et al. Correlation of smoking-associated DNA methylation changes in buccal cells with

- DNA methylation changes in epithelial cancer. *JAMA Oncol.* 2015;1(4):476–85.
11. Baglietto L, Ponzi E, Haycock P, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *Int J Cancer.* 2017;140(1):50–61. <https://doi.org/10.1002/ijc.30431>.
  12. Fasanelli F, Baglietto L, Ponzi E, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun.* 2015;6:10192. <https://doi.org/10.1038/ncomms10192>.
  13. Zhang Y, Yang R, Burwinkel B, Breitling LP, Brenner H. F2RL3 methylation as a biomarker of current and lifetime smoking exposures. *Environ Health Perspect.* 2014;122(2):131–7. <https://doi.org/10.1289/ehp.1306937>.
  14. Dugué P-A, Jung C-H, Joo JE, et al. Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. *Epigenetics.* 2020;15(4):358–68. <https://doi.org/10.1080/15592294.2019.1668739>.
  15. Zhao N, Ruan M, Koestler DC, et al. Epigenome-wide scan identifies differentially methylated regions for lung cancer using pre-diagnostic peripheral blood. *Epigenetics.* 2021;1–13.
  16. Lund E, Dumeaux V, Braaten T, et al. Cohort profile: the Norwegian women and cancer study—NOWAC—Kvinner og kreft. *Int J Epidemiol.* 2007;37(1):36–41.
  17. Dumeaux V, Børresen-Dale A-L, Frantzen J-O, Kumle M, Kristensen VN, Lund E. Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. *Breast Cancer Res.* 2008;10(1):R13.
  18. Chadeau-Hyam M, Vermeulen R, Hebls D, et al. Prediagnostic transcriptomic markers of chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis. *Ann Oncol.* 2014;25(5):1065–72.
  19. McHale CM, Zhang L, Lan Q, et al. Global gene expression profiling of a population exposed to a range of benzene levels. *Environ Health Perspect.* 2011;119(5):628–40.
  20. Joehanes R, Just Allan C, Marioni Riccardo E, et al. Epigenetic signatures of cigarette smoking. *Circ Cardiovasc Genet.* 2016;9(5):436–47. <https://doi.org/10.1161/CIRCGENETICS.116.001506>.
  21. Remen T, Pintos J, Abrahamowicz M, Siemiatycki J. Risk of lung cancer in relation to various metrics of smoking history: a case-control study in Montreal. *BMC Cancer.* 2018;18(1):1275. <https://doi.org/10.1186/s12885-018-5144-5>.
  22. Chadeau-Hyam M, Campanella G, Jombart T, et al. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen.* 2013;54(7):542–57.
  23. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated regions in the human genome. *Epigenet Chromatin.* 2015;8(1):6. <https://doi.org/10.1186/1756-8935-8-6>.
  24. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Ser B.* 2010;72(4):417–73.
  25. Bodinier B, Filippi S, Nøst TH, Chiquet J, Chadeau-Hyam M. Automated calibration for stability selection in penalised regression and graphical models: a multi-OMICs network application exploring the molecular response to tobacco smoking. *arxiv.* 2021.
  26. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform.* 2012;13(1):86.
  27. Battram T, Richmond RC, Baglietto L, et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. *Int J Epidemiol.* 2019;48(5):1493–504. <https://doi.org/10.1093/ije/dyz190>.
  28. Sandanger TM, Nøst TH, Guida F, et al. DNA methylation and associated gene expression in blood prior to lung cancer diagnosis in the Norwegian women and cancer cohort. *Sci Rep.* 2018;8(1):1–10.
  29. Chen Z, Xiong S, Li J, et al. DNA methylation markers that correlate with occult lymph node metastases of non-small cell lung cancer and a preliminary prediction model. *Transl Lung Cancer Res.* 2020;9(2):280–7. <https://doi.org/10.21037/tlcr.2020.03.13>.
  30. He R, Zuo S. A robust 8-gene prognostic signature for early-stage non-small cell lung cancer. *Front Oncol.* 2019;9:693. <https://doi.org/10.3389/fonc.2019.00693>.
  31. Sehic E, Hadrovic A, Dogan S. Detection and analysis of stable and flexible genes towards a genome signature framework in cancer. *Bioinformatics.* 2019;15(10):772–9. <https://doi.org/10.6026/97320630015772>.
  32. Fei L-R, Huang W-J, Wang Y, et al. PRDM16 functions as a suppressor of lung adenocarcinoma metastasis. *J Exp Clin Cancer Res.* 2019;38(1):35. <https://doi.org/10.1186/s13046-019-1042-1>.
  33. Hu C, Zhang Q, Tang Q, et al. CBX4 promotes the proliferation and metastasis via regulating BMI-1 in lung cancer. *J Cell Mol Med.* 2020;24(1):618–31.
  34. Muscarella LA, D'Alessandro V, la Torre A, et al. Gene expression of somatostatin receptor subtypes SSTR2a, SSTR3 and SSTR5 in peripheral blood of neuroendocrine lung cancer affected patients. *Cell Oncol.* 2011;34(5):435–41. <https://doi.org/10.1007/s13402-011-0025-9>.
  35. Saha B, Gil C. High expression of TTYH3 is related to poor clinical outcomes in human gastric cancer. *J Clin Med.* 2019;8(11):1762. <https://doi.org/10.3390/jcm8111762>.
  36. Bisgaard H. Persistent wheezing in very young preschool children reflects lower respiratory inflammation. *Am J Respir Crit Care Med.* 2001;163(6):1290–1. <https://doi.org/10.1164/ajrccm.163.6.ed1801b>.
  37. Morales E, Bustamante M, Vilahur N, et al. DNA hypomethylation at ALOX12 is associated with persistent wheezing in childhood. *Am J Respir Crit Care Med.* 2012;185(9):937–43. <https://doi.org/10.1164/rccm.201105-0870OC>.
  38. Natarajan R, Esworthy R, Bai W, Gu J-L, Wilczynski S, Nadler J. Increased 12-lipoxygenase expression in breast cancer tissues and cells. Regulation by epidermal growth factor. *J Clin Endocrinol Metab.* 1997;82(6):1790–8.
  39. Stephen R, Crabtree J, Yoshimura T, Clayton C, Dixon M, Robinson P. Increased zinc finger protein zFOC1 transcripts in gastric cancer compared with normal gastric tissue. *Mol Pathol.* 2003;56(3):167.
  40. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinform.* 2016;54(1):1–30.
  41. Fawdar S, Trotter EW, Li Y, et al. Targeted genetic dependency screen facilitates identification of actionable mutations in FGFR4, MAP3K9, and PAK5 in lung cancer. *Proc Natl Acad Sci.* 2013;110(30):12426–31.
  42. Uhlen M, Zhang C, Lee S, et al. A pathology atlas of the human cancer transcriptome. *Science.* 2017. <https://doi.org/10.1126/science.aan2507>.
  43. Wang S, Bian C, Yang Z, et al. miR-145 inhibits breast cancer cell growth through RTKN. *Int J Oncol.* 2009;34(5):1461–6.
  44. Li W, Li J-F, Qu Y, et al. Comparative proteomics analysis of human gastric cancer. *World J Gastroenterol WJG.* 2008;14(37):5657–64. <https://doi.org/10.3748/wjg.14.5657>.
  45. Feng F, Huang L, Zhou G, Wang J, Zhang R, Li Z, Zhang Y, Ba Y. GPR61 methylation in cord blood: a potential target of prenatal exposure to air pollutants. *Int J Environ Health Res.* 2020;32(2):463–72. <https://doi.org/10.1080/09603123.2020.1773414>.

46. He W, Lu J. MiR-338 regulates NFATc1 expression and inhibits the proliferation and epithelial-mesenchymal transition of human non-small-cell lung cancer cells. *Mol Genet Genomic Med.* 2020;8(2):e1091. <https://doi.org/10.1002/mgg3.1091>.
47. Xian H, Li Y, Zou B, et al. Identification of TIMELESS and RORA as key clock molecules of non-small cell lung cancer and the comprehensive analysis. *BMC Cancer.* 2022;22(1):1–17.
48. Ping M, Wang S, Chen Y, Jia J. The short non-coding RNA 1251–5p regulates stemness transformation and inhibits aggression of lung malignant tumor cells. *J Biomater Tissue Eng.* 2021;11(5):982–9.
49. Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 2003;13(9):2129–41. <https://doi.org/10.1101/gr.772403>.
50. Marek L, Ware KE, Fritzsche A, et al. Fibroblast growth factor (FGF) and FGF receptor-mediated autocrine signaling in non-small-cell lung cancer cells. *Mol Pharmacol.* 2009;75(1):196–207.
51. Caccavari F, Valdembrì D, Sandri C, Bussolino F, Serini G. Integrin signaling and lung cancer. *Cell Adh Migr.* 2010;4(1):124–9.
52. Mazieres J, He B, You L, Xu Z, Jablons DM. Wnt signaling in lung cancer. *Cancer Lett.* 2005;222(1):1–10.
53. Bae G-Y, Choi S-J, Lee J-S, et al. Loss of E-cadherin activates EGFR-MEK/ERK signaling, which promotes invasion via the ZEB1/MMP2 axis in non-small cell lung cancer. *Oncotarget.* 2013;4(12):2512.
54. Herbst RS, Onn A, Sandler A. Angiogenesis and lung cancer: prognostic and therapeutic implications. *J Clin Oncol.* 2005;23(14):3243–56.
55. Lu C, Huang T, Chen W, Lu H. GnRH participates in the self-renewal of A549-derived lung cancer stem-like cells through upregulation of the JNK signaling pathway. *Oncol Rep.* 2015;34(1):244–50.
56. Yun H, Park M, Kim D, et al. Loss of presenilin 2 is associated with increased iPLA2 activity and lung tumor development. *Oncogene.* 2014;33(44):5193–200.
57. Logue MW, Smith AK, Wolf EJ, et al. The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics.* 2017;9(11):1363–71.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Dusan Petrovic<sup>1,2,3</sup> · Barbara Bodinier<sup>1</sup> · Sonia Dagnino<sup>1</sup> · Matthew Whitaker<sup>1</sup> · Maryam Karimi<sup>1,4,5</sup> · Gianluca Campanella<sup>1</sup> · Therese Haugdahl Nøst<sup>6</sup> · Silvia Polidoro<sup>7</sup> · Domenico Palli<sup>8</sup> · Vittorio Krogh<sup>9</sup> · Rosario Tumino<sup>10</sup> · Carlotta Sacerdote<sup>11</sup> · Salvatore Panico<sup>12</sup> · Eiliv Lund<sup>6,13</sup> · Pierre-Antoine Dugue<sup>14,15,16</sup> · Graham G. Giles<sup>14,15,16</sup> · Gianluca Severi<sup>17</sup> · Melissa Southey<sup>14,16,18</sup> · Paolo Vineis<sup>1</sup> · Silvia Stringhini<sup>2,3</sup> · Murielle Bochud<sup>2</sup> · Torkjel M. Sandanger<sup>6</sup> · Roel C. H. Vermeulen<sup>1,19,20</sup> · Florence Guida<sup>1,21</sup> · Marc Chadeau-Hyam<sup>1,19</sup> 

<sup>1</sup> Department of Epidemiology and Biostatistics, MRC Centre for Environment and Health, School of Public Health, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, UK

<sup>2</sup> Department of Epidemiology and Health Systems (DESS), University Centre for General Medicine and Public Health (UNISANTE), Lausanne, Switzerland

<sup>3</sup> Department and Division of Primary Care Medicine, University Hospital of Geneva, Geneva, Switzerland

<sup>4</sup> Bureau de Biostatistique et d'Épidémiologie, Institut Gustave Roussy, Université Paris-Saclay, Villejuif, France

<sup>5</sup> Oncostat U1018, Inserm, Équipe Labellisée Ligue Contre Le Cancer, Université Paris-Saclay, Villejuif, France

<sup>6</sup> Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway

<sup>7</sup> Italian Institute for Genomic Medicine, Turin, Italy

<sup>8</sup> Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute-ISPO, Florence, Italy

<sup>9</sup> Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

<sup>10</sup> Hyblean Association for Epidemiological Research, AIRE- ONLUS, Ragusa, Italy

<sup>11</sup> Unit of Cancer Epidemiology Città Della Salute e della Scienza University-Hospital, Via Santena 7, 10126 Turin, Italy

<sup>12</sup> Department of Clinical Medicine and Surgery, Federico II University, Naples, Italy

<sup>13</sup> The Norwegian Cancer Registry, Oslo, Norway

<sup>14</sup> Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Australia

<sup>15</sup> Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia

<sup>16</sup> Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Australia

<sup>17</sup> Centre for Research in Epidemiology and Population Health, Inserm (Institut National de La Sante Et de a Recherche Medicale), Villejuif, France

<sup>18</sup> Department of Clinical Pathology, Melbourne Medical School, The University of Melbourne, Melbourne, Australia

<sup>19</sup> Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands

<sup>20</sup> Julius Centre for Health Sciences and Primary Care, University Medical Centre, Utrecht, Utrecht, The Netherlands

<sup>21</sup> Group of Genetic Epidemiology, International Agency for Research on Cancer (IARC) – World Health Organization (WHO), Lyon, France